

Multidimensional Opinion Mining from Social Data

Keith Cortis

B.Sc. IT (Hons)(Melit.), M.Appl.Sc.

A dissertation submitted in partial fulfilment of the
requirements for the award of

Doctor of Philosophy (PhD)

to the



Dublin City University

School of Computing

Supervisor: Dr. Brian Davis

August 2022

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: 

Keith Cortis

ID No: 19215923

Date: 29th August 2022

Contents

List of Abbreviations and Acronyms	v
List of Tables	vi
List of Figures	viii
Dedication	ix
Abstract	x
Acknowledgements	xi
1 Introduction	1
1.1 Problem Specification	3
1.2 Motivation	4
1.3 Research Questions	6
1.4 Contributions and Publications	11
1.5 Thesis Structure	16
2 Opinion Mining from Social Data: A Literature Analysis	17
2.1 Introduction	17
2.2 Context and Research Method of Systematic Review	18
2.2.1 Research Questions	19
2.2.2 Search Strategy	20
2.2.3 Search Application	22
2.2.4 Study Selection	24
2.2.5 Extraction of data	26
2.2.6 Synthesis of data	27
2.3 Social Opinion Mining Analysis	28
2.3.1 Opinion Mining vs. Social Opinion Mining	28
2.3.2 Issues and Challenges	29
2.3.3 Existing Approaches	32
2.3.4 Social Media Platforms	41
2.3.5 Social Datasets	43
2.3.6 Language	47
2.3.7 Modality	48
2.3.8 Tools and Technologies	51
2.3.9 Natural Language Processing Tasks	54
2.3.10 Latest research	58

2.4	Social Opinion Dimensions	61
2.4.1	Context	61
2.4.2	Different Social Opinion Dimensions	62
2.4.3	Impact of Sarcasm and Irony on Social Opinions	66
2.5	Application Areas	69
2.6	Research Gaps Conclusions	73
2.7	Conclusion	75
3	Social Opinion Data	77
3.1	Introduction	77
3.2	Dataset 1: Twitter Sentiment Gold Standard for the Brexit Referendum . .	80
3.2.1	Generation Process	80
3.2.2	Statistics and Discussion	83
3.2.3	Data Reuse	85
3.2.4	Applicability	85
3.2.5	Impact and Limitations	85
3.3	Dataset 2: Fine-grained Sentiment Analysis on Financial Microblogs and News Statements and Headlines	86
3.3.1	Generation Process	87
3.3.2	Data Reuse	92
3.3.3	Applicability	92
3.3.4	Impact and Limitations	94
3.4	Dataset 3: Multidimensional and Multilingual Social Opinions for Malta’s Annual Government Budget	95
3.4.1	Generation Process	96
3.4.2	Statistics and Discussion	106
3.4.3	Data Reuse	120
3.4.4	Applicability	121
3.4.5	Impact	122
3.5	Conclusion	123
4	Multidimensional Opinion Classification	125
4.1	Introduction	125
4.2	Social Opinion Classification Literature	126
4.2.1	Machine Learning and Deep Learning-based Algorithms	127
4.2.2	Neural Language Models	129
4.3	Baseline Classification Models	131
4.3.1	Dataset	131
4.3.2	Machine Learning-based Models	132
4.3.3	Neural-based Models	133
4.4	Experiments	134
4.4.1	Machine Learning-based Models	134
4.4.2	Neural-based Models	136
4.5	Results and Discussion	137
4.5.1	Machine Learning-based Models	137
4.5.2	Neural-based Models	141
4.6	Impact	144
4.7	Conclusion	144

5	Multidimensional Opinion Search and Summarisation	146
5.1	Introduction	146
5.2	Opinion Summarisation Literature	148
5.2.1	Automatic Summarisation	149
5.2.2	Opinion Summarisation of Social Data	150
5.2.3	Aspect-based Opinion Summarisation	152
5.2.4	Advancing over Aspect-based Opinion Summarisation of Social Data	156
5.3	Opinion Search and Summarisation System	157
5.3.1	Overview	157
5.3.2	Dataset	158
5.3.3	Technologies	158
5.3.4	Context behind Visualisations of Opinion Summaries	159
5.4	Human Evaluation Setup	161
5.4.1	Methodology	162
5.4.2	Participants' Sample	164
5.4.3	Systems' Overview	164
5.4.4	System Evaluation Strategy	165
5.4.5	Extrinsic Evaluation	167
5.4.6	Usability Evaluation	168
5.4.7	Subject Matter Expert Evaluation	169
5.5	Results and Analysis	169
5.5.1	Participants' Background	170
5.5.2	Participants' Knowledge	170
5.5.3	Extrinsic Evaluation Outcome	171
5.5.4	Usability Evaluation Outcome	174
5.5.5	Subject Matter Expert Evaluation Outcome	174
5.5.6	Feedback and Suggestions	176
5.6	Impact	178
5.7	Conclusion	181
6	Conclusion	182
6.1	Conclusions	182
6.2	Future Directions	189
6.3	Final Remarks	193
	Bibliography	194
	Appendices	249
A	Malta Government Budget Dataset Annotation Guidelines	249
B	Full Evaluation Results for each Social Opinion Dimension using Machine Learning-based Models	255
C	Full Evaluation Results for each Social Opinion Dimension using Neural-based Models	266
D	DCU Research Ethics Committee Application	285
E	Human Evaluation - Pre-Evaluation Questionnaire	304

F Human Evaluation - Overview of Opinion Search and Summarisation System	307
G Human Evaluation - Information About Systems	311
H Human Evaluation - Topic-based Tasks	313
I Human Evaluation - Opinion Summarisation Results' Questionnaire	329
J Human Evaluation - Generic Opinion Search and Summarisation System Usability	332
K Human Evaluation - Other Questions	334
L Human Evaluation - Subject Matter Expert Interview Questions	335

List of Abbreviations and Acronyms

API	Application Programming Interface
ANN	Artificial Neural Network
AUC ROC	Area Under the Curve Receiver Operating Characteristics
BERT	Bidirectional Encoder Representations from Transformers
CNB	Complement Naïve Bayes
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CSV	Comma Separated Values
DT	Decision Tree
EQF	European Qualifications Framework
GAN	Generative Adversarial Network
GPU	Graphical Processing Unit
HMTL	HyperText Markup Language
IE	Information Extraction
LR	Linear Regression
MBNB	Multivariate Bernoulli Naïve Bayes
MCC	Matthews Correlation Coefficient
MOS	Mean Opinion Score
NB	Naïve Bayes
NER	Named Entity Recognition
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
OPU	Online Post Usefulness
POMS	Profile of Mood States
POS	Part of Speech
RF	Random Forest
RNN	Recurrent Neural Network
RoBERTa	Robustly optimized BERT approach
ROC	Receiver Operating Characteristic
RSS	Really Simple Syndication
SemEval	Semantic Evaluation
SOU	Source Usefulness
STS	Stanford Twitter Sentiment
SUI	Summary Informativeness
SUS	System Usability Scale
SUU	Summary Usefulness
SVC	Support Vector Classification
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
XML	Extensible Markup Language

List of Tables

2.1	Metadata fields used in search application	23
2.2	Primary studies selection procedure from the electronic libraries	25
2.3	Approaches used in the studies analysed	32
2.4	Non-English languages supported by studies in this review analysis	49
2.5	Studies adopting a multimodal approach	49
2.6	Studies performing aspect-based Social Opinion Mining	57
2.7	Studies focusing on two or more social opinion dimensions	63
2.8	Studies adopting sarcasm and/or irony	68
3.1	Three social datasets generated for the purposes of this thesis	79
3.2	Inter-rater reliability measures for each annotation type	82
3.3	Agreement between annotators	83
3.4	Distribution of sentiment annotations	83
3.5	Distribution of strength annotations	84
3.6	Distribution of context dependence annotations	84
3.7	Details of Newswires data for each dataset	97
3.8	Details of Twitter data for each dataset	99
3.9	Malta Government Budget 2018 - Inter-rater reliability measures for each annotation type	102
3.10	Malta Government Budget 2019 - Inter-rater reliability measures for each annotation type	102
3.11	Malta Government Budget 2020 - Inter-rater reliability measures for each annotation type	102
3.12	Details of online posts for each topic annotation	105
3.13	Distribution of subjectivity annotations	106
3.14	Distribution of sentiment polarity annotations	107
3.15	Distribution of emotion annotations	109
3.16	Distribution of sarcasm annotations	111
3.17	Distribution of irony annotations	112
3.18	Distribution of negation annotations	113
3.19	Distribution of off-topic annotations	113
3.20	Distribution of language annotations	114
3.21	Distribution of annotations by data source	117
3.22	Online posts statistics - Times of Malta	118
3.23	Online posts statistics - MaltaToday	118
3.24	Online posts statistics - The Malta Independent	118
3.25	Online posts statistics - Twitter	118
4.1	Class distribution for each annotation per dataset	131

4.2	Classification model results - English dataset	139
4.3	Classification model results - Maltese-English and Maltese dataset	140
4.4	Evaluation results of all the pre-trained models for each social opinion dimension	142
5.1	Groups of participants by system and topic	166
5.2	Highest education of the participants based on the EQF	170
5.3	Mean Opinion Score (MOS) results for the summary usefulness (SUU), source usefulness (SOU), online post usefulness (OPU), and summary informativeness (SUI) extrinsic quality measures	172

List of Figures

2.1	Primary Studies by Year	26
2.2	Application Areas	70
5.1	OPINIONCLOUD opinion summary as extracted from [Potthast and Becker, 2010]	151
5.2	Example of a feature-based summary as extracted from [Hu and Liu, 2004]	152
5.3	Opinion Observer’s comparison screen as extracted from [Liu et al., 2005]	153
5.4	A sample opinion summary generated the sentiment summarisation system as extracted from [Bahrainian and Dengel, 2013]	153
5.5	An example of an extractive opinion summary as extracted from [Condori and Pardo, 2017]	154
5.6	An example of an abstractive opinion summary as extracted from [Condori and Pardo, 2017]	154
5.7	WikiAsp aspect-based summarisation model as extracted from [Hayashi et al., 2021]	155
5.8	Opinion Summary Visualisation about the “Innovation” topic using the full-text search feature	159
5.9	Opinion Summary Visualisation about the “Transport” topic and “Bicycle” sub-topic using the faceted search feature	160
5.10	Opinion Summary Visualisation - real-world use case	180

To my beloved daughter Emilja Marì who has motivated me to finish this PhD to spend more quality time with her. You mean the world to me and I promise you that you'll never walk alone.

Keith Cortis

Multidimensional Opinion Mining from Social Data

Abstract

Social media popularity and importance is on the increase due to people using it for various types of social interaction across multiple channels. This thesis focuses on the evolving research area of Social Opinion Mining, tasked with the identification of multiple opinion dimensions, such as subjectivity, sentiment polarity, emotion, affect, sarcasm, and irony, from user-generated content represented across multiple social media platforms and in various media formats, like textual, visual, and audio. Mining people's social opinions from social sources, such as social media platforms and newswires commenting sections, is a valuable business asset that can be utilised in many ways and in multiple domains, such as Politics, Finance, and Government. The main objective of this research is to investigate how a multidimensional approach to Social Opinion Mining affects fine-grained opinion search and summarisation at an aspect-based level and whether such a multidimensional approach outperforms single dimension approaches in the context of an extrinsic human evaluation conducted in a real-world context: the Malta Government Budget, where five social opinion dimensions are taken into consideration, namely *subjectivity*, *sentiment polarity*, *emotion*, *irony*, and *sarcasm*. This human evaluation determines whether the multidimensional opinion summarisation results provide added-value to potential end-users, such as policy-makers and decision-takers, thereby providing a nuanced voice to the general public on their social opinions on topics of a national importance. Results obtained indicate that a more fine-grained aspect-based opinion summary based on the combined dimensions of subjectivity, sentiment polarity, emotion, and sarcasm or irony is more informative and more useful than one based on sentiment polarity only. This research contributes towards the advancement of intelligent search and information retrieval from social data and impacts entities utilising Social Opinion Mining results towards effective policy formulation, policy-making, decision-making, and decision-taking at a strategic level.

Acknowledgements

The decision to pursue a PhD has been a great learning experience at both a personal and academic level which will stay with me forever. This 6 year PhD journey started in 2016 with the University of Passau in Germany, continued in 2018 at Maynooth University in Ireland, and finally settled at Dublin City University in 2019.

Firstly, I would really like to thank my supervisor Dr. Brian Davis who in 2018 provided me with the opportunity to continue my PhD in Ireland after a very challenging start to this journey. He has always provided me with valuable insights, relevant feedback, the support needed, and has always believed in me throughout my PhD. I would also like to mention that this research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106 P2 at the ADAPT SFI Research Centre at Dublin City University.

I would like to thank my parents Philip and Marthese who have always supported me and took great interest in my studies. The same applies to my brother Kelvin, his wife Ingrid, my nephews Eric and Ian, and my in-laws who have always supported me in the best way possible. I would also like to thank the friends that supported me and anyone who has contributed in some way or another, including the participants who accepted to take part in the human evaluation carried out as part of this PhD.

I am eternally grateful to my wife Stefania who has continuously encouraged me, supported me, motivated me, cared for me, and showed great patience during this long and tough journey. She has always showed me her love, lent an ear in times of need, and provided me with the adequate advice.

Last but not least, a special mention goes to my 9-month-old daughter Emilja Marì who has always gave me affectionate, infectious, and dazzling smiles each time she visited me in my office whilst writing this thesis. Those amazing smiles of yours have encouraged me to wrap up this chapter in life and helped me realise where my priorities stand post-PhD!

Sláinte

Chapter 1

Introduction

Finding out what other people think about a product or service has always been a very important part of an individual's and/or organisation's information gathering behaviour especially during a decision-making process. Before the World Wide Web, people asked their friends and colleagues about recommendations for an automobile mechanic, or about whom they plan to vote for in the upcoming elections, and checked with the consumer reports before buying a house appliance. Traditionally during a decision-making process, entities such as organisations carried out opinion polls, surveys, and focus groups to capture the general public opinion concerning their products and services [Liu, 2010]. In this regard, an **opinion** is defined as “a positive or negative sentiment, attitude, emotion or appraisal about an entity or an aspect of the entity from an opinion holder” [Liu and Zhang, 2012], where the opinion holder refers to the authors of the postings.

Social media usage is one of the most popular online activities due to its increased popularity and its importance for carrying out specific activities. In fact, public and enterprise interest in social media has been growing over the past years [Farzindar and Inkpen, 2015]. This is principally due to the large number of people who make use of different social media platforms for various types of social interaction, such as content creation and sharing [Farzindar and Inkpen, 2015]. According to [Kaplan and Haenlein, 2010], social media is defined as “a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content”. This definition fully reflects that social media

platforms are essential for online users to submit their views and read online posts about various aspects and/or entities, such as opinions about a political party they are supporting in an upcoming election, recommendations of products to buy, restaurants to eat in, and holiday destinations to visit.

Opinion Mining is defined as “the computational study of people’s opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes” [Liu and Zhang, 2012]. Opinion Mining [Pang and Lee, 2008, Liu and Zhang, 2012] as a research area is popular and extremely valuable especially for the exploitation of user-generated content extracted from social sources (all content is considered as social data). This evolving research area, also called **Social Opinion Mining**, deals with the identification of several opinion dimensions (**multidimensional**), such as *subjectivity*, *sentiment polarity*, *emotion*, *sarcasm*, and *irony*, from noisy social data derived from heterogeneous sources, such as social media platforms and online newswires commenting sections. The definition for each opinion dimension is as follows:

- **subjectivity**: determines whether a sentence expresses an opinion –in terms of personal feelings or beliefs– or not, in which case a sentence expresses objectivity;
- **sentiment polarity**: determines the polarity i.e., positive, negative, or neutral, of an expressed opinion;
- **emotion**: refers to a person’s subjective feelings and thoughts, such as love, joy, surprise, anger, sadness, and fear;
- **irony**: used to convey the opposite meaning of the actual things one says, but its purpose is not intended to hurt another person; and
- **sarcasm**: holds the “characteristic” of meaning the opposite of what one says, but unlike irony, it is used to hurt the other person towards whom the sarcasm is directed.

This chapter is structured as follows: Section 1.1 discusses the problem specification, with Section 1.2 presenting the motivation behind this thesis. The research questions are defined and discussed in Section 1.3, whereas the main contributions discussed in this thesis and the resulting publications are presented in Section 1.4. Section 1.5 provides an overview of the structure of this thesis.

1.1 Problem Specification

Opinion Mining is considered a challenging Natural Language Processing (NLP) problem, especially when applied on social data due to its semi-structured and/or unstructured format. People’s social opinions are not always taken into consideration and/or utilised to their full potential/extent in certain domains. Mining of people’s opinions, which are usually expressed in various media formats, such as textual (e.g., online posts, newswires), visual (e.g., images, videos), and audio, is a valuable business asset that can be utilised in many ways ranging from marketing strategies to product or service improvement [Cambria et al., 2013]. However as indicated in [Ravi and Ravi, 2015], dealing with unstructured data, such as video, speech, audio, and text, creates crucial research challenges. In fact, an aspect which makes such challenges even more pronounced is the lack of quality datasets in this field such as multilingual datasets or multidimensional opinion datasets (which go beyond sentiment analysis), notwithstanding *multilingual multidimensional* opinion datasets. Generating such datasets is challenging due to the annotation process being costly in terms of resources, domain knowledge, and expertise. Moreover, most of the existing Opinion Mining research that makes use of social data caters for either the English or the Chinese language (refer to Section 2.3.6 for more information) and focuses on subjectivity detection and sentiment analysis only, whereas the areas of emotion analysis, sarcasm detection, and irony detection are still evolving and hence not in a mature state. As a result of this, there is a lack of research that supports multiple languages or targets multilingual data on multiple Opinion Mining tasks, such as:

- **Opinion Classification:** classifies an opinion document (e.g., online post) as expressing a positive, negative, or neutral opinion (in terms of sentiment polarity) [Liu and Zhang, 2012];
- **Aspect-based Opinion Mining:** determines whether the opinion on an aspect, feature, topic, or entity is positive, negative, or neutral (in terms of sentiment polarity) [Liu and Zhang, 2012];
- **Opinion Search:** provides a general search for opinions [Liu, 2007]; and

- **Opinion Summarisation:** produces a structured summary from a large number of opinions on an object (e.g., product, service, topic, etc.) [Liu, 2007].

Furthermore, there is also lack of literature that caters for the multiple opinion dimensions (multidimensional), since as previously mentioned the most common form of Social Opinion Mining is based on subjectivity and/or sentiment polarity, which are referred to as subjectivity detection and sentiment analysis, respectively. Multidimensional Social Opinion Mining is important since the context-rich and fine-grained nature of the social opinions provides a nuanced voice to the general public given that current practice in Social Opinion Mining mostly limits to the use of a single opinion dimension (i.e., sentiment polarity) which results in a limited and simplified view of social opinions. Regarding the applicability of Social Opinion Mining, most of the research targets common areas and domains such as Finance, Marketing, Advertising, Sales, and Politics. Other research in non-traditional application areas and domains (e.g., Government) can help in determining whether Social Opinion Mining can influence certain multiple application areas.

1.2 Motivation

User-generated content is invaluable for certain needs, such as improving an entity's service or perception and tracking citizen opinion to aid policy-makers and decision-takers [Hilts and Yu, 2010]. Opinion-rich resources have been growing both in terms of availability and popularity. The year of 2001 marked the beginning of widespread awareness of the research problems and opportunities for Opinion Mining and Sentiment Analysis [Pang and Lee, 2008]. Online review sites and personal blogs were early examples of such opinionated resources, whereas social networking (e.g., Facebook¹), microblogging (e.g., Twitter²), travel (e.g., TripAdvisor³), and newswire (e.g., Reuters⁴) services are nowadays the most popular. The advent of the Social Web has created new opportunities and challenges for Opinion Mining, especially on user-generated content gathered from newswires commenting sections and social networking services. This allows users to create and share

¹<https://www.facebook.com>

²<https://www.twitter.com>

³<http://www.tripadvisor.com>

⁴<https://www.reuters.com>

content and their opinions directly to the public, thus circumventing possible forms of bias by acquaintance of experts only. Moreover, the summarisation and visualisation of such user-generated content can play a very important role in public Opinion Mining and decision-making [Sobhani et al., 2015].

Currently, Opinion Mining is used in several real-world scenarios, namely chatbots [Androusoy et al., 2019] that can understand a customer’s sentiment, emotion, and/or mood, and are able respond accordingly (e.g., offer an efficient and smooth service to dissatisfied customers to ensure their retention), adaptive customer online service based on identified customer sentiment and emotion [Yadollahi et al., 2017] (e.g., quick escalation to negative and angry users), tracking of overall customer satisfaction for a product or service [Zhao et al., 2019] (e.g., to quantify how effective your service is towards increasing satisfaction), and detection of changes in customer opinion towards a brand, product, or service [Geetha et al., 2017] (e.g., use insights to monitor trends and identify if overall opinion on your entity rises or drops). Moreover, Social Opinion Mining is carried out to understand opinions as expressed by humans which can help in the advancement of several real-world application areas such as Politics (e.g., election prediction), Marketing, Advertising, and Sales (e.g., brand/product awareness), Technology (e.g., company perception), and Finance (e.g., stock market prediction); industries such as Hospitality (e.g., restaurant recommendations); and domains such as Sports (e.g., fan sentiment), E-commerce (e.g., product recommendations), and Government (e.g., government-citizen interaction). In this regard, people’s social opinions as expressed through various social sources can be beneficial in several domains, used in several applications, and applied in real-life scenarios. Due to this rapid and intense increase in relevance, Social Opinion Mining has become an important research area with a large impact potential, therefore motivating the direction for this thesis.

Notwithstanding the potential cross-domain impact, the majority of Social Opinion Mining research as yet still focuses on high-resourced languages such as English, and less attention has been given to low-resourced languages such as Maltese, and code-switched languages such as Maltese-English. In fact, Malta’s Strategy and Vision for Artificial Intelligence [Schembri, 2019] is currently investing in the development of Maltese language

resources and tools to counter the threat of “digital extinction” for the Maltese language, which has low technological support available in comparison with other European languages [Rosner et al., 2012].

Opinions expressed by humans in textual form are multidimensional semantic artefacts. People usually portray their sentiments and emotions via opinions. Therefore, one’s personal opinion about a particular topic or entity indicates that certain opinion dimensions, such as emotions and sentiment polarities, are mutually influenced by each other [Troussas et al., 2016]. Therefore, opinions greatly influence people’s behaviour and are central to the majority of human activities e.g., when one wants to make a decision this is usually based on the opinions of other humans [Liu, 2012]. In this regard, multiple social opinion dimensions, such as subjectivity, sentiment polarity, emotion, irony, and sarcasm, contribute to the formulation of a more context-rich and fine-grained representation of the real-life complexity of human opinions.

The above discussion is the main motivation behind the exploration of this in-depth research in the Social Opinion Mining research area. The research of this thesis focuses on multidimensional and multilingual Social Opinion Mining for fine-grained opinion search and summarisation at an aspect-based level and has been applied for a real-world use case in the Government domain, namely the Malta Government Budget. Multidimensional opinion summarisation provides a nuanced voice to the general public i.e., citizens and residents of Malta, on their opinions about a topic that is related to the annual Malta Government Budget. These opinions are summarised for the Government of Malta in terms of different social opinion dimensions, namely *subjectivity*, *sentiment polarity*, *emotion*, *sarcasm*, and *irony*, to help in any decision-making and/or policy-making processes.

1.3 Research Questions

Following the motivation discussed in Section 1.2, the **main research question** to be answered in this thesis is defined as follows:

How do multiple Social Opinion Mining dimensions enable a better understanding of the complexity of an aspect-level social opinion?

The main research question of this thesis is aimed towards exploring how the use of multidimensional Social Opinion Mining affects fine-grained opinion search and summarisation at an aspect-based level, and whether this multidimensional approach results in a more accurate social opinion with regard to reflecting the real-life complexity than one based on a single dimension (e.g., sentiment polarity only). It is important to note that there is lack of literature that explores Social Opinion Mining based on the multiple dimensions mentioned, since the most common form of Social Opinion Mining is based on subjectivity and/or sentiment polarity, which are referred to as subjectivity detection and sentiment analysis, respectively. Given that this research question is quite generic, more specific research questions have been defined to better direct the research of this thesis. The contributions presented within this thesis are also defined for each question.

The first aim of this thesis is to have a better understanding of the current Opinion Mining solutions with a specific focus on research that makes use of social data. This focus is mainly due to the fact that this research area has evolved due to the popularity of social media usage and its increase in popularity for carrying out certain activities. Hence, our first research question is the following:

Research Question 1: *What are the existing Opinion Mining approaches that make use of social data and what are the different dimensions of Social Opinion Mining used?*

The main objective of this research question is to identify, analyse, and evaluate existing Opinion Mining approaches that make use of social data in terms of social media platforms, techniques, social datasets, language, modality, tools and technologies, NLP tasks, and other aspects. The different dimensions of Social Opinion Mining, such as subjectivity, sentiment polarity, and emotion, are also identified from the analysed Opinion Mining solutions together with their main application areas. This analysis, carried out through an extensive and comprehensive systematic review, will provide us with a clear direction on the social opinion dimensions that are under-represented in literature, whether existing work caters for multiple dimensions such as sentiment polarity and emotion, and if current approaches accurately portray the real-life complexity of human opinions. The

main challenges faced by the existing Opinion Mining approaches are also identified.

All of this information will identify the current literature gaps and provide a pathway for future research directions within the Social Opinion Mining field of study. This leads towards the next research question which focuses on the application areas of this field of study:

Research Question 2: *How does Social Opinion Mining influence multiple application areas?*

The majority of the current Social Opinion Mining literature targets a particular real-world application area, with the following being the most popular: Politics, Marketing, Advertising, Sales, Technology, Finance, Film, and Healthcare. The main objective of this research question is to determine whether Social Opinion Mining can influence multiple application areas, specifically Politics, Finance, and Government. In this regard, we generate three datasets spanning the application areas of: Politics (targeting Referendums), Finance (targeting Stock Trading), and Socio-Economic and Government (targeting Government Budgets).

Even though a number of social datasets exist for conducting Social Opinion Mining research in multiple application areas, current literature shows that there is still a lack of datasets that cater for multiple social opinion dimensions and ones that target multilingual data. In fact, most of the social datasets available for Opinion Mining purposes are either in English or Chinese and annotated for sentiment polarity only. Moreover, another research gap identified is related to classification models for certain social opinion dimensions especially ones using code-switched, low-resourced, and/or multilingual social data. In this regard, the next research question targets the use of multidimensional and multilingual social opinion data for the development of social opinion classification models:

Research Question 3: *How can multidimensional and multilingual social opinion data be used for the development of classification models?*

The majority of available datasets applicable for Social Opinion Mining purposes are monolingual e.g., English, and they usually target one social opinion dimension e.g., sentiment polarity. The generation of new social datasets poses several challenges since the annotation process is costly in terms of resources (time, funding, and personnel), requires domain knowledge, and expertise. All of these challenges are a major cause behind the low availability of such datasets.

To answer this research question we generate a social dataset in the Socio-Economic and Government domains, specifically Malta’s annual Government Budget, that is made up of multilingual data gathered from newswires and social networking services, which we annotated for multiple social opinion dimensions, namely subjectivity, sentiment polarity, emotion, irony, and sarcasm. Moreover, we explore how this multidimensional and multilingual social dataset can be used for classification purposes. In this respect, this dataset is validated through the development of different classification models for each social opinion dimension, namely subjectivity, sentiment polarity, emotion, irony, and sarcasm, and for different language levels, namely monolingual (English), code-switched (Maltese, Maltese-English), and multilingual (English, Maltese, Maltese-English).

The multidimensional social opinion dataset is valuable for multiple research applications. In fact, the next research question explores how the multidimensional social opinion dataset can be used for fine-grained aspect-based opinion search and summarisation:

Research Question 4: *How are fine-grained opinion search and summarisation affected through the use of a multidimensional Social Opinion Mining approach at an aspect-based level?*

Through this research question we explore whether a multidimensional social opinion provides end-users, such as policy-makers and decision-takers, with a more context-rich and fine-grained representation of the real-life complexity of human opinions, and whether it can result in better informed decisions about a particular domain-specific entity or aspect. Existing opinion summarisation literature focuses on the sentiment polarity dimension only, whereas the proposed approach takes into consideration five social opinion

dimensions, namely subjectivity, sentiment polarity, emotion, irony, and sarcasm.

In this respect, a human evaluation is carried out to determine whether fine-grained opinion search and summarisation at an aspect-based and multidimensional level can be used for a specific purpose, in our case for policy formulation, policy-making, decision-making, and decision-taking, within the context of the Government domain. This human evaluation consists of an extrinsic evaluation, usability evaluation, and a subject matter evaluation, with the former evaluation being based on existing quality measures that measure for usefulness and informativeness. This evaluation shall indicate whether a more fine-grained aspect-level opinion summary based on the combined dimensions of subjectivity, sentiment polarity, emotion, and sarcasm or irony is more informative and more useful than an opinion summary based on sentiment polarity only.

Social Opinion Mining can help in the advancement of several real-world application areas such as Politics (e.g., election prediction), Marketing, Advertising, and Sales (e.g., brand/product awareness), Technology (e.g., company perception), and Finance (e.g., stock market prediction); industries such as Hospitality (e.g., restaurant recommendations); and domains such as Sports (e.g., fan sentiment), E-commerce (e.g., product recommendations), and Government (e.g., government-citizen interaction). For the purposes of this thesis, Social Opinion Mining has been applied in the Government domain which leads us to the last research question:

Research Question 5: *How can multidimensional Social Opinion Mining be applied to the Government domain?*

The proposed multidimensional Social Opinion Mining model has been validated using a real-world use case within the Government domain, specifically the Malta Government Budget. This use case was chosen since it is a different application area than the ones Social Opinion Mining is traditionally applied to, such as Finance, Marketing, Advertising, Sales, and Politics. Moreover, this application area highlights the potential of utilising Social Opinion Mining for multiple domains which are catered for within the Malta Government Budget, such as Taxation, Social Security, Health, and Transport. The research in

this thesis evaluates whether multidimensional Social Opinion Mining at an aspect-based level affects fine-grained opinion search and summarisation. This is carried out using an Opinion Search and Summarisation system that leverages the novel multidimensional and multilingual social opinion dataset discussed in **Research Question 3**. In this respect, the fine-grained aspect-level opinion summaries are based on the use case being considered.

Moreover, the same dataset is used to build baseline monolingual (English), code-switched (Maltese, Maltese-English), and multilingual (English, Maltese, Maltese-English) classification models for subjectivity detection, sentiment analysis, emotion analysis, sarcasm detection, and irony detection, which can be used by the Government of Malta for policy formulation, policy-making, decision-making, and decision-taking. The main objective is to show that opinion summaries provide a nuanced-voice to the general public i.e., citizens and residents of Malta, on their opinions about a topic that is related to the annual Malta Government Budget.

It is important to note that this multidimensional Social Opinion Mining research can be applied to similar Government initiatives and to other application areas, such as Politics, Finance, Marketing, Advertising, Sales and Education.

1.4 Contributions and Publications

The following is a list of contributions presented in this thesis based on the research questions defined in Section 1.3.

1. A systematic analysis of existing Social Opinion Mining solutions

A systematic review was carried out on the evolving research area of Social Opinion Mining based on **Research Question 1**. Throughout this review, a large number of existing Social Opinion Mining solutions were analysed in-depth based on the use of social media platforms, techniques, social datasets, language, modality, tools and technologies, NLP tasks including aspect-based Social Opinion Mining, and other aspects identified within the published literature. Moreover, the different Social Opinion Mining dimensions, namely subjectivity, sentiment polarity, emotion, affect, irony, and sarcasm are identified from the analysed Opinion Mining solutions,

together with their main application areas. The latest research developments, advancements, and current literature gaps within the Social Opinion Mining research area are also identified. As a contribution this was crucial since it forms the basis of this thesis and identified which research gaps to further explore in context of this thesis. This research is discussed in detail in Chapter 2. This contribution is published in the following publication: [Cortis and Davis, 2021b].

2. **Three social datasets spanning the Politics, Finance, and Socio-Economic and Government application areas**

The following three social datasets were generated: Politics targeting the Referendums, Finance targeting Stock Trading, and Socio-Economic and Government targeting Government Budgets. In terms of Social Opinion Mining dimensions, the Political dataset focuses on sentiment in terms of polarity and strength, the Financial dataset focuses on sentiment in terms of score, and the Socio-Economic and Government dataset focuses on subjectivity, sentiment polarity, emotion, irony, and sarcasm. In response to **Research Question 2**, these datasets were generated to determine how Social Opinion Mining can influence multiple application areas. Whilst all three datasets cater for one common social opinion dimension i.e., sentiment polarity, since it is the most explored dimension within the Opinion Mining research area, the Socio-Economic and Government dataset also focuses on additional opinion dimensions. This was done in order to cater for the lack of multidimensional datasets for Social Opinion Mining purposes. However, given that all the three datasets cater for one common social opinion dimension in terms of sentiment polarity, the datasets can still be evaluated in terms of the influence of Social Opinion Mining in multiple application areas. The generation of the above-mentioned three social datasets enabled us to understand the importance of multiple social opinion dimensions that target multilingual data, which is currently largely unexplored in the Social Opinion Mining research area. For this reason, the research direction in this thesis proceeded towards contributing in this domain. This research is discussed in detail in Chapter 3. This contribution is published in the following publications: [Hürlimann et al., 2016, Cortis et al., 2017, Cortis and Davis, 2019, Cortis and Davis, 2021a].

3. **A Dataset of Multidimensional and Multilingual Social Opinions for Malta’s Annual Government Budget**

This social opinion dataset in the Socio-Economic and Government domains, comprises social data on Malta’s annual Government Budgets of 2018, 2019, and 2020. Overall it contains over 6,000 online posts of user-generated content in English, Maltese, and Maltese-English, gathered from newswires and social networking services. This has been annotated for multiple opinion dimensions in subjectivity, sentiment polarity, emotion, irony, and sarcasm, and in terms of negation, topic, and language. In response to **Research Question 3**, this dataset is a valuable resource for developing Social Opinion Mining tools and Language Technologies, and can be used as a baseline for assessing the state-of-the-art and for developing new advanced analytical methods for Social Opinion Mining. Moreover, in response to **Research Question 4**, it can be used by the Government of Malta for policy formulation, policy-making, decision-making, and decision-taking. This research is discussed in detail in Section 3.4. This contribution is published in the following publications: [Cortis and Davis, 2019, Cortis and Davis, 2021a].

4. **Baseline classification models for subjectivity, sentiment polarity, emotion, irony, and sarcasm**

Several baseline classification models using traditional machine learning techniques and state-of-the-art deep neural network models were developed. In response to **Research Questions 3 and 5**, the novel multidimensional and multilingual social opinion dataset in the Socio-Economic and Government domains, specifically Malta’s annual Government Budget, is used for the development of the said models. This social opinion dataset is validated through the development of different classification models for each social opinion dimension, namely subjectivity, sentiment polarity, emotion, irony, and sarcasm, and for different language levels, namely monolingual for English –a high-resourced language; code-switched for Maltese –a low-resourced language, and Maltese-English –a code-switched language; and multilingual for the English, Maltese, Maltese-English languages. This research is discussed in detail in Chapter 4. This contribution is published in the following publications: [Cortis

et al., 2021, Cortis and Davis, 2022].

5. A Fine-grained Opinion Search and Summarisation system

An Opinion Search and Summarisation system was developed in order to determine whether multidimensional Social Opinion Mining at an aspect-based level provides benefits for fine-grained opinion search and summarisation. This system is pre-loaded with the novel multidimensional and multilingual social opinion dataset previously discussed. Therefore, any new annotated data (manually, semi-automatically, and/or automatically) shall need to be primarily pre-loaded in this system prior to be used for opinion search and summarisation purposes. In response to **Research Question 5**, this system was evaluated for a real-world use case in the Government domain, namely the annual Malta Government Budget. The goal was to determine whether the aspect-based multidimensional approach affects decision-making and/or policy-making purposes. A human evaluation comprising an extrinsic evaluation, a usability evaluation, and a subject matter expert evaluation was carried out to determine whether the fine-grained opinion search and summarisation at an aspect-based and multidimensional level provides added-value for a specific purpose, in our case for policy formulation, policy-making, decision-making, and decision-taking. The extrinsic evaluation was based on four quality measures, improving upon existing state-of-the-art with the introduction of an additional measure and an adaptation of an existing one to be representative of the social data used within our study. In response to **Research Question 4**, the results obtained indicate that a more fine-grained aspect-level opinion summary based on the combined dimensions of subjectivity, sentiment polarity, emotion, and sarcasm or irony is more informative and more useful than an opinion summary based on sentiment polarity only. The proposed multidimensional opinion search and summarisation approach contributes towards the formulation of a more context-rich and fine-grained representation of the real-life complexity of human opinions. This highlights our contribution over current state-of-the-art opinion summarisation research which focuses on sentiment polarity only. This research is discussed in detail in Chapter 5. The aforementioned contribution has been submitted for review as the following publication: [Cortis and

Davis, 2023].

The work discussed in this thesis is partially covered or is derived by the following publications:

1. **Keith Cortis**, Brian Davis. “Over a decade of social opinion mining: a systematic review.” *Artificial intelligence review* 54, no. 7 (2021): 4873-4965.
2. Manuela Hürlimann, Brian Davis, **Keith Cortis**, André Freitas, Siegfried Handschuh, Sergio Fernández. “A twitter sentiment gold standard for the brexit referendum.” In *Proceedings of the 12th international conference on semantic systems*, pp. 193-196. 2016.
3. **Keith Cortis**, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, Brian Davis. “Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news.” *Association for Computational Linguistics (ACL)*, 2017.
4. **Keith Cortis**, Brian Davis. “A social opinion gold standard for the Malta government budget 2018.” In *Proceedings of the 5th workshop on noisy user-generated text (W-NUT 2019)*, pp. 364-369. 2019.
5. **Keith Cortis**, Brian Davis. “A dataset of multidimensional and multilingual social opinions for malta’s annual government budget.” In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, pp. 971-981. 2021.
6. **Keith Cortis**, Kanishk Verma, Brian Davis. “Fine-tuning Neural Language Models for Multidimensional Opinion Mining of English-Maltese Social Data.” In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pp. 309-314. 2021.
7. **Keith Cortis**, Brian Davis. “Baseline English-Maltese Classification Models for Subjectivity Detection, Sentiment Analysis, Emotion Analysis, Sarcasm Detection, and Irony Detection.” In *Proceedings of the LREC 2022 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages (SIGUL 2022)*, pp. 161-168. 2022.

8. **Keith Cortis**, Brian Davis. “Social Opinion Search and Summarisation: A Multidimensional Approach.” (*under review*). 2023.

1.5 Thesis Structure

This thesis consists of six chapters. Chapter 1 provides an introduction in terms of problem specification, motivation, research questions, contributions, and publications. In Chapter 2, we present a literature analysis based on a comprehensive systematic review on the research area of Social Opinion Mining. This chapter answers Research Question 1. Chapter 3 presents three different Social Opinion datasets that were generated for the purposes of this thesis. This chapter answers Research Question 2, part of Research Question 3, and part of Research Question 5. Chapter 4 discusses multidimensional opinion classification baseline models built on a novel multidimensional and multilingual social opinion dataset in the Socio-Economic and Government domains. This chapter answers Research Question 3 and Research Question 5. Chapter 5 discusses multidimensional opinion search and summarisation for a real-world use case in the Government domain, namely the annual Malta Government Budget. This chapter answers Research Question 4 and Research Question 5. Finally, Chapter 6 concludes this thesis and offers future research directions.

Chapter 2

Opinion Mining from Social Data: A Literature Analysis

2.1 Introduction

The research area of Opinion Mining is evolving due to the rise of social media platforms, where several work already exists on the analysis of sentiment polarity. Moreover, researchers can gauge widespread opinions from user-generated content and better model and understand human beliefs and their behaviour. Opinion Mining is regarded as a challenging NLP problem, in particular for social data obtained from social media platforms, such as Twitter, and also for transcribed text. Standard linguistic processing tools were built and developed on newswires and review-related data due to such data following more strict grammar rules. These differences should be taken in consideration when performing any kind of analysis [Balazs and Velásquez, 2016]. Therefore, social data is difficult to analyse due to the short length in text, the non-standard abbreviations used, the high sparse representation of terms and difficulties in finding out the synonyms and any other relations between terms, emoticons, and hashtags used, lack of punctuations, use of informal text, slang, non-standard shortcuts, and word concatenations. Hence, typical NLP solutions are not likely to work well for Opinion Mining.

Opinion Mining –presently a very popular field of study– is defined by Liu and Zhang as “the computational study of people’s opinions, appraisals, attitudes, and emotions

toward entities, individuals, issues, events, topics and their attributes” [Liu and Zhang, 2012]. **Social** is defined by the Merriam-Webster Online dictionary¹ as “of or relating to human society, the interaction of the individual and the group, or the welfare of human beings as members of society”. In light of this, we define **Social Opinion Mining** as “the study of user-generated content by a selective portion of society be it an individual or group, specifically those who express their opinion about a particular entity, individual, issue, event and/or topic via social media interaction”. Therefore, the research area of Social Opinion Mining is tasked with the identification of several opinion dimensions, such as sentiment polarity, emotion, sarcasm, irony, and mood, from social data which is represented in structured, semi-structured, and/or unstructured data formats.

The structure of this chapter is as follows: Section 2.2 presents the context and research method of a systematic review carried out on the research area of Social Opinion Mining, followed by Section 2.3 which provides a thorough analysis of this research area. This is followed by Section 2.4 which discusses the different dimensions of Social Opinion Mining, and Section 2.5 which presents the different application areas where this research area is used. Lastly, Section 2.6 discusses the current literature gaps and future research directions within the research area of Social Opinion Mining.

Sections 2.2-2.6 of this chapter are based on the following publication:

- **Keith Cortis**, Brian Davis. ”Over a decade of social opinion mining: a systematic review.” *Artificial intelligence review* 54.7 (2021): 4873-4965.

This systematic review was carried out on the evolving research area of Social Opinion Mining, which study forms the basis of this thesis.

2.2 Context and Research Method of Systematic Review

This section discusses the context behind a systematic review carried out on the research area of Social Opinion Mining and the research method used in this regard. It is worth noting that no systematic review within this newly defined domain exists even though there are several good Opinion Mining survey papers, such as [Liu and Zhang, 2012, Tsytsarau

¹<http://www.merriam-webster.com/>

and Palpanas, 2012, Medhat et al., 2014, Ravi and Ravi, 2015]. The research paper by Bukhari et al. [Bukhari et al., 2016] is closest to a systematic review in this domain, whereby the authors performed a search over the ScienceDirect and SpringerLink electronic libraries for the “sentiment analysis”, “sentiment analysis models”, “sentiment analysis of microblogs” terms. As a result, we felt that the Social Opinion Mining domain well and truly deserves a thorough systematic review that captures all of the relevant research conducted over the last decade and beyond. This systematic review also identifies the current literature gaps within this popular and constantly evolving research domain.

A systematic methodology was used to gather the relevant literature on Social Opinion Mining. This empirical research process was based on the guidelines and procedures proposed by [Kitchenham, 2004, Brereton et al., 2007, Dyba et al., 2007, Attard et al., 2015] which were focused on the software engineering domain. The procedure undertaken was structured as follows:

1. Specification of research questions;
2. Generation of search strategy which includes the identification of electronic sources (libraries) and selection of relevant search terms;
3. Application of the relevant search;
4. Choice of primary studies via the utilisation of inclusion and exclusion criteria on the obtained results;
5. Extraction of required data from primary studies; and
6. Synthesis of data.

2.2.1 Research Questions

A systematic literature review is usually characterised by an appropriate generic “research question, topic area, or phenomenon of interest” [Kitchenham, 2004]. This question can be expanded into a set of sub-questions that are more clearly defined, whereby all available research relevant to these sub-questions are identified, evaluated, and interpreted.

As part of the overall objectives of this thesis, the goal of this systematic review is to identify, analyse, and evaluate current Opinion Mining solutions that make use of social data (data extracted from social media platforms). In light of this, the following generic research question is defined:

- **What are the existing Opinion Mining approaches which make use of user-generated content obtained from social media platforms?**

The following are specific sub-questions that the generic question above can be subdivided into:

1. What are the existing approaches that make use of social data for Opinion Mining and how can they be classified²?
2. What are the different dimensions/types of Social Opinion Mining?
3. What are the challenges faced when performing Opinion Mining on social data?
4. What techniques, datasets, tools/technologies, and resources are used in the current solutions?
5. What are the application areas of Social Opinion Mining?

2.2.2 Search Strategy

The search strategy for this systematic review is primarily directed via the use of published papers which consist of journals, conference/workshop proceedings, or technical reports. The following electronic libraries were identified for use, due to their wide coverage of relevant publications within our domain: ACM Digital Library³, IEEE Xplore Digital Library⁴, ScienceDirect⁵, and SpringerLink⁶.

The first three electronic libraries listed were used by three out of the four systematic reviews that our research process was based on (and which made use of a digital source),

²Classification in this context refers to the dimension of Opinion Mining being conducted, such as subjectivity detection, sentiment analysis, emotion analysis, sarcasm detection, and irony detection.

³<https://dl.acm.org/>

⁴<http://ieeexplore.ieee.org/>

⁵<https://www.sciencedirect.com/>

⁶<https://link.springer.com/>

whereas SpringerLink is one of the most popular sources for publishing work in this domain (as will be seen in Section 2.2.4 below). Moreover, three other electronic libraries were considered for use, two –Web of Science⁷ and Ei Compendex⁸– which the host university did not have access to and Google Scholar⁹ which was not included since content is obtained from the electronic libraries listed above (and more), thus making the process redundant.

The relevant search terms were identified for answering the research questions defined in Section 2.2.1. In addition, these questions were also used to perform some trial searches before the following list of relevant search terms was determined:

- “Social opinion mining”;
- “Social sentiment analysis”;
- “Opinion mining social media”;
- “Sentiment analysis social media”;
- “Microblog opinion mining”;
- “Microblog sentiment analysis”;
- “Social network sentiment”;
- “Social network opinion”;
- “Social data sentiment analysis”;
- “Social data opinion mining”;
- “Twitter sentiment analysis”;
- “Twitter opinion mining”; and
- “Social data analysis”.

The following are important justifications behind the search terms selected above:

⁷<https://webofknowledge.com/>

⁸<https://www.elsevier.com/solutions/engineering-village/content/compendex>

⁹<http://scholar.google.com/>

- “Opinion Mining” and “Sentiment Analysis”: are both included due to the fact that these key terms are used interchangeably to denote the same field of study [Pang and Lee, 2008, Cambria et al., 2013], even though their origins differ and hence do not refer to the same concept [Serrano-Guerrero et al., 2015]; and
- “microblog”, “social network”, and “Twitter”: the majority of the Opinion Mining and/or Sentiment Analysis research and development efforts target these two kinds of social media platforms, in particular the Twitter microblogging service.

2.2.3 Search Application

The “OR” Boolean operator was chosen to formulate the search string. The search terms were all linked using this operator in a composite manner (e.g., “social network opinion”), making the search query simple and easy to use across multiple electronic libraries. Therefore, a publication only had to include any one of the search terms to be retrieved [Attard et al., 2015]. In addition, this operator is more suitable for the defined search terms given that this study is not a general one e.g., about Opinion Mining in general, but is focused about Opinion Mining in a social context. Construction of the correct search string (and terms) is very important, since this eliminates noise (i.e., false positives) as much as possible and at the same time still retrieves potential relevant publication which increases recall.

Several other factors had to be taken in consideration during the application of search terms on the electronic libraries. The following is a list of factors relevant to our study, identified in [Brereton et al., 2007] and verified during our search application process:

- Electronic library search engines have different underlying models, thus not always provide required support for systematic searching;
- Same set of search terms cannot be used for multiple engines e.g., complex logical combination not supported by the ACM Digital Library but is by the IEEE Xplore Digital Library;
- Boolean search string is dependent on the order of terms, independent of brackets;

- Inconsistencies in the order or relevance in search results (e.g., IEEE Xplore Digital Library results are sorted in order of relevance); and
- Certain electronic libraries treat multiple words as a Boolean term and look for instances of all the words together (e.g., “social opinion mining”). In this case, the use of the “AND” Boolean operator (e.g., “social AND opinion AND mining”) looks for all of the words not necessary together.

In this case, it was very important to select a search strategy that is more appropriate to the research question of the systematic review, which could be applied to the selected electronic libraries.

When applying the relevant search on top of the search strategy defined in Section 2.2.2, another important element was to identify appropriate metadata fields upon which the search string can be executed. Table 2.1 presents the ones applied in this review.

Metadata field	ACM	IEEE Xplore	ScienceDirect	SpringerLink
title	✓	✓	✓	✓
abstract	✓	✓	✓	✓
keywords	✓	✓	✓	

Table 2.1: Metadata fields used in search application

Applying the search on the title metadata field alone would result in several missed and/or incorrect results. Therefore, using the abstract and/or keywords in the search is very important to reduce the number of irrelevant results. In addition, this ensures that significant publications that lack any of the relevant search terms within their title are returned.

A separate search method was applied for each electronic library, since they all offer different functionalities and have different underlying models. Each method is detailed below:

- ACM: Separate searches for each metadata field were conducted and results were merged (duplicates removed). Reason being that the metadata field search functionality “ANDs” all metadata fields, whereas manual edition of the search query does not work well when amended.

- IEEE: Separate searches for each metadata field were conducted and results were merged (duplicates removed).
- ScienceDirect: One search that takes in consideration all the chosen metadata fields.
- SpringerLink: By entering a search term or phrase, a search is conducted over the title, abstract, and full-text (including authors, affiliations, and references) of every article and book chapter. This was noted in the large amount of returned papers which resulted in a high amount of false positives.

2.2.4 Study Selection

A manual study selection was performed on the primary studies obtained from the search application defined in Section 2.2.3. This was required to eliminate any studies that might be irrelevant even though the search terms appear in either of the metadata fields defined in Table 2.1 above. Therefore, inclusion and exclusion criteria (listed below) were defined.

Published papers that meet any of the following inclusion criteria were chosen as primary studies:

- I1. A study that targeted at least one social networking service and/or utilised a social dataset besides other social media services, such as blogs, chats, and wikis. Please note that only work performed on social data from social networking services is taken in consideration for the purposes of this review;
- I2. A study published from the year 2007 onwards. This year was chosen, since the mid-2000s saw the evolution of several social networking services, in particular Facebook's growth (2007), which currently contains the highest monthly active users; and
- I3. A study published in the English language.

Published papers that satisfy any of the exclusion criteria from the following list, were removed from the systematic review:

- E1. A study published before 2007;

- E2. A study that does not focus on performing any sort of Opinion Mining on social media services, even though it mentions some of the search terms;
- E3. A study that focuses on Opinion Mining or Sentiment Analysis in general i.e., no reference in a social context;
- E4. A study that is only focused on social data sources obtained from online forums, communities, blogs, chats, social news websites (e.g., Slashdot¹⁰), and review websites (e.g., IMDb¹¹); and
- E5. A study that consists of either a paper’s front cover and/or title page i.e., a result that does not contain the study itself.

Selection of the primary studies for this systematic review was carried out in 2019. Therefore, studies indexed or published from 2019 onwards, are not included in this review.

Primary studies	ACM	IEEE Xplore	ScienceDirect	SpringerLink
Search application	106	242	57	456
False positives	39	83	17	262
Study selection	67	159	40	194
No full paper access	0	0	5	4
Full paper access	67	159	35	190
Total	451			

Table 2.2: Primary studies selection procedure from the electronic libraries

Table 2.2 shows the results for each electronic library at each step of the procedure used for selecting the final set of primary studies. The results included one proceedings, which was resolved by including all the published papers within the track relevant to this study, since the other papers were not relevant thus not included in the initial results. The search application phase resulted in a total of 861 published papers. False positives, which consisted of duplicate papers and papers that met any of the exclusion criteria were removed. This was done through a manual study selection which was performed on all the metadata fields considered i.e., the title, abstract, and keywords. In cases where we were still unclear of whether a published paper is valid or not, we went through the full text. After removing a total of 401 false positives, this study selection operation left us

¹⁰<https://slashdot.org/>

¹¹<https://www.imdb.com/>

with 460 published papers. Out of the final study selection published papers, we did not have full access to 9 published papers, thus reducing the total primary studies to 451.

In addition to the primary studies selected from the electronic libraries, a set of relevant studies were added –34 published papers (excluding survey papers)– for completeness sake which were either published in reputable venues within the Opinion Mining community or were highly cited. Therefore, the final set of primary studies totalled 485 published papers.

2.2.5 Extraction of data

The main objective of this review was to conduct a systematic analysis of the current literature in the research area of Social Opinion Mining. Each published paper in this review was analysed in terms of the following information/parameters: social media platforms, techniques and approaches, social datasets, language, modality, tools and technologies, (other) NLP tasks, application areas, and Opinion Mining dimensions. It is important to note that this information was manually extracted from each published paper. The overall statistics about the relevant primary studies that resulted from the study selection phase of this systematic review are discussed below.

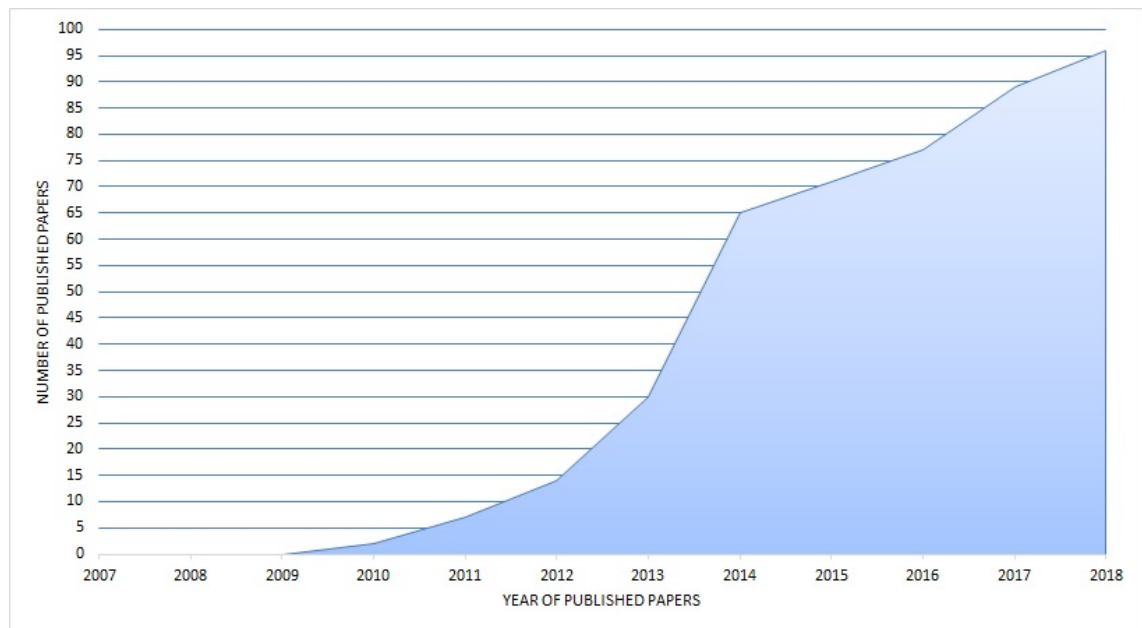


Figure 2.1: Primary Studies by Year

Figure 2.1 shows that the first three years of this evaluation period, i.e., 2007-2009, did not return any relevant literature. It is important to note that 2006 and 2007 was the period when Opinion Mining emerged in Web applications and weblogs within multiple domains, such as Politics and Marketing [Pang and Lee, 2008]. However, 2010 –which coincides with the introduction of various social media platforms and a major increase in Facebook and Twitter usage¹²– resulted in the first literature, with figures increasing in the following years. It is interesting to note that the final year in evaluation, which is 2018, contains literature that was published or indexed until the 31st December 2018. From the twelve full years evaluated, 2018 produced the highest number of relevant literature items. This shows the importance of Opinion Mining on social data, and therefore the continuous increase in social media usage and popularity, in particular social networking services.

The additional set of studies included in this systematic review, were published in the period between the years of 2009 and 2014. These ranged from various publishers, namely the four selected for this study (ACM, IEEE Xplore, ScienDirect, and SpringerLink) and other popular ones, such as Association for the Advancement of Artificial Intelligence (AAAI)¹³, Association for Computational Linguistics (ACL)¹⁴, and Wiley Online Library¹⁵.

2.2.6 Synthesis of data

The data synthesis highlighted in this detailed analysis is based on the extracted data mentioned in Section 2.2.5 above. This in-depth analysis is focused on the social media platforms, techniques, social datasets, language, modality, tools and technologies, NLP tasks including aspect-based Social Opinion Mining, and other aspects used across the published papers.

While the systematic review covered studies until 2018, some recent literature in terms of developments and advancements from 2019 until 2021 were also researched and analysed. This was necessary due to the fast research development in Social Opinion Mining which has kept evolving at an incredibly fast rate, thus reiterating its validity and popu-

¹²<https://www.techinasia.com/social-media-timeline-2010>

¹³<https://aaai.org/Library/library.php>

¹⁴<http://aclweb.org/anthology/>

¹⁵<https://onlinelibrary.wiley.com/>

larity as a research area.

2.3 Social Opinion Mining Analysis

This section discusses the relevant literature for the research area of Social Opinion Mining. Due to word limitations of this thesis, the sections below provide an aggregation of in-depth analysis carried out as part of the systematic review.

2.3.1 Opinion Mining vs. Social Opinion Mining

In 2008, Pang and Lee had already identified the relevance between the field of “social media monitoring and analysis” and the body of work reviewed in [Pang and Lee, 2008], which deals with the computational treatment of opinion, sentiment, and subjectivity in text. This work is nowadays known as **Opinion Mining**, **Sentiment Analysis**, and/or **Subjectivity Analysis** [Pang and Lee, 2008]. Other phrases, such as **Review Mining** and **Appraisal Extraction** have also been used in the same context, whereas some connections have been found to **Affective Computing** (where one of its goals is to enable computers in recognising and expressing emotions) [Pang and Lee, 2008]. Merriam-Webster’s Online Dictionary defines that the terms¹⁶ “opinion”, “view”, “belief”, “conviction”, “persuasion”, and “sentiment” mean a judgement one holds as true. This shows that the distinctions in common usage between these terms can be quite subtle. In light of this, three main three research areas –Opinion Mining, Sentiment Analysis, and Subjectivity Analysis– are all related and use multiple techniques taken from NLP, information retrieval, and structured and unstructured data mining [Ravi and Ravi, 2015]. However, even though these three concepts are broadly used as synonyms, thus used interchangeably, it is worth noting that their origins differ. Some authors also consider that each concept presents a different understanding [Serrano-Guerrero et al., 2015] and also have different notions [Tsytsarau and Palpanas, 2012]. We are in agreement with this, hence why we felt that a new terminology is required to properly specify what Social Opinion Mining means, as defined in Section 2.1.

According to Cambria et al., sentiment analysis can be considered as a very restricted

¹⁶<http://www.merriam-webster.com/dictionary/opinion>

NLP problem, where the polarity (negative/positive) of each sentence and/or target entities or topics needs to be understood [Cambria et al., 2013]. On the other hand, Liu discusses that “opinions are usually subjective expressions that describe people’s sentiments, appraisals or feelings toward entities, events and their properties” [Liu, 2010]. He further identifies two sub-topics of sentiment and subjectivity analysis, namely sentiment classification (or document-level sentiment classification) and subjectivity classification. Social Opinion Mining requires such classification methods to determine an opinion dimension, such as objectivity/subjectivity and sentiment polarity. For example, subjectivity classification is required to classify whether user-generated content, such as a product review, is objective or subjective, whereas sentiment classification is performed on subjective content to find the sentiment polarity (positive/negative) as expressed by the author of the opinionated text. In cases where the user-generated content is made up of multiple sentences, sentence-level classification needs to be performed to determine the respective opinion dimension. In addition, sentence-level classification is not suitable for compound sentences, i.e., a sentence that expresses more than one opinion. For such cases, aspect-based Opinion Mining needs to be performed.

2.3.2 Issues and Challenges

The authors in [Pang and Lee, 2008], had already identified that the writings of Web users can be very challenging in their own way due to numerous factors, such as the quality of written text, discourse structure, and the order in which different opinions are presented. The effects of the latter factor can result in a completely opposite overall sentiment polarity, where the order effects can completely overwhelm the frequency effects. This is not the case in traditional text classification, where if a document refers to the term “car” in a frequent manner, the document will probably somewhat be related to cars. Therefore, order dependence manifests itself in a more fine-grained level of analysis.

In [Liu, 2010], the author mentions that complete sentences (for reviews) are more complex than short phrases and contain a large amount of noise, thus making it more difficult to extract features for feature-based sentiment analysis. Even though we agree that with more text comes a higher probability of spelling mistakes, etc., we tend to

disagree that shorter text, such as microposts, contain less noise.

The process of mining user-generated content posted on the Web is very intricate and challenging due to the nature of short textual content limit (e.g., tweets allowed up to 280 characters which was updated from 140 characters in October 2017), which at times forces a user to resort in using short words, such as acronyms and slang, to make a statement. These often lead to further issues in the text, such as misspellings, incomplete content, jargon, incorrect acronyms and/or abbreviations, emoticons, and content misinterpretation [Cortis, 2013]. Other noteworthy challenges include swear words, irony, sarcasm, negatives, conditional statements, grammatical mistakes, use of multiple languages, incorrect language syntax, syntactically inconsistent words, and different discourse structures. In the case of discourse structures within social data, discourse interpretation in terms of discourse relations, argumentation, and threads in social media platform conversations (e.g., Twitter) are important to cater for computational linguistics issues and to enhance certain NLP tasks, such as sentiment analysis [Benamara et al., 2018]. When informal language is used in the user-generated content, the grammar and lexicon varies from the standard language normally used [Dashtipour et al., 2016]. Moreover, user-generated text exhibits more language variation due to it being less grammatical than longer posts, where the aforementioned use of emoticons, abbreviations together with hashtags, and inconsistent capitalisation, can form an important part of the meaning [Maynard et al., 2012]. In the latter paper the authors also point out that microposts are in some sense the most challenging type of text for text mining tools especially for Opinion Mining, since they do not contain a lot of contextual information and assume much implicit knowledge. Another issue is ambiguity, since microposts such as tweets, do not follow a conversation thread. Therefore, this isolation from other tweets makes it more difficult to make use of coreference information, unlike in blog posts and comments. Due to the short textual content, features can also be sparse to find and use, in terms of text representation [Wang et al., 2014a]. In addition, the majority of microposts usually contain information about a single topic due to the length limitation, which is not the case in traditional blogs, where they contain information on more than one topic given that they do not face the same length limitations [Giachanou and Crestani, 2016]. Moreover, if the dominant social opin-

ion dimension (e.g., sentiment polarity) needs to be determined from social data, certain NLP tasks like Word Sense Disambiguation can be used for a better understanding of the word meanings in certain contexts to counter problems, such as lexical ambiguity which can be of a syntactic or semantic nature [Sumanth and Inkpen, 2015]. Stance classification –tasked with defining “an overall position held by a person toward an object, idea or proposition” [Somasundaran and Wiebe, 2010]– carried out on user-generated content such as online news comments is challenging to carry out in cases when certain news is considered as the target to investigate the overall position toward it [Sobhani et al., 2015]. It is important to mention that stance classification differs to sentiment classification, since stance is not always expressed in terms of affective words. Therefore, determining the sentiment polarity within the text is not sufficient, since a stance classification system should detect favourability toward a particular target which may be different from the opinion target [Sobhani et al., 2015].

Big data challenges, such as handling and processing large volumes of streaming data, are also encountered when analysing social data [Bravo-Marquez et al., 2014]. Limited availability of labelled data and dealing with the evolving nature of social streams usually results in the target concept changing which would require the learning models to be constantly updated [Guerra et al., 2014].

In light of the above, social networking services bring several issues and challenges with them and the way in how content is generated by their users. Therefore, several Information Extraction (IE) tasks, such as Named Entity Recognition (NER) and Coreference Resolution, might be required to carry out multidimensional Social Opinion Mining. In fact, several shared evaluation tasks are being organised to try and reach a standard mechanism towards performing IE tasks on noisy text which is very common in user-generated social media content. As already discussed in detail above, such tasks are much harder to solve when they are applied on micro-text like microposts [Ravi and Ravi, 2015]. This problem presents serious challenges on several levels, such as performance. Examples of such tasks are “Named Entity Recognition in Twitter”¹⁷.

In terms of content, social media-based studies present only analysis and results from

¹⁷<http://noisy-text.github.io/2016/ner-shared-task.html>

a selective portion of society, since not everyone uses social media. Moreover, several cross-cultural differences and factors determine the social media usage in each country and hence the results of such studies. For example for the Political domain, these services are used predominantly by young and politically active individuals or by ones with strong political views. This could be easily reflected in the Brexit results, where the majority of younger generation (age 18-44) voted to remain in the European Union as opposed to people over age 45. Such a result falls in line with the latest United Kingdom social media statistics, such as for Twitter, where 72% of the users are between the age of 15-44, whilst for Facebook the most popular age group is 25-34 (26% of users) [Hürlimann et al., 2016]. However, results of similar studies in other cultures and languages might differ due to different use of social words to reflect a general opinion, sentiment polarity, and/or emotion [Lin et al., 2018].

2.3.3 Existing Approaches

All the studies analysed in the systematic review were categorised under the following approaches in terms of techniques used to carry out the Opinion Mining process on social data: Lexicon (Lx), Machine Learning (ML), Deep Learning (DL), Statistical (St), Probabilistic (Pr), Fuzziness (Fz), Rule (Rl), Graph (Gr), Ontology (On), Hybrid (Hy) –a combination of more than one technique, Manual (Mn), and Other (Ot). Table 2.3 provides the yearly statistics for all the respective approaches adopted.

Year	Lx	ML	DL	St	Pr	Fz	Rl	Gr	On	Hy	Mn	Ot
2007	0	0	0	0	0	0	0	0	0	0	0	0
2008	0	0	0	0	0	0	0	0	0	0	0	0
2009	0	1	0	0	0	0	0	0	0	1	0	0
2010	2	2	0	0	0	0	0	0	0	3	0	0
2011	2	3	1	0	0	0	0	0	0	7	1	0
2012	6	5	0	0	0	0	0	1	0	10	0	1
2013	6	14	2	1	1	0	2	0	1	21	0	0
2014	14	20	2	1	3	1	1	0	1	41	0	3
2015	16	15	4	1	1	0	0	1	0	42	0	0
2016	13	21	4	3	0	0	0	0	0	38	2	4
2017	20	22	9	2	1	1	0	0	0	50	2	5
2018	17	18	13	1	0	0	1	2	0	69	1	4
Total	96	121	35	9	6	2	4	4	2	282	6	17

Table 2.3: Approaches used in the studies analysed

The results show that a hybrid approach is the most popular one, with over half of the studies adopting such an approach. This is followed by machine learning and lexicon

techniques, which are usually chosen to perform any form of Opinion Mining. An overview of these results is explained in more detail in the sections below.

2.3.3.1 Lexicon

The majority of the lexicons used were specifically related to opinions and are well known in this domain, whereas the others that were not can still be used for conducting Opinion Mining. The following are the top lexicons based on use:

1. SentiWordNet¹⁸ [Baccianella et al., 2010];
2. Hu and Liu¹⁹ [Hu and Liu, 2004];
3. AFINN²⁰ [Årup Nielsen, 2011] and SentiStrength²¹ [Thelwall et al., 2012];
4. MPQA - Subjectivity²² [Wilson et al., 2005];
5. HowNet Sentiment Analysis Word Library (HowNetSenti)²³;
6. NRC Word-Emotion Association Lexicon (also known as NRC Emotion Lexicon or EmoLex)²⁴ [Mohammad and Turney, 2010, Mohammad and Turney, 2013];
7. WordNet²⁵ [Miller, 1995]; and
8. Wikipedia - list of emoticons²⁶.

In addition to the lexicons mentioned above, some studies used lexicons that they created as part of their work or specifically focused on creating Social Opinion Mining lexicons, such as [Årup Nielsen, 2011] who created the AFINN word list for sentiment analysis in microblogs, [Javed et al., 2014] who built a bilingual sentiment lexicon for English and Roman Urdu, [Santarcangelo et al., 2015] the creators of the first Italian sentiment thesaurus, [Wu et al., 2016] for Chinese sentiment analysis, and [Bandhakavi et al.,

¹⁸<https://github.com/aesuli/SentiWordNet>

¹⁹<https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>

²⁰<https://github.com/fnielsen/afinn>

²¹<http://sentistrength.wlv.ac.uk/>

²²https://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

²³Original dataset points to <http://www.keenage.com/>

²⁴<https://www.saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>

²⁵<https://wordnet.princeton.edu/>

²⁶https://en.wikipedia.org/wiki/List_of_emoticons

2016] for sentiment analysis on Twitter. These lexicons varied from social media focused lexicons, to sentiment and/or emoticon lexicons, and extensions of existing state-of-the-art lexicons, such as [Li et al., 2016] who extended HowNetSenti with words manually collected from the internet, and [Pandarachalil et al., 2015] who built a sentiment lexicon from SenticNet²⁷ [Cambria et al., 2020] and SentiWordNet for slang words and acronyms.

2.3.3.2 Machine Learning

Several supervised and unsupervised algorithms were used by the studies that adopted a machine learning-based approach to perform a form of Social Opinion Mining.

In terms of supervised machine learning algorithms, the Naïve Bayes [Lewis, 1998] and Support Vector Machine [Cortes and Vapnik, 1995] algorithms are clearly the most popular in this domain, especially for text classification. The Logistic Regression [McCullagh, 1984] statistical technique is also widely used in machine learning for binary classification problems. Decision Tree [Quinlan, 1986] learning has also been very much in use, which model uses a Decision Tree for both classification and regression problems. The Maximum Entropy [Jaynes, 1957], a probabilistic classifier, is also used for text classification problems, such as sentiment analysis. More specifically, it is a generalisation of Logistic Regression for multi-class scenarios [Yu et al., 2011]. Random Forest [Breiman, 2001] was used in some studies, where this supervised learning algorithm –which can be used for both classification and regression tasks –creates a forest (which is an ensemble of Decision Trees) and makes it somehow random. Moreover, some studies used the K-Nearest Neighbors [Altman, 1992] algorithm, one of the simplest classification algorithms where no learning is required, since the model structure is determined from the entire dataset. The SentiStrength [Thelwall et al., 2012] algorithm, is also used in both supervised and unsupervised cases, since the authors developed a version for each learning case. Conditional Random Fields [Lafferty et al., 2001] –a type of discriminative classifier that model the decision boundary amongst different classes, Linear Regression [Cook, 1977], the SANE optimization algorithm [Hu et al., 2013], and the Stochastic Gradient Descent [Bottou, 2010] algorithm, were also used in some studies. In addition, the following are other al-

²⁷<https://www.sentic.net/>

gorithms which were used in some studies: Passive Aggressive [Crammer et al., 2006], Bootstrap Aggregating [Breiman, 1996], Bayesian Network [Heckerman et al., 1995], Conjunctive Rule Based [Clark and Niblett, 1989], Adaptive Boosting [Freund et al., 1999], Hidden Markov Model [Baum and Petrie, 1966], Dictionary Learning [Ramirez et al., 2010], Support Vector Machine with Naïve Bayes features [Wang and Manning, 2012], Multiclass Classifier [Witten et al., 2016], and Iterative Classifier Optimizer [Witten et al., 2016].

In terms of unsupervised machine learning algorithms, the following were used in some studies: K-Means [Lloyd, 1982], Expectation Maximization [Dempster et al., 1977], Continuous Bag-Of-Word [Mikolov et al., 2013], and Denoising Autoencoder [Vincent et al., 2008].

Other studies proposed their own supervised, semi-supervised, and/or unsupervised algorithms, with some of the already established algorithms discussed above playing an important role in their implementation. Moreover, some studies proposed an ensemble method or evaluated ensemble-based classifiers. Ensembles created usually result in providing more accurate classification answers when compared to individual classifiers, i.e., classic learning approaches. In addition, ensembles reduce the overall risk of choosing a wrong classifier especially when applying it on a new dataset [Da Silva et al., 2014].

2.3.3.3 Deep Learning

Deep learning is a subset of machine learning based on Artificial Neural Networks (ANNs) –algorithms inspired by the human brain– where there are connections, layers, and neurons for data to propagate. Several studies adopted a deep learning-based approach to perform a form of Social Opinion Mining, where supervised and unsupervised algorithms were used.

The Long Short-Term Memory [Hinton et al., 2012], a prominent variation of the Recurrent Neural Network (RNN) [Graves and Schmidhuber, 2005] which makes it easier to remember past data in memory, was the most popular deep learning algorithm amongst the evaluated studies. Some further studies used the Bidirectional Long Short-Term Memory [Graves and Schmidhuber, 2005], an extension of the traditional Long Short-Term Memory which can improve model performance on sequence classification problems.

The Convolutional Neural Network (CNN) [LeCun et al., 1990] algorithm –a variant

of the ANN– was used by numerous studies. This algorithm is made up of neurons that have learnable weights and biases, where each neuron receives an input, performs a dot product, and optionally follows it with non-linearity. RNNs, a powerful set of ANNs useful for processing and recognising patterns in sequential data such as natural language, were also used for Social Opinion Mining. In particular, the Recursive Neural Tensor Network [Socher et al., 2013], a special network modification of the RNN, was used in some studies.

Some studies used a simple type of ANN, such as the feedforward neural network. Moreover, the Multilayer Perceptron [Hornik et al., 1989], a class of feedforward ANN, was used in a few studies. Similarly, some studies proposed methods based on the Autoencoder [Rumelhart et al., 1985] unsupervised learning algorithm which is used for representation learning. Lastly, the Gated Recurrent Units [Greff et al., 2017] and Dynamic Architecture for ANN [Ghiassi and Saidane, 2005] algorithms were also sometimes used.

2.3.3.4 Statistical

A few studies adopted a statistical approach to perform a form of Social Opinion Mining. In particular, one of the approaches proposed in [Arslan et al., 2017] uses the term frequency-inverse document frequency (TF-IDF) [Salton and McGill, 1986] numerical statistic to find out the important words within a tweet and to dynamically enrich Twitter specific dictionaries created by the authors. The TF-IDF is also one of several statistical-based techniques used in [Wang et al., 2018b] for comparing the proposed novel feature weighting approach for Twitter sentiment analysis. Moreover, [Raja and Swamynathan, 2016] focuses on a statistical sentiment score calculation technique based on adjectives, whereas the authors in [Yang et al., 2014] use a variation of the point-wise mutual information to measure the opinion polarity of an entity and its competitors, which method is different from the traditional Opinion Mining way.

2.3.3.5 Probabilistic

A probabilistic approach was used in some studies to perform a form of Social Opinion Mining. In particular, [Ou et al., 2014] propose a novel probabilistic model in the Content and Link Unsupervised Sentiment Model, where the focus is on microblog sentiment

classification incorporating link information, namely behaviour, same user, and friend.

2.3.3.6 Fuzziness

A few studies adopted a fuzzy-based approach to perform a form of Social Opinion Mining. In [D’Asaro et al., 2017], the authors present a sentiment evaluation and analysis system based on fuzzy linguistic textual analysis. In [Del Bosque and Garza, 2014], the authors assume that aggressive text detection is a sub-task of sentiment analysis, which is closely related to document polarity detection given that aggressive text can be seen as intrinsically negative. This approach considers the document’s length and the number of swear words as inputs, with the output being an aggressiveness value between 0 and 1.

2.3.3.7 Rule-based

Rule-based approaches were used in some studies to perform a form of Social Opinion Mining. Notably, the authors in [Bosco et al., 2013] applied an approach for automatic emotion annotation of ironic tweets. This relies on sentiment lexicons (words and expressions) and sentiment grammar expressed by compositional rules.

2.3.3.8 Graph

Some studies adopted a graph-based approach to perform a form of Social Opinion Mining. The study in [Vilarinho and Ruiz, 2018] presents a word graph-based method for Twitter sentiment analysis using global centrality metrics over graphs to evaluate sentiment polarity. In [Dritsas et al., 2018], a graph-based method is proposed for sentiment classification at a hashtag level. Moreover, the authors in [Chen et al., 2015] compare their proposed multimodal hypergraph-based microblog sentiment prediction approach with a combined hypergraph-based method [Huang et al., 2010]. Lastly, [Rabelo et al., 2012] used link mining techniques to infer the opinions of users.

2.3.3.9 Ontology

An ontology-based approach was adopted in some studies to perform a form of Social Opinion Mining. In particular, the technique developed in [Kontopoulos et al., 2013]

performs more fine-grained sentiment analysis of tweets where each subject within the tweets is broken down into a set of aspects, with each one being assigned a sentiment score.

2.3.3.10 Hybrid

Hybrid approaches are very much in demand for performing different Opinion Mining tasks, where more than half of the studies analysed adopted this approach. The majority of these studies used two different techniques (213 out of 282) within their hybrid approach, whereas 62 used three, and 7 studies used four different techniques.

The lexicon and machine learning-based techniques were mostly used, where they accounted for 40% of the hybrid approaches, followed by lexicon and statistical-based (7.8%), machine learning and statistical-based (7.4%), and lexicon, machine learning, and statistical-based (7.4%) techniques.

Moreover, out of the 282 hybrid approaches, 232 used lexicons, 205 used machine learning, and 39 used deep learning. These numbers reflect the importance of these three techniques within the Social Opinion Mining research and development domain. In light of these, a list of lexicons, machine learning algorithms, and deep learning algorithms used in these studies has been compiled, similar to Sections 2.3.3.1, 2.3.3.2, and 2.3.3.3 above. These were either used in the proposed method/s and/or for comparison purposes in the respective studies.

In terms of state-of-the-art lexicons, the top ones align with the results obtained from the lexicon-based approaches in Section 2.3.3.1. The following are the lexicons used for more than ten times across the hybrid approaches:

1. SentiWordNet;
2. MPQA - Subjectivity;
3. Hu and Liu;
4. WordNet;
5. AFINN;

6. SentiStrength;
7. HowNetSenti;
8. NRC Word-Emotion Association Lexicon;
9. NRC Hashtag Sentiment Lexicon²⁸ [Mohammad et al., 2013];
10. SenticNet;
11. Sentiment140 Lexicon (also known as NRC Emoticon Lexicon)²⁹ [Mohammad et al., 2013];
12. National Taiwan University Sentiment Dictionary [Ku et al., 2006]; and
13. Wikipedia list of emoticons.

Further to the quoted lexicons, a number of studies used lexicons that they created as part of their work. Some studies composed their lexicons from emoticons/emojis that were extracted from a dataset, combined publicly available emoticon lexicons/lists or mapped emoticons to their corresponding polarity, and others used seed/feeling/emotional words to establish a microblog typical emotional dictionary. Additionally, some authors constructed or used sentiment lexicons some of which are domain or language specific, others extended state-of-the-art lexicons, and some made the lexicons created available to the research community such as the Distributional Polarity Lexicon [Castellucci et al., 2016].

Several machine learning algorithms were used within the hybrid approaches. The Support Vector Machine and Naïve Bayes algorithms were mostly used for supervised learning, which result corresponds to the machine learning-based approaches in Section 2.3.3.2 above. Moreover, the Logistic Regression, Decision Trees, Random Forest, Maximum Entropy, and SentiStrength (used in both supervised and unsupervised settings) algorithms were also used in various studies. Notably, some additional algorithms from the ones used in the machine learning-based approaches in Section 2.3.3.2 above, were used in a hybrid approach, in particular, Support Vector Regression [Drucker et al., 1997], Extremely Randomised Trees [Geurts et al., 2006], Least Median of Squares Regression

²⁸<http://saifmohammad.com/WebDocs/NRC-Hashtag-Sentiment-Lexicon-v0.1.zip>

²⁹<http://saifmohammad.com/Lexicons/Sentiment140-Lexicon-v0.1.zip>

[Rousseeuw, 1984], Maximum Likelihood Estimation [Fisher, 1925], Hyperpipes [Witten et al., 2016], Extreme Learning Machine [Huang et al., 2006], Domain Adaptation Machine [Duan et al., 2009], RIPPER [Cohen, 1995], Affinity Propagation [Frey and Dueck, 2007], Multinomial inverse regression [Taddy, 2013], Apriori [Agrawal et al., 1994], Distant Supervision [Go et al., 2009], and Label Propagation [Zhu and Ghahramani, 2002].

Given that deep learning is a subset of machine learning, the algorithms used within the hybrid approaches are also discussed. Besides the deep learning algorithms mentioned in Section 2.3.3.3, the following are the others used:

- Deep Belief Network [Hinton and Salakhutdinov, 2006], a probabilistic generative model that is composed of multiple layers of stochastic, latent variables;
- Generative Adversarial Networks (GANs) [Goodfellow et al., 2014], deep neural net architectures composed of a two networks, a generator and a discriminator, pitting one against the other;
- Conditional Generative Adversarial Networks [Mirza and Osindero, 2014], a conditional version of GANs that can be constructed by feeding the data that needs to be conditioned on both the generator and discriminator; and
- Hierarchical Attention Network, a neural architecture for document classification [Yang et al., 2016].

Further to the quoted algorithms, several studies used ensemble learning methods in their work, where they combined the output of several base machine learning and/or deep learning methods. In particular, [Gonçalves et al., 2013] compared eight popular lexicon and machine learning-based sentiment analysis algorithms, and then developed an ensemble that combines them, which in turn provided the best coverage results and competitive agreement. Moreover, [Ghosal et al., 2018] proposes a Multilayer Perceptron-based ensemble network that combines Long Short-Term Memory, CNN, and feature-based Multilayer Perceptron models, with each model incorporating character, word, and lexicon level information, to predict the degree of intensity for sentiment and emotion. Lastly, the Random Forest ensemble learning method was used in a number of studies.

2.3.3.11 Manual and Other

Some of the studies did not adopt any of the previous approaches (discussed in Sections 2.3.3.1-2.3.3.10). This is mainly due to three reasons: no information provided by the authors, use of an automated approach, or use of a manual approach to perform a form of Social Opinion Mining. Regarding the former, the majority of them were not specifically focused on Social Opinion Mining (this was secondary), in contrast to the others. As for the automated approaches, some of them used cloud services, such as Microsoft Azure Text Analytics³⁰ or out-of-the-box functionality provided by existing tools/software libraries, such as the TextBlob³¹ Python library.

2.3.4 Social Media Platforms

Social data refers to online data generated from any type of social media platform be it from microblogging, social networking, blogging, photo/video sharing, and crowdsourcing. Given that this systematic review focuses on Opinion Mining approaches that make use of social networking and microblogging services, the social media platforms used in the studies were identified and are listed below:

1. Twitter: a microblogging platform that allows publishing of short text updates (“microposts”);
2. Sina Weibo³²: a Chinese microblogging platform that is like a hybrid of Twitter and Facebook;
3. Facebook: a social networking platform that allows users to connect and share content with family and friends online;
4. YouTube³³: a video sharing platform;
5. Tencent Weibo³⁴: a Chinese microblogging platform;

³⁰<https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/>

³¹<https://textblob.readthedocs.io/en/dev/>

³²<https://weibo.com/>

³³<https://www.youtube.com/>

³⁴Original platform points to <http://t.qq.com/> which is not online anymore.

6. TripAdvisor: a travel platform that allows people to post their reviews about hotels, restaurants, and other travel-related content, besides offering accommodation bookings;
7. Instagram³⁵: a platform for sharing photos and videos from a smartphone;
8. Flickr³⁶: an image- and video-hosting platform that is popular for sharing personal photos;
9. Myspace³⁷: a social networking platform for musicians and bands to show and share their talent and connect with fans;
10. Digg³⁸: a social bookmarking and news aggregation platform that selects stories to the specific audience;
11. Foursquare³⁹: formerly a location-based service and nowadays a local search and discovery service mobile application known as Foursquare City Guide;
12. Stocktwits⁴⁰: a social networking platform for investors and traders to connect with each other;
13. LinkedIn⁴¹: a professional networking platform that allows users to communicate and share updates with colleagues and potential clients, job searching, and recruitment;
14. Plurk⁴²: a social networking and microblogging platform;
15. Weixin⁴³: a Chinese multi-purpose messaging and social media app developed by Tencent;
16. PatientsLikeMe⁴⁴: a health information sharing platform for patients;

³⁵<https://www.instagram.com>

³⁶<https://www.flickr.com/>

³⁷<https://myspace.com/>

³⁸<http://digg.com/>

³⁹<https://foursquare.com/>

⁴⁰<https://stocktwits.com/>

⁴¹<https://www.linkedin.com/>

⁴²<https://www.plurk.com>

⁴³<https://weixin.qq.com/>

⁴⁴<https://www.patientslikeme.com/>

17. Apontador⁴⁵: a Brazilian platform that allows users to share their opinions and photos on social networks and also book hotels and restaurants; and
18. Google+⁴⁶: formerly a social networking platform (shut down in April 2019) that included features such as posting photos and status updates, group different relationship types into Circles, organise events, and location tagging.

Overall, Twitter was the most popular with 371 Opinion Mining studies making use of it, followed by Sina Weibo with 46, and Facebook with 30. Other popular platforms such as YouTube (12), Tencent Weibo (8), TripAdvisor (7), Instagram (6), and Flickr (5) were also used in a few studies. These results show the importance and popularity of microblogging platforms, such as Twitter and Sina Weibo, which are also very frequently used for research and development purposes in this domain. Such microblogging platforms provide researchers the possibility of using an Application Programming Interface (API) to access social data, which plays a crucial role in selecting them for their studies. On the other hand, data retrieval from other social media platforms such as Facebook, is becoming more challenging due to ethical concerns. For example, Facebook access to the Public Feed API⁴⁷ is restricted and users cannot apply for it.

2.3.5 Social Datasets

Numerous datasets were used across the studies evaluated for the systematic review. These consisted of Social Opinion Mining datasets released online for public use –which have been widely used across the studies– and newly collected datasets, some of which were made available for public use or else for private use within the respective studies. In terms of data collection, the majority of them used the respective platform’s API, such as the Twitter Search API⁴⁸, either directly or through a third-party library e.g., Twitter4J⁴⁹. Due to the large number of datasets, only the ones mostly used shall be discussed within this section. In addition, only social datasets are mentioned irrespective of whether other

⁴⁵<https://www.apontador.com.br/>

⁴⁶<https://plus.google.com/>

⁴⁷Original API pointed to https://developers.facebook.com/docs/public_feed/ which is not available anymore.

⁴⁸<https://developer.twitter.com/en/docs/tweets/search/overview>

⁴⁹<http://twitter4j.org/en/> - a Java library for the Twitter API

non-social datasets (e.g., news, movies, etc.) were used, given that the main focus of the review is on social data.

The following are the top ten social datasets used across all studies:

1. **Stanford Twitter Sentiment (STS)** [Go et al., 2009]: 1,600,000 training tweets collected via the Twitter API, that is made up of 800,000 tweets containing positive emoticons and 800,000 tweets containing negative emoticons. These are based on various topics, such as Nike, Google, China, Obama, Kindle, San Francisco, North Korea, and Iran.
2. **Sanders**⁵⁰: 5513 hand-classified tweets about four topics: Apple, Google, Microsoft, and Twitter. These tweets are labelled as follows: 570 positive, 654 negative, 2,503 neutral, and 1,786 irrelevant.
3. **SemEval 2013 - Task 2**⁵¹ [Nakov et al., 2013]: Training, development, and test sets for Twitter and SMS messages were annotated with positive, negative, and objective/neutral labels via the Amazon Mechanical Turk crowdsourcing platform. This was done for 2 subtasks focusing on an expression-level and message-level.
4. **SemEval 2014 - Task 9**⁵² [Rosenthal et al., 2014]: Continuation of SemEval 2013 - Task 2, where three new test sets from regular and sarcastic tweets, and LiveJournal sentences were introduced.
5. **STS Gold (STS-Gold)** [Saif et al., 2013]: A subset of STS, which was annotated manually at a tweet and entity-level. The tweet labels were either positive, negative, neutral, mixed, or other.
6. **Health care reform** [Speriosu et al., 2011]: Dataset contains tweets about the 2010 health care reform in the USA. A subset of these are annotated for polarity with the following labels: positive, negative, neutral, irrelevant. The polarity targets, such as health care reform, conservatives, democrats, liberals, republicans, Obama, Stupak,

⁵⁰Original dataset points to <http://www.sananalytics.com/lab/twitter-sentiment/> which is not online anymore.

⁵¹Original dataset points to <https://www.cs.york.ac.uk/semeval-2013/task2/> which is not online anymore.

⁵²<http://alt.qcri.org/semeval2014/task9/>

and Tea Party, were also annotated. All were distributed into training, development, and test sets.

7. **Obama-McCain Debate** [Shamma et al., 2009]: 3,238 tweets about the first presidential debate held in the USA for the 2008 campaign. The sentiment labels of the tweets are acquired by [Diakopoulos and Shamma, 2010] using Amazon Mechanical Turk, and are rated as either positive, negative, mixed, or other.
8. **SemEval 2015 - Task 10**⁵³ [Rosenthal et al., 2015]: This continues on datasets number 3 and 4, with three new subtasks. The first two target sentiment about a particular topic in one tweet or collection of tweets, whereas the third targets the degree of prior polarity of a phrase.
9. **SentiStrength Twitter** [Thelwall et al., 2012]: Six human-coded databases from BBC, Digg, MySpace, Runners World, Twitter, and YouTube annotated for sentiment polarity strength i.e., negative between -1 (not negative) and -5 (extremely negative), and positive between 1 (not positive) and 5 (extremely positive).
10. **SemEval 2016 - Task 4**⁵⁴ [Nakov et al., 2016]: This is a re-run of dataset 7, with three new subtasks. The first one replaces the standard two-point scale (positive/negative) or three-point scale (positive/negative/neutral) with a five-point scale (very positive/positive/OK/negative/very negative). The other two subtasks replaced tweet classification with quantification (i.e., estimating the distribution of the classes in a set of unlabelled items) according to a two-point and five-point scale, respectively.
11. **NLPCC 2012**⁵⁵: Chinese microblog sentiment dataset (sentence level) from Tencent Weibo provided by the First Conference on Natural Language Processing and Chinese Computing (NLP&CC 2012). It consists of a training set of microblogs about two topics, and a test set about 20 topics, where the subjectivity (subjective/objective) and the polarity (positive/negative/neutral) was assigned for each.

⁵³<http://alt.qcri.org/semEval2015/task10/>

⁵⁴<http://alt.qcri.org/semEval2016/task4/>

⁵⁵<http://tcci.ccf.org.cn/conference/2012/>

12. **NLPCC 2013**⁵⁶: Dataset from Sina Weibo used for the Chinese Microblog Sentiment Analysis Evaluation task in the second conference on NLP&CC 2013. The Chinese microblogs were classified into 7 emotion types: anger, disgust, fear, happiness, like, sadness, and surprise. Test set contains 10,000 microblogs, where each text is labelled with a primary emotion type and a secondary one (if possible).
13. **Sentiment Evaluation (SE-Twitter)** [Narr et al., 2012]: Human annotated multilingual dataset of 12,597 tweets from 4 languages, namely English, German, French, and Portuguese. Polarity annotations with labels: positive, negative, neutral, and irrelevant, were conducted manually using Amazon Mechanical Turk.
14. **SemEval 2017 - Task 4** [Rosenthal et al., 2017]: This dataset continues with a re-run of dataset 10, where two new changes were introduced; inclusion of the Arabic language for all subtasks and provision of profile information of the Twitter users that posted the target tweets.

All the datasets above are textual, with the majority of them composed of social data from Twitter. From the datasets above, in terms of language, only the SE-Twitter (number 13) social dataset can be considered as multilingual, with the rest targeting English (majority) or Chinese microblogs, whereas SemEval 2017 - Task 4 (number 14) introduced a new language in Arabic. An additional dataset is the one produced by Mozetič et al., which contains 15 Twitter sentiment corpora for 15 European languages [Mozetič et al., 2016]. Some studies such as [Munezero et al., 2015] used one of the English-based datasets above (STS-Gold) for multiple languages, given that they adopted a lexicon-based approach. Moreover, these datasets had different usage within the respective studies, with the most common being used as a training/test set, the final evaluation of the proposed solution/lexicon, or for comparison purposes. Evaluation challenges like the International Workshop on Semantic Evaluation (SemEval)⁵⁷ are important to generate social datasets such as the ones mentioned above and [Cortis et al., 2017], since these can be used by the Opinion Mining community for further research and development.

⁵⁶<http://tcci.ccf.org.cn/conference/2013/>

⁵⁷<https://semeval.github.io/>

2.3.6 Language

Multilingual/bilingual Social Opinion Mining is very challenging, since it deals with multicultural social data. For example, analysing Chinese and English online posts can bring a mixed sentiment on such posts. Therefore, it is hard for researchers to make a fair judgement in cases where online posts' results from different languages contradict each other [Yan et al., 2014].

The majority of the studies (354 out of 465) considered for the systematic review analysis support one language in their Social Opinion Mining solutions. A total of 80 studies did not specify whether their proposed solution is language-agnostic or otherwise, or else their modality was not textual-based. Lastly, only 31 studies cater for more than one language, with 18 being bilingual, 1 being trilingual, and 12 proposed solutions claiming to be multilingual. Regarding the latter, the majority were tested on a few languages at most, with [Castellucci et al., 2015a, Castellucci et al., 2015b] on English and Italian; [Montejo-Raez et al., 2014] on English and Spanish; [Erdmann et al., 2014] on English and Japanese; [Radhika and Sankar, 2017] on English and Malayalam⁵⁸; [Baccouche et al., 2018] on English, French, and Arabic; [Munezero et al., 2015] on keyword sets for different languages (e.g., Spanish, French); [Wehrmann et al., 2017] on English, Spanish, Portuguese, and German; [Cui et al., 2011] on Basic Latin (English) and Extended Latin (Portuguese, Spanish, German); [Teixeira and Laureano, 2017] on Spanish, Italian, Portuguese, French, English, and Arabic; [Zhang et al., 2017] on 8 languages, namely English, German, Portuguese, Spanish, Polish, Slovak, Slovenian, and Swedish; and [Gao et al., 2016] on 11 languages, namely English, Dutch, French, German, Italian, Polish, Portuguese, Russian, Spanish, Swedish, and Turkish.

The list below specifies the languages supported by the 19 bilingual and trilingual studies:

- English and Italian;
- English and German;
- English and Spanish;

⁵⁸Malayalam is a Dravidian language spoken in the Indian state of Kerala and the union territories of Lakshadweep and Puducherry by the Malayali people.

- English and Brazilian Portuguese;
- English and Chinese;
- English and Dutch;
- English and Greek;
- English and Hindi;
- English and Japanese;
- English and Roman-Urdu;
- English and Swedish;
- English and Korean; and
- English, German, and Spanish.

Some studies above, such as [D’Avanzo and Pilato, 2015, Anjaria and Guddeti, 2014, Tumasjan et al., 2010] translated their input data into an intermediate language, mostly English, to perform Social Opinion Mining.

Moreover, Table 2.4 provides a list of the non-English languages identified from the 354 studies that support one language. Authors in [Chou et al., 2017] claim that their method can be easily applied to any ConceptNet⁵⁹ supported language, with [Wang et al., 2016] similarly claiming that their method is language independent, whereas the solution by [Wang and Wu, 2015] is multilingual given that emoticons are used in the majority of languages.

2.3.7 Modality

The majority of the studies in this systematic review and in the state-of-the-art focus on Social Opinion Mining on the textual modality, with only 15 out of 465 studies applying their work on more than one modality. However, other modalities, such as visual (image, video), and audio information is often ignored, even though it contributes greatly towards

⁵⁹<http://conceptnet.io/> – an open, multilingual knowledge graph

Language	Total Studies
Chinese	53
Spanish	11
Indonesian	8
Italian	5
Arabic	5
Portuguese	3
Brazilian Portuguese	3
Japanese	3
Korean	2
French	2
French - Bambara	1
Bulgarian	1
German	1
Roman Urdu	1
Russian	1
Swiss German	1
Thai	1
Persian	1
Bengala	1
Vietnamese	1

Table 2.4: Non-English languages supported by studies in this review analysis

expressing user emotions [Chen et al., 2015]. Moreover, when two or more modalities are considered together for any form of social opinion, such as emotion recognition, they are often complementary, thus increase the system’s performance [Caschera et al., 2016]. Table 2.5 lists the multimodal studies within the review analysis, with the ones catering for two modalities –text and image– being the most popular.

Text	Image	Video	Audio	Studies
✓	✓			[Ortis et al., 2018, Rai et al., 2018, Saini et al., 2018, Chen et al., 2017, Chen et al., 2015, Baecchi et al., 2016, Liu et al., 2015, Zhang et al., 2015, Wang et al., 2014a, Flaes et al., 2016, Cai and Xia, 2015, Yuan et al., 2015]
		✓	✓	[Song and Gruzd, 2017]
✓		✓	✓	[Caschera et al., 2016]
✓	✓	✓	✓	[Poria et al., 2016]

Table 2.5: Studies adopting a multimodal approach

2.3.7.1 Datasets

Current available datasets and resources for Social Opinion Mining are restricted to the textual modality only. The following are the non-textual social datasets (not listed in Section 2.3.5) used across the mentioned studies:

1. **YouTube Dataset** [Morency et al., 2011] used in [Poria et al., 2016]: 47 videos targeting various topics, such as politics, electronics, and product reviews.
2. **SentiBank Twitter Dataset**⁶⁰ [Borth et al., 2013] used in [Baecchi et al., 2016, Cai and Xia, 2015]: Image dataset from Twitter annotated for polarity using Amazon Mechanical Turk. Tweets with images related to 21 hashtags (topics) resulted in 470 being positive and 133 being negative.
3. **SentiBank Flickr Dataset** [Borth et al., 2013] used in [Cai and Xia, 2015]: 500,000 image posts from Flickr labeled by 1,553 adjective noun pairs based on Plutchik’s Wheel of Emotions (psychological theory) [Plutchik, 1980].
4. **You Image Dataset** [You et al., 2015] used in [Cai and Xia, 2015]: Image dataset from Twitter consisting of 769 positive and 500 negative tweets with images, annotated using Amazon Mechanical Turk.
5. **Katsurai and Sotoh Image Dataset**⁶¹ [Katsurai and Satoh, 2016] used in [Ortis et al., 2018]: Dataset of images from Flickr (90,139) and Instagram (65,439) with their sentiment labels.

2.3.7.2 Observations

The novel methodology by [Poria et al., 2016], is the only multimodal sentiment analysis approach which caters for four different modalities, namely text, vision (image and video), and audio. Sentiments are extracted from social Web videos. In [Caschera et al., 2016], the authors propose a method whereby machine learning techniques need to be trained on different and heterogeneous features when used on different modalities, such as polarity and intensity of lexicons from text, prosodic features from audio, and postures, gestures, and

⁶⁰<http://www.ee.columbia.edu/ln/dvmm/vso/download/sentibank.html>

⁶¹<http://mm.doshisha.ac.jp/senti/CrossSentiment.html>

expressions from video. The sentiment of video and audio data in [Song and Gruzd, 2017] was manually coded, which task is labour intensive and time consuming. The addition of images to the microblogs' textual data reinforces and clarifies certain feelings [Wang et al., 2014a, Baecchi et al., 2016], thus improving the sentiment classifier with the image features [Liu et al., 2015, Zhang et al., 2015, Wang et al., 2014a, Cai and Xia, 2015]. Similarly, [Chen et al., 2015] also demonstrates superiority with their multimodal hypergraph method when compared to single modality (in this case textual) methods. Moreover, these results are further supported by the method in [Poria et al., 2016] –which caters for more than two modalities, in audio, visual, and textual– where it shows that accuracy improves drastically when such modalities are used together.

The authors in [Flaes et al., 2016] apply their multimodal (text, images) method in a real world application area, which research shows that several relationships exist between city liveability indicators collected by the local government and sentiment that is extracted automatically. For example, a negative linear association of detected sentiment from Flickr data is related with people living on welfare checks. Results in [Rai et al., 2018] show that there is a high correlation between sentiment extracted from text-based social data and image-based landscape preferences by humans. In addition, results in [Yuan et al., 2015] show some correlation between image and textual tweets. However, the authors mention that more features and robust data is required to determine the exact influence of multimedia content in the social domain. The work in [Chen et al., 2017] adopts a bimodal approach to solve the problem of cross-domain image sentiment classification by using textual features and visual features from the target domain and measuring the text/image similarity simultaneously.

Therefore, multimodality in the Social Opinion Mining domain is one of numerous research gaps identified in the systematic review. This provides researchers with an opportunity towards further research, development, and innovation in this area.

2.3.8 Tools and Technologies

An analysis was also carried out on the tools and technologies that were used across all studies of the systematic review. This was done for various Opinion Mining operations

conducted on social data, such as NLP, machine learning, and Opinion Mining-specific tools. The sub-sections below provide respective lists for the ones mostly used across the studies for the various operations required.

2.3.8.1 NLP

The following are the top 5 NLP tools used across all studies for various NLP tasks:

- Natural Language Toolkit (NLTK)⁶²: a platform that provides lexical resources, text processing libraries for classification, tokenisation, stemming, tagging, parsing, and semantic reasoning, and wrappers for industrial NLP libraries;
- TweetNLP⁶³: consists of a tokeniser, Part-of-Speech (POS) tagger, hierarchical word clusters, and a dependency parser for tweets, besides annotated corpora and web-based annotation tools;
- Stanford NLP⁶⁴: software that provides statistical NLP, deep learning NLP, and rule-based NLP tools, such as Stanford CoreNLP, Stanford Parser, and Stanford POS Tagger;
- NLP-ICTCLAS⁶⁵: a Chinese word segmentation system that includes keyword extraction, POS tagging, NER, and microblog analysis, amongst other features; and
- word2vec⁶⁶: an efficient implementation of the continuous bag-of-words and skip-gram architectures for computing vector representations of words.

2.3.8.2 Machine Learning

The top 5 machine learning tools used across all studies are listed below:

- Weka⁶⁷: a collection of machine learning algorithms for data mining tasks, including tools for data preparation, classification, regression, clustering, association rules mining, and visualisation;

⁶²<https://www.nltk.org/>

⁶³<http://www.cs.cmu.edu/~ark/TweetNLP/>

⁶⁴<https://nlp.stanford.edu/software/>

⁶⁵<http://ictclas.nlpir.org/>

⁶⁶<https://code.google.com/archive/p/word2vec/>

⁶⁷<https://www.cs.waikato.ac.nz/ml/weka/>

- scikit-learn⁶⁸: consists of a set of tools for data mining and analysis, such as classification, regression, clustering, dimensionality reduction, model selection, and pre-processing;
- LIBSVM⁶⁹: an integrated software for support vector classification, regression, distribution estimation, and multi-class classification;
- LIBLINEAR⁷⁰: a linear classifier for data with millions of instances and features; and
- SVM-Light⁷¹: an implementation of SVMs for pattern recognition, classification, regression, and ranking problems.

2.3.8.3 Opinion Mining

Certain studies used Opinion Mining tools in their research to either conduct their main experiments or for comparison purposes to their proposed solution/s. The following are the top 3 Opinion Mining tools used:

- SentiStrength⁷²: a sentiment analysis tool that is able to conduct binary (positive/negative), trinary (positive/neutral/negative), single-scale (-4 very negative to very positive +4), keyword-oriented, and domain-oriented classifications;
- Sentiment140⁷³: a tool that allows you to discover the sentiment of a brand, product, or topic on Twitter; and
- Valence Aware Dictionary and sEntiment Reasoner (VADER)⁷⁴: a lexicon and rule-based sentiment analysis tool that is specifically focused on sentiments expressed in social media.

⁶⁸<https://scikit-learn.org/>

⁶⁹<https://www.csie.ntu.edu.tw/~cjlin/libsvm/>

⁷⁰<https://www.csie.ntu.edu.tw/~cjlin/liblinear/>

⁷¹<http://svmlight.joachims.org/>

⁷²<http://sentistrength.wlv.ac.uk/>

⁷³<http://www.sentiment140.com/>

⁷⁴<https://github.com/cjhutto/vaderSentiment>

2.3.9 Natural Language Processing Tasks

This section presents information about other NLP tasks that were conducted to perform Social Opinion Mining.

2.3.9.1 Overview

An element of NLP is performed in over 60% of the studies analysed, either for pre-processing, feature extraction (machine learning), or one of the processing parts within their Social Opinion Mining solution. The most common and important NLP tasks range from Tokenisation, Segmentation and POS, to NER and Language Detection.

It is important to mention that the NLP tasks mentioned above together with Anaphora Resolution, Parsing, Sarcasm, and Sparsity, are some other challenges faced in the Social Opinion Mining domain [Khan et al., 2014]. Moreover, online posts with complicated linguistic patterns are challenging to deal with [Li and Xu, 2014].

However, the authors in [Koto and Adriani, 2015c] showcase the importance and potential of NLP within this domain, where they investigated the pattern or word combination of tweets in subjectivity and polarity by considering their POS sequence. Results reveal that subjective tweets tend to have word combinations consisting of adverb and adjective, whereas objective tweets tend to have a word combination of nouns. Moreover, negative tweets tend to have a word combination of affirmation words which often appear as a negation word.

2.3.9.2 Pre-processing and negations

The majority (over 75%) of the studies performed some sort of pre-processing in their studies. Different methods and resources were used for such a process, such as NLP tasks (e.g., tokenisation, stemming, lemmatisation, NER), dictionaries for stop words, acronyms for slang words, and others (e.g., noslang.com, noswearing.com, Urban Dictionary, Internet lingo).

Negation handling is one of the most challenging issues faced by Social Opinion Mining solutions. However, 25% of the studies cater for negations within their approach. Several different methods are used, such as negation replacement, negation transformation,

negation dictionaries, textual features based on negation words, and negation models.

2.3.9.3 Emoticons/Emojis

Social media can be seen as a sub-language that uses emoticons/emojis mixed with text to show emotions [Min et al., 2013]. Emoticons/emojis are commonly used in tweets irrespective of the language, therefore are sometimes considered as being domain and language independent [Khan et al., 2014], thus useful for multilingual Social Opinion Mining [Cui et al., 2011].

Even though some researchers remove emoticons/emojis as part of their pre-processing stage (depending on what the authors want to achieve), many others have utilised the respective emotional meaning within their Social Opinion Mining process. This has led to emoticons/emojis in playing a very important role within 44% solutions of the analysed studies especially when the focus is on emotion recognition.

Results obtained from the emoticon networks model in [Zhang et al., 2013] show that emoticons can help in performing sentiment analysis. This is further supported by [Jiang et al., 2015] who found that emoticons are a pure carrier of sentiment. This is further supported by the results obtained by the emoticon polarity-aware method in [Li et al., 2018] which show that emoticons can significantly improve the precision for identifying the sentiment polarity. In the case of hybrid (lexicon and machine learning) approaches, emoticon-aided lexicon expansion improve the performance of lexicon-based classifiers [Zhou et al., 2014]. From an emotion classification perspective, Porshnev et al. [Porshnev et al., 2014] analysed users' emoticons on Twitter to improve the accuracy of predictions for the Dow Jones Industrial Average and S&P 500 stock market indices. Other researchers [Cvijikj and Michahelles, 2011] were interested in analysing how people express emotions, displayed via adjectives or usage of internet slang i.e., emoticons, interjections, and intentional misspelling.

Several emoticon lists were used in these studies, with the Wikipedia and DataGenetics⁷⁵ ones commonly used. Moreover, emoticon dictionaries, such as [Agarwal et al., 2011, Aisopos et al., 2012, Becker et al., 2013], consisting of emoticons and their corre-

⁷⁵<http://www.datagenetics.com/blog/october52012/index.html>

sponding polarity class were also used in certain studies.

2.3.9.4 Word embeddings

Word embeddings, a type of word representation which allows words with a similar meaning to have a similar representation, were used by several studies adopting a learning-based (machine learning, deep learning, and statistical) or hybrid approach. These studies used word embedding algorithms, such as word2vec, fastText⁷⁶, and/or GloVe⁷⁷. Such a form of learned representation for text is capable of capturing the context of words within a piece of text, syntactic patterns, semantic similarity, and relation with other words, amongst other word representations. Therefore, word embeddings are used for different NLP problems, with Social Opinion Mining being one of them.

2.3.9.5 Aspect-based Social Opinion Mining

Sentence-level Social Opinion Mining approaches tend to fail in discovering an opinion dimension, such as sentiment polarity about a particular entity and/or its aspects [Cambria et al., 2013]. Therefore, an aspect-level (also referred to as feature/topic-based) [Hu and Liu, 2004] approach –where an opinion is made up of targets and their associated opinion dimension (e.g., sentiment polarity)– has been used in some studies to overcome such issues. Certain NLP tasks, such as a parsing, POS tagging, and NER, are usually required to extract the entities or aspects from the respective social data.

From all the studies analysed, only 39 performed aspect-based Social Opinion Mining as categorised in Table 2.6 based on the respective dimension of Opinion Mining conducted, with 37 focusing on aspect-based sentiment analysis, 1 on aspect-based sentiment and emotion analysis, and 1 on aspect-based affect analysis. In particular, the Twitter aspect-based sentiment classification process in [Lek and Poo, 2013] consists of the following main steps: aspect-sentiment extraction, aspect ranking and selection, and aspect classification, whereas Lau et al. [Lau et al., 2014] use NER to parse product names to determine their polarity. The aspect-based sentiment analysis approach in [Hagge et al., 2017] leveraged POS tagging and dependency parsing. Moreover, [Zainuddin et al., 2016a]

⁷⁶<https://fasttext.cc/>

⁷⁷<https://nlp.stanford.edu/projects/glove/>

proposed a hybrid approach to analyse aspect-based sentiment of tweets. As the authors claim, it is more important to identify the opinions of tweets rather than finding the overall polarity which might not be useful to organisations. In [Zainuddin et al., 2018], the same authors used association rule mining augmented with a heuristic combination of POS patterns to find single and multi-word explicit and implicit aspects. Results in [Jiang et al., 2011] show that classifiers incorporating target-dependent features significantly outperform target-independent ones. In contrast to the studies discussed, [Weichselbraun et al., 2017] introduced an aspect-based analysis approach that integrates affective (includes sentiment polarity and emotions) and factual knowledge extraction to capture opinions related to certain aspects of brands and companies. The social data analysed is classified in terms of sentiment polarity and emotions, aligned with the “Hourglass of Emotions” [Susanto et al., 2020].

Dimension	Number of Studies	References of Studies
sentiment analysis	37	[Bansal and Srivastava, 2018, Dragoni, 2018, Gandhe et al., 2018, Ghiassi and Lee, 2018, Kao and Huang, 2018, Katz et al., 2018, Liu et al., 2018, Rathan et al., 2018, Wang et al., 2018a, Zainuddin et al., 2018, Abdullah and Zolkepli, 2017, Dambhare and Karale, 2017, Hagge et al., 2017, Ray and Chakrabarti, 2017, Rout et al., 2017, Tong et al., 2017, Vo et al., 2017, Zhou et al., 2017, Zimbra et al., 2016, Zainuddin et al., 2016a, Zainuddin et al., 2016b, Kokkinogenis et al., 2015, Lima et al., 2015, Hridoy et al., 2015, Castellucci et al., 2015a, Averchenkov et al., 2015, Tan et al., 2014, Lau et al., 2014, Del Bosque and Garza, 2014, Varshney and Gupta, 2014, Unankard et al., 2014, Lek and Poo, 2013, Wang and Ye, 2013, Min et al., 2013, Kontopoulos et al., 2013, Jiang et al., 2011, Prabowo and Thelwall, 2009]
sentiment analysis and emotion analysis	1	[Aoudi and Malik, 2018]
affect analysis	1	[Weichselbraun et al., 2017]

Table 2.6: Studies performing aspect-based Social Opinion Mining

In terms of techniques, the majority of the aspect-based studies used a hybrid ap-

proach, where only 5 studies used deep learning for such a task. In particular, the study by [Averchenkov et al., 2015] used a deep learning approach based on RNNs for aspect-based sentiment analysis. A comparative review of deep learning for aspect-based sentiment analysis published by [Do et al., 2019] discusses current research in this domain. It focuses on deep learning approaches, such as CNN, Long Short-Term Memory, and Gated Recurrent Units, for extracting both syntactic and semantic features of text without the need for in-depth requirements for feature engineering as required by classical NLP.

2.3.10 Latest research

Given that this systematic review covers studies till 2018, some recent developments and advancements from 2019 till 2021 shall be discussed within this sub-section. This shows the fast research turnaround in Social Opinion Mining which has kept evolving at an incredibly fast rate, thus reiterating its validity and popularity as a research area.

The number of studies using deep learning approaches continued to increase, especially ones using certain deep learning techniques, such as CNNs, RNNs, Long Short-Term Memory, Gated Recurrent Units, and Deep Belief Networks [Yadav and Vishwakarma, 2020, Wadawadagi and Pagi, 2020], and with the introduction of new techniques, such as Transfer Learning. This is supported by numerous studies [Carvalho and Plastino, 2021, Eke et al., 2020] who have noted that researchers are shifting from using traditional machine learning techniques to deep learning ones. The authors in [Carvalho and Plastino, 2021] focus on sentiment analysis on tweets, [Xu et al., 2020] focus on emotion classification on tweets, [Akhtar et al., 2020] focus on sentiment and emotion intensity, [Cignarella et al., 2020] focus on irony detection of English, Spanish, French, and Italian tweets, whereas in [Eke et al., 2020] the authors focus on sarcasm detection with Twitter also being the social media platform mostly used in this research area.

Transfer learning is a deep learning technique where a model is trained for one or more tasks (source tasks), in which learnt knowledge is applied to a related second task (target task) [Pan and Yang, 2009]. In particular, the Transformer model architecture introduced in [Vaswani et al., 2017], is based on attention mechanisms and is designed to handle sequential data like natural language for NLP tasks, such as sentiment analysis

and text summarisation. This has coincided with the advancement of Social Opinion Mining for different opinion dimensions, such as sentiment polarity [Nguyen et al., 2020, Naseem et al., 2020], emotion [Acheampong et al., 2021], and irony [Nguyen et al., 2020], especially studies focused on adaptation to new domains and/or knowledge transfer from one language to another. The latter application is extremely reliable for cross-lingual adaptations where a labelled dataset is available in one language e.g., English, which is then applied to another language, such as low-resourced languages [Ruder, 2017].

With respect to language, more Social Opinion Mining studies supporting languages other than the popular ones (such as English and Chinese) are on the rise. In [Rani and Kumar, 2019], the authors discuss the growth of research work in the fields of sentiment and emotion analysis for Indian languages. Moreover, [Buechel et al., 2020] created emotion lexicons for 91 languages for sentiment and emotion analysis. Other recent studies have focused on languages, such as Urdu for sentiment analysis [Mukhtar and Khan, 2019], Indonesian for sentiment analysis [Koto et al., 2020], Portuguese for sentiment and emotion analysis [Pereira, 2021], and Arabic for sentiment and emotion analysis [Alhumoud and Al Wazrah, 2021]. Studies on code-switched languages are also on the increase, with [Bansal et al., 2020] demonstrating how Hindi-English code-switching patterns from tweets can be used to improve sarcasm detection, and [Appidi et al., 2020] analysing code-switched Kannada-English from tweets for emotion classification.

In terms of modality, the visual modality is gaining more interest in the Social Opinion Mining research community. In [Akhtar et al., 2019], the authors propose a deep multi-task learning framework that carries out sentiment and emotion analysis from the textual, acoustic, and visual frames of video data obtained from YouTube. On the other hand, [Kumar and Garg, 2019] propose a multimodal sentiment analysis model for Twitter, where the sentiment polarity and strength is extracted from tweets based on their text and images (typographic and/or infographic).

More research has been published on aspect-based Social Opinion Mining, where [Jiang et al., 2020] focused on sentiment polarity in both single-aspect and multi-aspect scenarios, whereas [Hyun et al., 2020] focused on sentiment polarity in the automotive domain for the English and Korean languages.

The ongoing coronavirus disease (COVID-19) global pandemic has led to a rise in Social Opinion Mining studies analysing social opinions in terms of different dimensions, such as sentiment polarity. The work in [Müller et al., 2020] released a COVID-19 Transformer-based model that was pre-trained on multiple datasets of tweets from Twitter. These datasets contained tweets on various topics, such as vaccine sentiment and maternal vaccine stance, and used other well known datasets, such as SemEval 2016 - Task 4 which was previously discussed in Section 2.3.5. This model was pre-trained to carry out sentiment analysis on tweets written in other languages, such as Arabizi – a written form of spoken Arabic that relies on Latin characters and digits [Baert et al., 2020]. On the other hand, Kruspe et al. [Kruspe et al., 2020] presented sentiment analysis results of 4.6 million European tweets for the initial period of COVID-19 (December 2019 till April 2020), which results were aggregated by country (Italy, Spain, France, Germany, United Kingdom) and averaged over time. An ANN was trained to carry out sentiment analysis, which model was compared with several pre-trained models, such as the Bidirectional Encoder Representations from Transformers (BERT) which is trained on BookCorpus and English Wikipedia data [Devlin et al., 2018], a multilingual version of BERT trained on COVID-19 tweets [Müller et al., 2020], and the Embeddings from Language Models (ELMO) [Peters et al., 2018] trained on the 1 Billion Word Benchmark dataset.

In terms of NLP tools, Hugging Face⁷⁸ provides a state-of-the-art Transformer library for the Pytorch, TensorFlow, and JAX deep learning frameworks⁷⁹. In this regard, Hugging Face makes it possible to use Pytorch and TensorFlow to integrate out-of-the-box pre-trained and fine-tuned state-of-the-art models within a developed pipeline using a few lines of code. Therefore, it provides general-purpose architectures, such as BERT, GPT-2 [Radford et al., 2019], Robustly optimised BERT approach (RoBERTa) [Liu et al., 2019], cross-lingual language model (XLM) [Lample and Conneau, 2019], DistilBert [Sanh et al., 2019], and XLNET [Yang et al., 2019], for NLP tasks (like sentiment analysis), where over 32+ pre-trained models are available in 100+ languages. Similarly, TensorFlow Hub⁸⁰ provides a repository of trained machine learning models, with a variety of them using the

⁷⁸<https://huggingface.co/>

⁷⁹<https://huggingface.co/transformers/>

⁸⁰<https://www.tensorflow.org/hub>

Transformer architecture⁸¹, such as BERT.

The carbon footprint for training new deep learning models should always be taken in consideration especially if a large number of Central Processing Units (CPUs), Graphical Processing Units (GPUs), or Tensor Processing Units (TPUs) are needed. This in turn increases the related costs for model training, which is becoming very expensive and is expected to keep increasing in the future. In [Strubell et al., 2019], the authors mention that such costs amount to both the financial aspect in terms of hardware and electricity or cloud compute time, and the environmental aspect in terms of carbon footprint needed to fuel modern tensor processing hardware. Therefore, researchers should report the training time and computational resources needed in their published work, and they should prioritise computationally efficient algorithms and hardware that need less energy.

2.4 Social Opinion Dimensions

2.4.1 Context

An opinion describes a viewpoint or statement about a subjective matter. In many research problems, authors assume that an opinion is more specific and of a simpler definition. For example, sentiment analysis is considered to be a type of Opinion Mining even though it is only focused on extracting the sentiment score from a given text. Social data contains a wealth of signals to mine where opinions can be extracted over time. Different types of opinions require different modes of analysis [Agrawal et al., 2014]. This leads to opinions being multidimensional semantic artefacts. In fact, [Troussas et al., 2016] specify that “emotions and polarities are mutually influenced by each other, conditioning opinion intensities and emotional strengths”. Moreover, multiple studies applied different approaches, where the authors in [Bravo-Marquez et al., 2013] showed that a composition of polarity, emotion, and strength features, achieve significant improvements over single approaches, whereas [Koto and Adriani, 2015b] focused on finding the correlation between emotion –which can be differentiated by facial expression, voice intonation, and also words– and sentiment in social media. Similar in nature, the authors in [Buscaldi and

⁸¹https://tfhub.dev/google/collections/transformer_encoders_text

Hernandez-Farias, 2015] found out that finer-grained negative tweets potentially help in differing between negative feelings, e.g., fear (emotion).

Furthermore as discussed in [Porshnev and Redkin, 2014], mood, emotions, and decision-making are closely connected. Research on multidimensional sentiment analysis shows that the human mood is very rich in social media, where a piece of text may contain multiple moods, such as calm and agreement [Huang et al., 2015]. On the other hand, there are studies showing that one mood alone is already highly influential in encouraging people to rummage through Twitter feeds for predictive information. For example in [Weiss et al., 2015], “calmness” was highly correlated with stock market movement. Different opinion dimensions are also able to effect different entities, such as events. Results in [Zhang et al., 2012] show a strong correlation between emergent events and public moods. In such cases, new events can be identified by monitoring emotional vectors in microblogs. Moreover, work in [Thelwall et al., 2011] assessed if popular events are correlated with sentiment strength as it increases, which is likely to be the case.

This literature motivated us to pursue further research and development on the identification of different opinion dimensions that are present within social data, such as microblogs, published across heterogeneous social media platforms. A more fine-grained opinion representation and classification of this social data shall lead to a better understanding of the messages conveyed, thus potentially influencing multiple application areas.

2.4.2 Different Social Opinion Dimensions

The studies analysed in our systematic review focused on different opinion dimensions, namely: objectivity/subjectivity, sentiment polarity, emotion, affect, irony, sarcasm, and mood. These were conducted on different levels, such as document-level, sentence-level, and/or feature/aspect-based, depending on the study. In this regard, most of the evaluated studies focused on one social opinion dimension with sentiment analysis, specifically sentiment polarity classification being the most common. On the other hand 60 studies focused on more than one social opinion dimension, with 58 on two dimensions, 1 on three dimensions, and 1 on four dimensions. In this regard, Table 2.7 lists the different social opinion dimensions and the respective studies.

Dimensions	Studies
subjectivity and sentiment polarity	[Jiang et al., 2011, Blenn et al., 2012, Bravo-Marquez et al., 2013, Zhu et al., 2013, Wang and Ye, 2013, Cui et al., 2013, Li and Li, 2013, Rui et al., 2013, Bravo-Marquez et al., 2014, Tan et al., 2014, Garg and Chatterjee, 2014, Abdul-Mageed et al., 2014, Samoylov, 2014, Koto and Adriani, 2015b, Koto and Adriani, 2015c, Koto and Adriani, 2015a, Feng et al., 2015, Mansour et al., 2015, Wu et al., 2016, Zainuddin et al., 2016b, Er et al., 2016, Abdullah and Zolkepli, 2017, Hao et al., 2017, Ahuja and Dubey, 2017, Sahni et al., 2017, Moh et al., 2017, Dritsas et al., 2018, Gandhe et al., 2018, Nausheen and Begum, 2018]
sentiment polarity and emotion	[Cvijikj and Michahelles, 2011, Orellana-Rodriguez et al., 2013, Sheth et al., 2014, Yuan et al., 2015, Orellana-Rodriguez et al., 2015, Gallegos et al., 2016, Qaisi and Aljarah, 2016, Shukri et al., 2015, Munezero et al., 2015, Barapatre et al., 2016, Karyotis et al., 2017, Bouazizi and Ohtsuki, 2017, Radhika and Sankar, 2017, Abdullah and Hadzikadic, 2017, Zhang et al., 2017, Singh et al., 2018, Aoudi and Malik, 2018, Pai and Alathur, 2018, Ghosal et al., 2018, Rout et al., 2018, dos Santos et al., 2018, Stojanovski et al., 2018]
sentiment polarity and mood	[Bollen et al., 2011]
sentiment polarity and irony	[Reyes et al., 2013]
sentiment polarity and sarcasm	[Unankard et al., 2014]
sentiment polarity and affect	[Weichselbraun et al., 2017]
emotion and anger	[Delcea et al., 2014, Cotfas et al., 2015]
irony and sarcasm	[Fersini et al., 2015]
subjectivity, sentiment polarity, and emotion	[Jiang et al., 2015]
subjectivity, sentiment polarity, emotion, and irony	[Bosco et al., 2013]

Table 2.7: Studies focusing on two or more social opinion dimensions

The following sub-sections present the different social opinion dimensions mentioned above as identified in literature⁸².

2.4.2.1 Subjectivity

Subjectivity determines whether a sentence expresses an opinion –in terms of personal feelings or beliefs– or not, in which case a sentence expresses objectivity. Objectivity refers to sentences that express some factual information about the world [Liu, 2010].

In this domain, objective statements are usually classified as being neutral (in terms of polarity), whereas subjective statements are non-neutral. In the latter cases, sentiment analysis is performed to determine the polarity classification (more information on this is provided below). However, it is important to clarify that neutrality and objectivity are not the same. Neutrality refers to situations whereby a balanced view is taken, whereas objectivity refers to factual based i.e., true statements/facts that are quantifiable and measurable.

2.4.2.2 Sentiment Polarity

Sentiment determines the polarity (positive/negative/neutral) and strength/intensity (through a numeric rating score e.g., 1 to 5 stars, or level of depth e.g., low/high/medium) of an expressed opinion [Liu, 2010].

In some studies, such as [Sandoval-Almazan and Valle-Cruz, 2018, Bouazizi and Ohtsuki, 2017, Chou et al., 2017, Karyotis et al., 2017, Furini and Montangero, 2016, Gambino and Calvo, 2016, Jiang et al., 2015, Yuan et al., 2015], the sentiment polarity was derived from the emotion classification, such as, joy/love/surprise translated to positive, and anger/sadness/fear translated to negative.

2.4.2.3 Emotion

Emotion refers to a person’s subjective feelings and thoughts, such as love, joy, surprise, anger, sadness, and fear [Liu, 2010].

⁸²Note that some level categories are dependant on the domain.

A study [Munezero et al., 2015] mapped the observed emotions into two broad categories of enduring sentiments: ‘like’ and ‘dislike’. The former includes emotions that have a positive evaluation of the object, i.e., joy, trust, and anticipation, and the latter includes emotions that have a negative evaluation of the object, i.e., anger, fear, disgust, and sadness.

It is important to note that some of the emotion categories are based on published theories of emotion, with the most popular ones being Paul Ekman’s six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) [Ekman, 1992], and Plutchik’s eight primary emotions (anger, fear, sadness, disgust, surprise, anticipation, trust, and joy) [Plutchik, 1980].

Moreover, other studies have used emotion categories that are influenced from emotional state/psychological models, such as the Pleasure Arousal Dominance [Mehrabian, 1996], and the Ortony, Clore, and Collins (commonly referred to as OCC) [Ortony et al., 1988].

Several studies, such as [Xu et al., 2012, Furini and Montangero, 2016, Walha et al., 2016, Hubert et al., 2018] that targeted emotion classification incorrectly referred to such a task as sentiment analysis. Even though emotions and sentiment are highly related, the former are seen as enablers to the latter, i.e., an emotion/set of emotions affect the sentiment.

2.4.2.4 Affect

Affect refers to a set of observable manifestations of a subjectively experienced emotion. The basic tasks of affective computing are emotion recognition and sentiment polarity detection [Cambria, 2016].

The affect categories are based on the “Hourglass of Emotions”, which was inspired by Plutchik’s studies on human emotions. When using this affective model, sentiment is based on four independent dimensions, namely Pleasantness, Attention, Sensitivity, and Aptitude. The different levels of activation of these dimensions constitute the total emotional state of the mind [Hussain and Cambria, 2018]. The semi-supervised learning model proposed by [Hussain and Cambria, 2018] based on the merged use of multidimensional

scaling by means of random projections and biased SVM, has been exploited for the inference of semantics and sentics (conceptual and affective information). The latter are linked with concepts in a multidimensional vector space, in accordance with this affective model. This is used to carry out sentiment polarity detection and emotion recognition in cases when there is a lack of labelled common-sense data.

2.4.2.5 Irony

Irony is used to convey the opposite meaning of the actual things one says, but its purpose is not intended to hurt another person [Mer, a].

2.4.2.6 Sarcasm

Sarcasm holds the “characteristic” of meaning the opposite of what one says, but unlike irony, it is used to hurt the other person towards whom the sarcasm is directed [Mer, c].

2.4.2.7 Mood

Mood refers to a conscious state of mind or predominant emotional state of person or atmosphere of groups, people, or places, at any point in time [Mer, b]. Some of the mood categories are based on the profile of mood states (POMS) Bipolar questionnaire [McNair and Droppleman, 1971] which is designed by psychologists to assess human mood states, or on GPOMS [Bollen et al., 2011] which expands the POMS Bipolar questionnaire to capture a wider variety of naturally occurring mood terms in tweets.

2.4.2.8 Aggressiveness

The authors in [Del Bosque and Garza, 2014] assume that **aggressive** text detection is a sub-task of sentiment analysis, which is closely related to document polarity detection. Their reasoning is that aggressive text can be seen as intrinsically negative.

2.4.3 Impact of Sarcasm and Irony on Social Opinions

Sarcasm and irony are two social opinion dimensions from the ones listed in Section 2.4.2 which are worth further analysis. Sarcasm and irony are often confused and/or misused.

This leads to their classification in being very difficult even for humans [Unankard et al., 2014, Buscaldi and Hernandez-Farias, 2015], with most users holding negative views on such messages [Unankard et al., 2014]. The study by [Buscaldi and Hernandez-Farias, 2015] is a relevant example, whereby a large number of false positives were identified in the tweets classified as ironic. Moreover, such tasks are also very time consuming and labour intensive particularly with the rapid growth in volume of online social data. Therefore, not many studies focused and/or catered for sarcasm and/or irony detection.

2.4.3.1 Challenges

The majority of the reviewed proposed approaches are not equipped to cater for traditional limitations, such as negation effects or ironic phenomena in text [Castellucci et al., 2015a]. Such Opinion Mining tasks face several challenges, with the main ones being:

- Different languages and cultures result in various ways of how an opinion is expressed on certain social media platforms. For example, Sina Weibo users prefer to use irony when expressing negative polarity [Wang et al., 2014b]. Future research is required for the development of cross-lingual/multilingual NLP tools that are able to identify irony and sarcasm [Yan et al., 2014].
- Presence of sarcasm and irony in social data, such as tweets, may affect the feature values of certain machine learning algorithms. Therefore, further advancement is required in the techniques used for handling sarcastic and ironic tweets [Pandey et al., 2017]. The work in [Sarsam et al., 2020] addresses the main challenges faced for sarcasm detection in Twitter and the machine learning algorithms that can be used in this regard.
- Classifying/rating a given sentence's sentiment is very difficult and ambiguous, since people often use negative words to be humorous or sarcastic.
- Sarcasm and/or irony annotation is very hard for humans and thus it should be presented to multiple persons for accuracy purposes. This makes it very challenging to collect large datasets that can be used for supervised learning, with the only possible way being to hire people to carry out such annotations [D'Asaro et al.,

2017]. Moreover, the differentiation between sarcasm and irony by human annotators results in a lack of available datasets and datasets with enough examples of ironic and/or sarcastic annotations. Such datasets are usually needed for “data hungry” computational learning methods [Sykora et al., 2020].

2.4.3.2 Observations

Table 2.8 lists the studies within the review analysis that focused on sarcasm and/or irony. These account for only 18 out of 465 reviewed papers. One can clearly note the research gap that exists within these research areas.

Sarcasm	Irony	Studies
✓		[Baccouche et al., 2018, Bouazizi and Ohtsuki, 2018, Ghiassi and Lee, 2018, Abdullah and Zolkepli, 2017, Bouazizi and Ohtsuki, 2017, Caschera et al., 2016, Tan et al., 2014, Unankard et al., 2014, Mejova et al., 2013, Bakliwal et al., 2013, Mejova and Srinivasan, 2012, Wang et al., 2012]
	✓	[Buscaldi and Hernandez-Farias, 2015, Hernandez-Farias et al., 2014, Bosco et al., 2013, Reyes et al., 2013]
✓	✓	[Fersini et al., 2015, Pandey et al., 2017]

Table 2.8: Studies adopting sarcasm and/or irony

The following is an overview of the studies’ main results and observations:

- [Bosco et al., 2013]: The authors found that irony is normally used together with a positive statement to express a negative statement, but seldomly the other way. Analysis shows that the Senti-TUT⁸³ corpus can be representative of a wide range of irony phenomena from bitter sarcasm to genteel irony.
- [Reyes et al., 2013]: The study describes a number of textual features used to identify irony at a linguistic level. These are mostly applicable for short texts, such as tweets. The developed irony detection model is evaluated in terms of representativeness and relevance. Authors also mention that there are overlaps in occurrences of irony, satire, parody, and sarcasm, with their main differentiators being tied to usage, tone, and obviousness.

⁸³<http://www.di.unito.it/~tutreeb/sentiTUT.html>

- [Mejova et al., 2013]: A multi-stage data-driven political sentiment classifier is proposed in this study. The authors found “that a humorous tweet is 76.7% likely to also be sarcastic”, whereas “sarcastic tweets are only 26.2% likely to be humorous”. Future work is required on the connection between sarcasm and humour.
- [Fersini et al., 2015]: Addresses the automatic detection of sarcasm and irony by introducing an ensemble approach based on Bayesian Model Averaging, that takes into account several classifiers according to their reliabilities and their marginal probability predictions. Results show that not all the features are equally able to characterise sarcasm and irony, whereby sarcasm is better characterised by POS tags, and ironic statements by pragmatic particles (such as emoticons and emphatic/onomatopoeic expressions, which represent those linguistic elements typically used in social media to convey a particular message).
- [Jiang et al., 2015]: The authors’ model classifies subjectivity, polarity, and emotion in microblogs. Results show that emoticons are a pure carrier of sentiment, whereas sentiment words have more complex senses and contexts, such as negations and irony.
- [Wang et al., 2012]: Post-facto analysis of user-generated content, such as tweets, show that political tweets tend to be quite sarcastic.
- [Ghiassi and Lee, 2018]: Certain keywords or hash-tagged words (e.g., “thanks”, “#smh”, “ #not”) that follow certain negative or positive sentiment markers in textual social data, usually indicate the presence of sarcasm.

The challenges faced in sarcasm and irony detection within context of Opinion Mining as discussed in Section 2.4.3.1 have motivated us to pursue further research in this direction.

2.5 Application Areas

Around half of the studies analysed as part of the systematic review discussed in Section 2.2 focused their Social Opinion Mining work on a particular real-world application area (or multiple ones), with Figure 2.2 showing the main areas. Note that each circle represents

an application area, where the size reflects the number of studies within the particular application area. The smallest circles represent a minimum of two studies that pertain to the respective application area, whereas the biggest circle reflects the most popular application area. Intersecting circles represent application areas that were identified as being related to each other based on the analysis conducted.

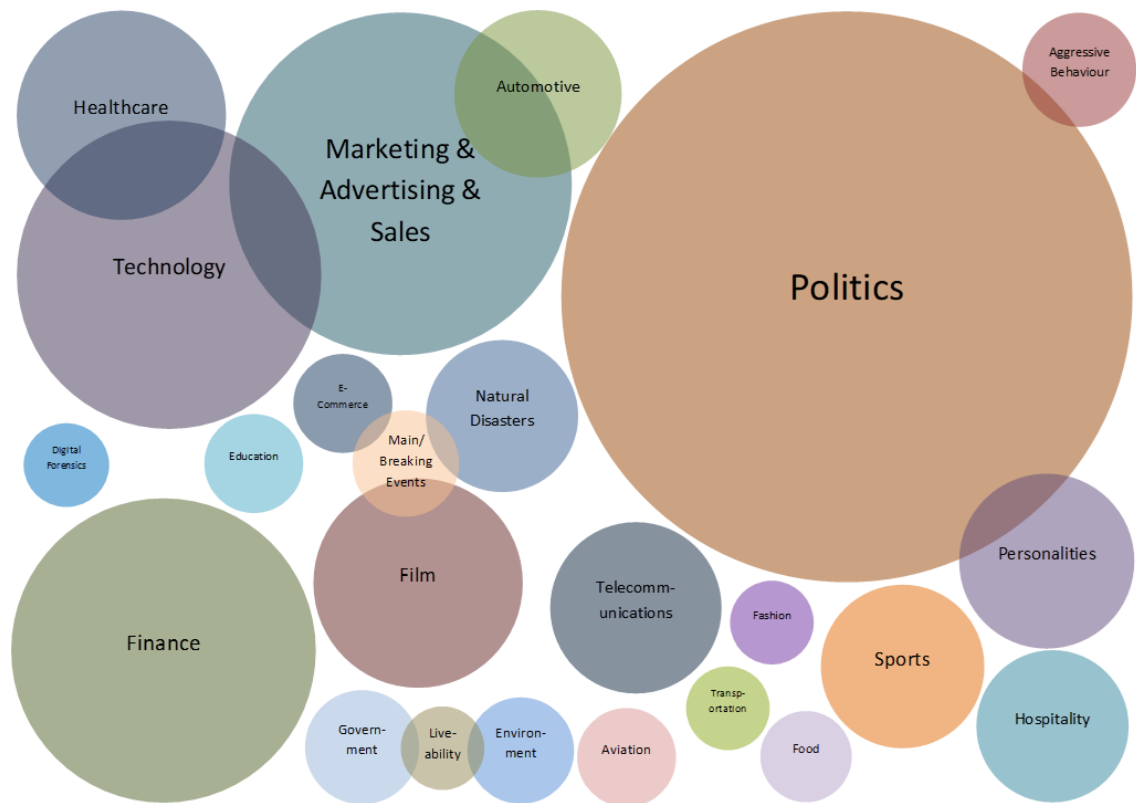


Figure 2.2: Application Areas

The **Politics** domain is the dominant application area with 45 studies applying Social Opinion Mining on different events, namely elections, reforms such as equality marriage, debates, referendums, political parties or politicians, and political events such as terrorism, protests, uprisings, and riots. In terms of **Marketing & Advertising & Sales**, 29 studies focused on brand/product management and/or awareness, products/services in general, local marketing, and online advertising.

The **Technology** industry-oriented studies (23) focused on: company perception, products such as mobile/smart phones, laptops, electronics, tablets, operating systems, cloud service providers, social media providers, and multiple technologies. All the 21

studies targeting the **Finance** domain applied Social Opinion Mining on demonitisation, currencies, and the stock market for risk management and predictive analytics.

Thirteen studies applied Social Opinion Mining on the **Film** industry for recommendations, box office predictions, or from a general perspective. Similarly, 13 studies focused on **Healthcare**, namely on epidemics/infectious diseases, drugs, hospitals, vaccines, public health such as epidemics, clinical science, mental health, and in general such as health-related tweets and health applications.

In terms of other industries, Social Opinion Mining was applied within the following:

- **Telecommunications** (e.g., telephony, television) on particular service providers;
- **Automotive** on car brand and/or car model perceptions and for vehicle sales predictions;
- **Hospitality** for restaurant recommendations and hotel/resort perceptions;
- **Aviation** on specific airline services e.g., customer relationship management, and air crashes;
- **Food** either in general or on safety; and
- **Fashion** on fast fashion retail company perception.

In terms of domains, the studies focused on:

- **Sports** on football/soccer, american football, basketball, cricket, and olympics;
- **Government** for smart cities and e-Government;
- **Environment** for policy-makers, urban mobility, wind energy, green initiatives, and peatland fires;
- **E-commerce** for product recommendations, crisis management, decision-making, and policy-making;
- **Education** for e-learning and on universities; and
- **Transportation** for ride hailing services and logistics and traffic conditions.

Moreover, other studies applied Social Opinion Mining in the following areas:

- **Personalities** on celebrity public perception;
- **Natural Disasters** on earthquakes, explosions, and in general;
- **Aggressive Behaviour** in relation to crime, cyberbullying, bullying, violence, and disorder;
- **Main/Breaking Events** such as Black Friday, Oscars, TV shows, product launch, earthquake, accidents e.g., shootings, and in general;
- **Liveability** in terms of place design to supports local authorities, urban designers and city planners, and government services such as welfare; and
- **Digital Forensics** for forensic analysts when investigating suspects.

Lastly, 19 further studies –not represented in Figure 2.2– focused on the following application areas: Human Development, Human Mobility, Public Facilities, Smart Cities, Web Publishing, Sponsorships, Countries, Industry, Entertainment, Refugee/Migrant crisis, Tourism, Music, Cryptocurrency, Economy, Social Issues, Law, Insurance/Social Security, Geographic Information, and Social Interactions.

Recent studies not covered within the systematic review show that the application areas above are still very popular, with research in new sub-domains emerging. In particular, several studies such as [Kapočiūtė-Dzikiene et al., 2019, Cresci et al., 2019, Guo and Li, 2019, Xing et al., 2020, Chen et al., 2020, Mishev et al., 2020] focus on the Finance domain. The authors in [Xing et al., 2020] identify common error patterns that cause financial sentiment analysis to fail, namely unrealistic mood, rhetoric, dependent opinion, unspecified aspects, unrecognised words, and external reference. On the other hand, in [Mishev et al., 2020] the authors evaluate sentiment analysis studies in the Finance domain by starting from lexicon-based approaches and finishes with the ones that use Transformers, such as BERT and RoBERTa.

2.6 Research Gaps Conclusions

Shared evaluation tasks such as SemEval usually focus on several research gaps that have been identified for Social Opinion Mining (such as the ones discussed in this section). These research gaps are very important to gather interest in this research area and shall help contribute to the advancement of Social Opinion Mining. Therefore, researchers are encouraged to engage in these tasks through their participation and/or organisation of new tasks. Moreover, the literature analysis carried out in this chapter targets different audiences, namely:

- Early-Stage Researchers who are interested in working within this evolving research field of study and/or are looking for an overview of this field;
- Experienced Researchers already working in Social Opinion Mining who would like to progress further on the technical side of their work and/or looking for weaknesses in the field of Social Opinion Mining; and
- Early-Stage and/or Experienced Researchers who are looking into applying Social Opinion Mining/their Social Opinion Mining work in a real-world application area.

The identification of the current literature gaps within the Social Opinion Mining field of study is one of the main contributions of this thesis. An overview below provides a pathway to future research and development work:

- **Social Media Platforms:** Most studies focus on data gathered from one social media platform, with Twitter being the most popular followed by Sina Weibo for Chinese targeted studies. It is encouraged to possibly explore information gathered from multiple data sources, subject to any existing API limitations⁸⁴. This should increase the variety and volume of data (two of the V's of Big Data) used for evaluation purposes, thus ensuring that results provide a more reflective picture of society in terms of opinions. The use of multiple data sources for studies focusing on the same real-world application areas are also beneficial for comparison purposes

⁸⁴Due to GDPR, API coverage in terms of which data can be accessed is being tightened in terms of control which can be a major issue faced by researchers.

and identification of any potential common traits, patterns, and/or results. Mining opinions from multiple sources of information also presents several advantages, such as higher authenticity, reduced ambiguity, and greater availability [Balazs and Velásquez, 2016].

- **Techniques:** The use of deep learning, statistical, probabilistic, ontology, and graph-based approaches should be further explored both as standalone and/or part of hybrid techniques, due to their potential and accessibility. In particular, the capabilities of deep learning have made several applications feasible, whereas ontologies and graph mining enable fine-grained Opinion Mining and the identification of relationships between opinions and their enablers (person, organisation, etc.). Moreover, ensemble machine learning and deep learning methods and fine-tuned Transformer-based models are still under-explored. In such a case, researchers should be attentive to the carbon footprint needed to train neural network models for NLP.
- **Language:** The majority of the studies support one language, with English and Chinese being the most popular. Studies that supporting two or more languages provides a major challenge in this domain due to numerous factors, such as cultural differences and lack of language-specific resources, e.g., lexicons, datasets, tools, and technologies. This domain also needs more studies that focus on code-switched languages and less-resourced languages, which shall enable the development of certain language resources needed for the respective code-switched and less-resourced languages.
- **Modality:** Bimodal/Multimodal Social Opinion Mining is another sub-domain that requires more research. Several studies cater for the text modality only, with the visual - image modality gaining more popularity. However, the visual - video and audio modalities are still in their early research phases with several aspects still unexplored. This also stems from a lack of available visual, audio, and multimodal datasets.
- **Aspect-based Social Opinion Mining:** Research in this sub-domain is increasing and developing, however, it is far from the finished article, especially when applied

in certain application areas. Further aspect-based research is encouraged on other opinion dimensions other than sentiment polarity, such as emotions and moods, which are still unexplored. Moreover, the use of deep learning approaches has the potential of advancing research in this direction.

- **Application areas:** Most studies target Politics, Marketing, Advertising, Sales, Technology, Finance, Film, and Healthcare. Research into other areas/sub-domains is encouraged to study and show the potential of Social Opinion Mining.
- **Social Datasets:** The majority of available datasets are either English or Chinese specific. This domain needs further social datasets published under a common open license for use by the public domain. These should target any of the following criteria: bilingual/multilingual data, and/or annotations of multiple opinion dimensions within the data, e.g., sentiment polarity, emotion, sarcasm, irony, mood, etc. Both requirements are costly in terms of resources (time, funding, and personnel), domain knowledge, and expertise.
- **Dimensions of Social Opinion Mining:** Most studies focus on subjectivity detection and sentiment analysis. The area of emotion analysis is increasing in popularity, however, sarcasm detection, irony detection, and mood analysis are still in their early research phases. Moreover, from the analysis of this literature it is evident that there is a lack of research on any possible correlations between the different opinion dimensions, e.g., emotions and sentiment. Lastly, no studies cater for all the different and/or majority of the Social Opinion Mining dimensions mentioned within Section 2.4.2 i.e., subjectivity, sentiment polarity, emotion, affect, irony, sarcasm, mood, and aggressiveness, within their work.

2.7 Conclusion

This chapter identifies, analyses, and evaluates existing Opinion Mining approaches that make use of social data in terms of social media platforms, techniques, social datasets, language, modality, tools and technologies, NLP tasks including aspect-based Social Opinion

Mining, and other aspects identified within literature. Moreover, the different Social Opinion Mining dimensions, namely subjectivity, sentiment polarity, emotion, affect, irony, and sarcasm were identified from the analysed Opinion Mining solutions, together with their main application areas, such as Politics, Marketing, Advertising, Sales, Technology, and Finance. The latest research developments, advancements, and current literature gaps within the Social Opinion Mining research area have also been identified and provide a pathway for future research directions within this field of study.

The content of this chapter answers **Research Question 1**, where the direction of this thesis is based on the identified research gaps listed in Section 2.6. In this regard, the research questions of this thesis are related to the following research gaps outlined: the need for aspect-based Social Opinion Mining studies that cater for multiple social opinion dimensions (**Research Questions 1, 4, 5**), research that determines whether Social Opinion Mining can influence multiple application areas (**Research Question 2**), research focused on code-switched languages e.g., Maltese-English and less-resourced languages e.g., Maltese (**Research Question 3**), collection of new social datasets from more than one data source for a real-world application area that target bilingual/multilingual data which data is annotated for multiple opinion dimensions (**Research Questions 3, 5**), and the need to develop Social Opinion Mining classification techniques based on machine learning and deep learning approaches (**Research Question 3**). Further literature for social opinion classification and opinion summarisation is discussed in Chapters 4 and 5 respectively, as relevant. The next chapter targets whether this field of study can influence multiple application areas and how the research gaps of multidimensional and multilingual social opinion data can be tackled.

Chapter 3

Social Opinion Data

3.1 Introduction

Even though a number of social datasets are available for conducting Opinion Mining research, there is still a lack of datasets that can be used to carry out Social Opinion Mining, especially in the context of exploring whether this research area can influence multiple application areas. Moreover, the available datasets are usually limited to one social opinion dimension namely sentiment polarity, they gather data from one social data source, they support one language mostly English, and they target the most common application areas such as Marketing, Advertising, and Sales.

This section describes three datasets that were generated towards answering the research questions of this thesis. These add to the social datasets discussed in Section 2.3.5.

In response to Research Question 2, three datasets spanning different domains, namely Politics (targeting Referendums), Finance (targeting Stock Trading), and Socio-Economic and Government (targeting Government Budgets), were generated to determine how Social Opinion Mining can influence multiple application areas.

In terms of Social Opinion Mining dimensions, the Political dataset (dataset 1 - refer to Section 3.2) focuses on sentiment in terms of polarity and strength, the Financial dataset (dataset 2 - refer to Section 3.3) focuses on sentiment in terms of score, and the Socio-Economic and Government dataset (dataset 3 - refer to Section 3.4) focuses on subjectivity, sentiment polarity, emotion, irony, and sarcasm.

Based on the current Social Opinion Mining research gaps identified in Section 2.6 and the main ones mentioned above, the three datasets discussed in this chapter contribute to the following Social Opinion Mining research gaps:

- **Social Datasets and Social Media Platforms:** Gathering of social data from more than one social data source, with dataset 2 (more than three data sources) and dataset 3 (four data sources) collecting data from social networking services and newswires;
- **Social Datasets and Application Areas:** Collection of a new social dataset for a real-world application area that is not common, with dataset 1 focusing on Referendums within the Political domain, and dataset 3 focusing on Government Budgets within the Socio-Economic and Government domains which in itself caters for multiple domains, such as Transport, Environment, and Taxation;
- **Social Datasets and Techniques:** Publishing of social datasets under a common open license, with all three datasets satisfying this criteria so that they can be used for further research purposes;
- **Social Datasets and Language:** Collection of datasets that target bilingual/multilingual data, where dataset 3 targets bilingual data (Maltese and English), a code-switched language (Maltese-English), and a less-resourced language (Maltese); and
- **Social Datasets and Aspect-based Social Opinion Mining and Dimensions of Social Opinion Mining:** Dataset that caters for multiple social opinion dimensions at an aspect-level (i.e., based on topic), with dataset 3 catering for five social opinion dimensions, namely subjectivity, sentiment polarity, emotion, irony, and sarcasm.

Table 3.1 presents an overview of the social datasets generated for the purposes of this thesis. Even through the catering of multiple social opinion dimensions was identified as a research gap, following the existing literature on the research area, the generated datasets cater for sentiment polarity as a common social opinion dimension. This allows for their

evaluation in terms of the influence of Social Opinion Mining in multiple application areas. Nevertheless, dataset 3 contributes towards this research gap by catering for multiple social opinion dimensions, as detailed further below.

	Dataset 1	Dataset 2	Dataset 3
Application Area	Politics - targeting Referendums	Finance - targeting Stock Trading	Socio-Economic and Government - targeting Government Budgets
Data Sources	Twitter	Twitter, Stock-Twits, News Statements and Headlines (e.g., Yahoo Finance)	Twitter, Times of Malta, MaltaToday, The Malta Independent
Open License	Yes	Yes	Yes
Languages	English	English	English, Maltese, Maltese-English
Aspect-level	No	No	Yes
Social Opinion Dimensions	Sentiment - Polarity and Strength	Sentiment Polarity (Score)	Subjectivity, Sentiment Polarity, Emotion, Irony, Sarcasm

Table 3.1: Three social datasets generated for the purposes of this thesis

The generation of the above-mentioned three social datasets enabled us to understand the importance of multiple social opinion dimensions that target multilingual data, which is currently largely unexplored in the Social Opinion Mining research area. For this reason, dataset 1 and dataset 2 are not used within the rest of this thesis, where priority was given to the multidimensional aspect to cater for the lack of multidimensional datasets for Social Opinion Mining purposes and to identify whether this can affect a particular domain and to what extent. Opinions expressed by humans are multidimensional semantic artefacts that are mutually influenced by each other e.g., emotions and sentiment polarities. In this regard, multiple social opinion dimensions contribute to the formulation of a more context-rich and fine-grained representation of the real-life complexity of human opinions. As a result, the dataset of Multidimensional and Multilingual Social Opinions for Malta’s Annual Government Budget presented in Section 3.4 was generated (dataset 3). This dataset caters for all of the main research gaps identified above and forms the basis for the rest of the work carried out in this thesis.

The sections of this chapter are in part based on the following publications:

- Manuela Hürlimann, Brian Davis, **Keith Cortis**, André Freitas, Siegfried Handschuh, Sergio Fernández. “A twitter sentiment gold standard for the brexit referendum.” In Proceedings of the 12th international conference on semantic systems, pp. 193-196. 2016. (Section 3.2)
- **Keith Cortis**, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, Brian Davis. “Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news.” Association for Computational Linguistics (ACL), 2017. (Section 3.3)
- **Keith Cortis**, Brian Davis. “A social opinion gold standard for the Malta government budget 2018.” In Proceedings of the 5th workshop on noisy user-generated text (W-NUT 2019), pp. 364-369. 2019. (Section 3.4)
- **Keith Cortis**, Brian Davis. “A dataset of multidimensional and multilingual social opinions for malta’s annual government budget.” In Proceedings of the International AAAI Conference on Web and Social Media, vol. 15, pp. 971-981. 2021. (Section 3.4)

3.2 Dataset 1: Twitter Sentiment Gold Standard for the Brexit Referendum

This section presents a sentiment-annotated Twitter gold standard targeting the historical event of the 2016 United Kingdom European Union membership referendum (“Brexit” referendum).

3.2.1 Generation Process

This section describes the multi-stage process used for building the dataset, namely the sampling strategy, methods used for data collection, the annotation process of the dataset, and the data quality measures carried out to consolidate the final dataset.

3.2.1.1 Sampling Strategy

In order to collect the dataset, a random sample of 2,000 tweets from a Twitter stream was set up to track 75 keywords, including hashtags and account names. Criteria for chosen keywords were based on the manual identification of common keywords associated with content relevant to Brexit, such as #eureferendum, #votein, #voteleave, #Brexit, and #brexitfears.

3.2.1.2 Data Collection

Data collection on this stream between May 4 and May 6, 2016 (inclusive) resulted in a population of 149,331 tweets. Before sampling, filters were applied to exclude spam and irrelevant content (discarding for example, very short content and tweeters with suspiciously high activity). Furthermore, only tweets published between 6am and 11pm GMT were considered in order to increase coverage of European postings. These measures reduced the population for sampling to 20,104 tweets.

3.2.1.3 Annotation

The 2,000 tweets sampled were presented to three raters, all proficient in English, who created the following annotations for each tweet:

1. **Sentiment:** assign one of the following values (or leave blank if one cannot decide):
 - *Stay*: the tweet is in favour of the UK remaining in the EU;
 - *Leave*: the tweet is in favour of the UK leaving the EU;
 - *Undecided*: the tweet expresses indecision about the EU referendum;
 - *Don't care/no sentiment*: the tweeter does not care about the outcome of the referendum, or there is no sentiment expressed; and
 - *Irrelevant*: the tweet is not about Brexit;
2. **Strength** (only for tweets classified as “stay” or “leave”): an integer value between 1 (very weak) and 5 (very strong) expressing the strength of the “stay” or “leave” sentiment; and

3. **Contextual dependency**: numerical value with 0 referring to an interpretation of sentiment in a tweet which does not depend on external sources, and 1 referring to an interpretation of sentiment in a tweet which depends on external sources (e.g., articles or images that are linked).

The five opinion (sentiment polarity-based) categories and strength annotations support a fine-grained view on the opinion landscape. Furthermore, the contextual dependency option provides an indication of the difficulty of scoring a tweet, which is a fundamental feature for the construction of Opinion Mining classifiers.

3.2.1.4 Quality

Table 3.2 presents the standard inter-rater reliability agreement scores for each of the annotations.

Annotation	Fleiss' kappa [Fleiss and Cohen, 1973]	Observed Agreement
Sentiment	0.394	0.537
Strength	0.240	0.439
Context	0.156	0.748

Table 3.2: Inter-rater reliability measures for each annotation type

A moderate Fleiss' agreement is achieved for sentiment and strength, and fair agreement for context. Average observed agreement gives an indication of the difficulty of this annotation task. The strength assignment is the most difficult, while context dependency is relatively straightforward to determine.

3.2.1.5 Consolidation

Table 3.3 shows the distribution of tweets with regard to the number of annotators who agreed on its opinion (sentiment polarity-based) annotation, providing a different view of agreement.

The annotation consolidation procedure is based on the three categories of agreement presented in Table 3.2, where the following conditions were taken in consideration: a) a majority vote for the opinion (sentiment-based) and contextual dependency annotations, and b) the average (rounded to the nearest integer) for the strength annotation for the

first two rows (unanimous and two different opinions). Cases where three different options were selected by the annotators were consolidated manually by a fourth rater who was not previously involved in the annotation process.

Agreement on “sentiment” annotation	Number of tweets	%
Unanimous	785	39.25
Two different opinions	870	43.50
Three different opinions	345	17.25
Total	2,000	100

Table 3.3: Agreement between annotators

3.2.2 Statistics and Discussion

The consolidated gold standard consists of a total of 2,000 tweets. The distribution of opinion annotations can be seen in Table 3.4.

Annotation	Number of tweets	%
stay	430	21.5
leave	816	40.8
undecided	82	4.1
no sentiment/don’t care	502	25.1
irrelevant	148	7.4
[left blank]	22	1.1

Table 3.4: Distribution of sentiment annotations

The large number of “leave” tweets in the dataset reflects the overall tweeting behaviour, as also identified by [Ontotext, 2016]. Very few tweets display an ‘undecided’ sentiment, in line with observations by [Barberá, 2015] that strong opinions predominate on Twitter. Our data displays a similar bimodal distribution if the “stay”, “leave”, and “undecided” annotations are considered.

The low percentage of irrelevant tweets shows the usefulness of tracking keywords used for retrieving content which is relevant to Brexit. Many of these irrelevant tweets are in languages other than English. Some hashtags are used ambiguously, e.g., “#takecontrol” is used in the context of yoga rather than the “Leave” campaign who coined the phrase. A rather large number of tweets were classified as not displaying any sentiment. These include tweets asking questions on the impact of Brexit or expressing non-opinionated

interest in the issue.

Table 3.5 shows the distribution of opinion strength annotations in the dataset. One can note that there is a greater tendency for “leave” tweets to display strong opinions, while the “stay” opinions tend towards the weaker end of the scale. Both opinions, however, span the entire continuum.

Strength	Number of “stay” tweets	% of “stay” tweets	Number of “leave” tweets	% of “leave” tweets
1	88	20.5	131	16.1
2	193	44.9	304	37.3
3	135	31.4	263	32.2
4	13	3.0	112	13.7
5	1	0.2	6	0.7

Table 3.5: Distribution of strength annotations

Table 3.6 shows a breakdown of context dependence by opinion annotation (sentiment polarity-based). In total, 268 tweets (13.4%) were annotated as depending on context, whereas 1,732 tweets (86.6%) were annotated as not depending on context. Moreover, it can be noted that the percentage of context-dependent tweets is rather stable across all opinion categories.

Annotation	Number of context-dependent tweets (%)	Number of context-independent tweets (%)	Total number of tweets
stay	59 (13.7%)	371 (86.3%)	430
leave	134 (16.4%)	682 (83.6%)	816
undecided	15 (18.3%)	67 (81.7%)	82
no sentiment/ don’t care	47 (9.4%)	455 (90.6%)	502
irrelevant	11 (7.4%)	137 (92.6%)	148
[left blank]	2 (9.1%)	20 (90.9%)	22
total	268 (13.4%)	1,732 (86.6%)	2,000

Table 3.6: Distribution of context dependence annotations

3.2.3 Data Reuse

The dataset has been published¹ under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) licence² for general use. All of the data from Twitter (tweets, creation dates, tweet ids) are covered by Twitter’s Terms of Service³.

3.2.4 Applicability

Referendum-type events, such as Brexit, require the application of different techniques and resources for Opinion Mining since these events have distinctive social dynamics and political discourse. The availability of language resources to ground the discourse analysis and the construction of supervised classification methods e.g., to carry out sentiment analysis, play a fundamental role for pushing forward the ability to develop systems that can support the interpretation of social media discourse. Therefore, the Brexit Sentiment Gold Standard presented can support the evolution of classification methods for sentiment analysis. Moreover, from a discourse perspective, this dataset provides a resource for observing the social and discourse dynamics behind referendum-type events, such as Brexit. This contrasts to most political corpora which as core discourse targets have politicians and parties.

3.2.5 Impact and Limitations

Popular referenda provide a rich setting for understanding the social and discourse dynamics behind a focused political discussion. Under these settings, Opinion Mining, such as sentiment analysis, over social media data provides a fundamental tool to provide systematic prospective and retrospective insights, supporting an analysis of the underlying political processes and dynamics at stake.

In terms of impact, this dataset is a contribution in a novel domain, i.e., a historical event that affected the United Kingdom in terms of economic impact and is currently influencing the United Kingdom-European Union relations. Moreover, it has influenced several Opinion Mining studies in the Politics domain such as [Jones et al., 2017, Agarwal

¹<https://bitbucket.org/ssix-project/brexit-gold-standard>

²<https://creativecommons.org/licenses/by-sa/4.0/>

³<https://twitter.com/en/tos>

et al., 2018, Georgiadou et al., 2020], research focusing on the 2016 Brexit referendum such as [Mancosu and Bobba, 2019, del Gobbo et al., 2021], sentiment analysis research targeting languages such as Bengali [Islam et al., 2020], and research focusing on application areas and domains such as the Common Core State Standards [Supovitz et al., 2018] and Political Event Analysis [Korakakis et al., 2017], Big Data Analysis such as [Andrešić et al., 2017], and other research areas such as Data Quality Assessment [Salvatore et al., 2021] and Link Prediction [Sotiropoulos et al., 2019].

With respect to the research gaps identified in Section 2.6, this dataset has some limitations given that the social data was collected from only one social data source in Twitter, supports only the English language, and is only annotated for the sentiment dimension in terms of polarity and strength. In light of these limitations, the direction of this thesis focused on the research gaps of multidimensional and multilingual social opinion data, where the Socio-Economic and Government dataset presented in Section 3.4 tackles all these gaps.

3.3 Dataset 2: Fine-grained Sentiment Analysis on Financial Microblogs and News Statements and Headlines

This section presents two datasets for fine-grained sentiment analysis, one collected from Microblogs and the second one covering News Statements and Headlines.

1. **Microblogs** were derived from two data sources:
 - (a) *StockTwits messages*: Consists of microblogs focusing on stock market events and assessments from investors and traders, exchanged via the StockTwits microblogging platform. Typical messages consist of references to company stock symbols (so-called cashtags - a stock symbol preceded by “\$”, e.g. “\$AAPL” for the company Apple Inc.), a short supporting text, or references to a link or pictures (typically containing charts showing stock values analysis).
 - (b) *Twitter messages*: Some stock market discussion also takes place on the Twitter platform. Twitter online posts containing company stock symbols (cashtags) were extracted.

- 2. News Statements and Headlines:** Sentences have been taken from news headlines as well as news text. The textual content was crawled from different sources on the Internet, such as Yahoo Finance⁴. These sentences were based on company names and abbreviations, as cashtags are not typically used in news statements and headlines.

These datasets were generated for the “Fine-Grained Sentiment Analysis on Financial Microblogs and News” task as part of SemEval-2017, specifically under the “Detecting sentiment, humour, and truth” theme. This task consisted of two tracks, with the first one about Microblogs (sub-task 1) and the second one (sub-task 2) about News Statements and Headlines. The main goal behind both tracks was to predict the sentiment score for each of the mentioned companies/stocks. The sentiment scores for each text instance adopted floating point values in the range of -1 (very negative/bearish) to 1 (very positive/bullish), with 0 designating a neutral sentiment.

3.3.1 Generation Process

This section describes the multi-stage process used for building each of the two datasets, namely the sampling strategy, methods used for data collection, the annotation process of each dataset, and the data quality measures carried out to consolidate the final datasets.

3.3.1.1 Sampling Strategy

The corpus of statements was created by conducting random sampling and an initial filtering process over a pool of StockTwits messages, tweets, and Really Simple Syndication (RSS) news feeds. While the random sampling ensured an unbiased set of statements, the filtering mechanism aimed at removing messages from the set microblogs which are spam. The filtering mechanism was based on a manual curation of the set of microblog users which are classified as spammers. The goal of data sampling is to come up with a most representative and manageable amount of data for manual annotation.

The first step was to apply a stratified random sampling by objects δ per the smallest time unit level θ determined (i.e., stock messages per day) to ensure that all different

⁴<http://finance.yahoo.com/>

objects are adequately represented in the sample with respect to their distribution in the population. Then, the random samples of a time-unit level θ_i are pooled into a time-unit level θ_{i+1} and randomly sampled.

The purpose of re-sampling at different time-unit levels is to make the resulted random sample more random, more balanced, and more representative of the entire time-span of the data. A general negative sentiment in a certain sub-sample will be counterbalanced by the other sub-samples.

3.3.1.2 Data Collection

StockTwits data was provided by StockTwits in a batch export and refers to the period from October 2011 to June 2015. The original pool before sampling contained 27 million StockTwits, from which 1,847 messages were sampled. Twitter data was collected between March 11th and 18th 2016 using the official Streaming APIs. Sampling was also applied to this data and resulted in a sample of 1,591 messages.

The News Statements and Headlines were collected from a pool of 20,000 RSS feeds in the period between August and November 2015 (e.g. Associated Press News⁵, Reuters, Handelsblatt⁶, Bloomberg⁷, and Forbes⁸). A final set of about 1,780 News Statements and Headlines was produced.

3.3.1.3 Annotation

To create a gold standard, the final set of data was annotated by three independent financial expert annotators using a Web platform developed for that purpose and according to the annotation guidelines defined. A fourth domain expert consolidated the ratings to create the final dataset. The total time the experts spent on annotating and consolidating the dataset is 120 hours (30 hours per expert).

Each statement (instance) is annotated with the following information:

- **Cashtag (sub-task 1) / Company (sub-task 2):** A stock company symbol (for

⁵<https://apnews.com/>

⁶<https://www.handelsblatt.com/>

⁷<https://www.bloomberg.com/>

⁸<https://www.forbes.com/>

microblogs) or reference to a company (for news statements/headlines) to which a sentiment score is assigned.

- **Sentiment Score:** A sentiment between -1 (very negative/bearish) and 1 (very positive/bullish), with 0 assigned to each cashtag or company representing neutral/no sentiment. The sentiment is assigned from the point of view of an investor and the sentiment annotation is carried out by domain experts. Textual data containing information implying a positive prospective trend for a company or stock, the markets, or the economy, in general, constitutes a positive sentiment, whereas information revealing negative trends constitutes a negative sentiment since it may impact companies, markets, or the economy negatively.
- **Span (sub-task 1):** extract of a text string in which sentiment is expressed.
- **Message (sub-task 1) / Title (sub-task 2):** text string in which the sentiment is expressed.
- **Source (sub-task 1):** textual value, either “twitter” or “stocktwits” depending on the origin of the text message.

The following are some examples of annotated microblogs and news headlines.

Microblogs

Este Lauder beats on Revenues and EPS and boosts dividend 25% - global growth in the Middle Class trend continues. \$EL \$NKE \$SBUX \$AAPL

- **Sentiment Score:**
 - \$EL: 0.95
 - \$NKE: 0.5
 - \$SBUX: 0.5

– \$AAPL : 0.5

- **Cashtag**

– \$EL

– \$NKE

– \$SBUX

– \$AAPL

- **Span**

– \$EL:

* (13, 38) - “beats on Revenues and EPS”

* (43, 62) - “boosts dividend 25%”

* (65, 144) - “global growth in the Middle Class trend continues”

– \$NKE, \$SBUX, \$AAPL:

* (65, 144) - “global growth in the Middle Class trend continues”

Awaiting These Sell Signals on the \$SPY & \$QQQ -
<https://t.co/GF9PRk5OUF> \$TQQQ \$SQQQ <https://t.co/W97yN4Zb4N>

- **Sentiment Score:**

– \$SPY: -0.25

– \$QQQ: -0.15

– \$TQQQ: -0.15

– \$SQQQ : 0.10

- **Cashtag**

- \$SPY
- \$QQQ
- \$TQQQ
- \$SQQQ

- **Span**

- \$SPY:
 - * (0, 41) - “Awaiting These Sell Signals on the \$SPY”
 - * (From the blog post) - “this bearish rising wedge for the next sell signal in the SPY”
 - * (From the blog post) - Chart shows a bearish rising wedge
- \$QQQ, \$TQQQ:
 - * The message and blog make reference to shorting the SPY, but as but indexes are strongly correlated so some of the sentiment for SPY could be transferred to these Exchange-Traded Funds.
- \$SQQQ:
 - * The message and blog make reference to shorting the SPY, but as indexes are strongly correlated so some of the sentiment for SPY could be transferred to this Exchange-Traded Fund but inverted.

News Statements & Headlines

First Solar, Vivint Solar Lead Short Interest Trend

- **Sentiment Score:**

- First Solar: -0.7
- Vivint Solar: -0.7

- **Company**

- First Solar
- Vivint Solar

3.3.1.4 Quality

The quality of the annotations was assessed following a similar methodology as proposed in [Takala et al., 2014], where inter-annotator agreements measures for continuous data is calculated for the sentiment classifications.

Spearman’s Rank Correlation [Gauthier, 2001] on sentiment scores was calculated for each pair of annotators, then averaged across annotator pairs. This yielded the following results: 0.54 for news headlines (three annotators, three pairs) and 0.69 for microblogs (four annotators, six pairs).

3.3.1.5 Consolidation

A fourth financial expert consolidated the ratings to create the gold standard datasets. The gold standard for subtask 1 consists of 2,510 Twitter and StockTwit messages, whereas the one for subtask 2 contains 1,647 Headlines and News Statements.

3.3.2 Data Reuse

The dataset has been published⁹ under the Creative Commons Attribution-Non-Commercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license¹⁰ for non-commercial use.

3.3.3 Applicability

The datasets created can be used to achieve the following goals:

⁹<https://bitbucket.org/ssix-project/semEval-2017-task-5-subtask-2>

¹⁰<https://creativecommons.org/licenses/by-nc-sa/4.0/>

- Develop state-of-the-art classification models for sentiment analysis on short text for the Financial domain;
- Incentivise the creation of new lexical resources for the Financial domain;
- Understand how state-of-the-art sentiment analysis performs on a domain-specific/highly technical corpus; and
- Improve the understanding of linguistic phenomena and the creation of semantic models for the Financial domain.

The Finance domain has unique linguistic and semantic features, whose interpretation depends on the formulation of semantic models which reflect the economic and mathematical tools used by the experts to assess financial information. Moreover, the accurate interpretation of financial text requires the orchestration of large volumes of common sense and domain-specific financial/economic knowledge. Additionally, as much of the financial discourse is mediated by terms which demand precise definitions, many times associated with the quantification of economic phenomena, the semantic interpretation processes in the financial domain require fine-grained semantic interpretation approaches. From a linguistic standpoint, the following are some topics of interest which can be further explored in terms of research:

- Low-level linguistic analysis tools for the Financial domain (e.g. tokenisation, POS tagging, parsing);
- Sentiment classification on Financial text;
- Understanding of linguistic phenomena associated with Financial tweets;
- New semantic models for Finance;
- Construction and application of distributional semantic models on Finance;
- Sentiment compositionality;
- Machine learning and deep learning approaches for sentiment classification; and
- Lexical resources for the Financial domain.

3.3.4 Impact and Limitations

The Financial domain is a high impact use case for Opinion Mining, in particular sentiment analysis, since it has been shown that sentiment polarity can affect market dynamics [Goonatilake and Herath, 2007, de Kauter et al., 2015]. Sentiments are in some cases derived from news which discuss macroeconomic factors, company-specific, or political information as all of these can be market-relevant [Sinha, 2014]. Good news tends to lift markets and increase optimism [de Kauter et al., 2015, Schuster, 2003]. Evidence has been found that both quantitative measures (e.g., the quantity of news, market fluctuation) and qualitative indicators, (e.g., linguistic style and tone) affect investors' behaviour [Tetlock et al., 2008, Loughran and McDonald, 2011, Takala et al., 2014]. The authors in [Bollen et al., 2011] showed that changes in public mood reflect value shifts in the Dow Jones Industrial Index three to four days later. Given the link between sentiment polarity and market dynamics, the analysis of public sentiment becomes a powerful method to predict the market reaction. The generation of this dataset is motivated by general interest in this application area and the great potential for improvement. It aims to assess the overall market sentiment as well as sentiment about specific stocks, thus making use of their predictive power.

Improving the accuracy of sentiment analysis in the Financial domain could drive the engagement of multiple communities, such as financial (banks, hedge funds, insurances), academic (business schools, universities, research institutes), public (newspapers, regulatory agencies, media and information providers), private (technology providers, stock-traded companies), and governmental entities (governments, commission, chambers of commerce). Moreover, enhancing the quality of sentiment analysis will directly benefit various groups and have an economic impact. This will also empower both the public and private sectors to develop innovative services and products that are able to leverage the large amounts of sentiment data which is constantly produced and published on various social sources, such as social media networks and newswires.

In terms of impact, this dataset was generated for a SemEval-2017 task focusing on "Fine-Grained Sentiment Analysis on Financial Microblogs and News". In this regard, a high amount of researchers have used this dataset either as participants of the said

task and/or for carrying out Opinion Mining research within the Finance domain [Akhtar et al., 2017, Chen et al., 2018b, Chen et al., 2018a, Akhtar et al., 2020, Mishev et al., 2020, Bos and Frasincar, 2022], for sentiment analysis research targeting languages such as Lithuanian [Kapočiūtė-Dzikienė et al., 2019] and Portuguese [Saias et al., 2018], as motivation for other research within the Finance domain such as spam and bot activity in stock microblogs [Cresci et al., 2019] and decision support from financial disclosures [Kraus and Feuerriegel, 2017], influenced research focusing on application areas and domains such as economic trends during the COVID-19 pandemic [Hossu and Parde, 2021], and other research areas such as Recommender Systems [Aramanda et al., 2021].

With respect to the research gaps identified in Section 2.6, this dataset has some limitations given that it supports only the English language, and is only annotated for the sentiment dimension in terms of score. In light of these limitations, the direction of this thesis focused on the research gaps of multidimensional and multilingual social opinion data, where the Socio-Economic and Government dataset presented in Section 3.4 tackles all these gaps.

3.4 Dataset 3: Multidimensional and Multilingual Social Opinions for Malta’s Annual Government Budget

This section presents three high quality datasets focusing on bilingual multidimensional Opinion Mining for the Maltese less-resourced language and English, in the Socio-Economic and Government domains, specifically Malta’s annual Government Budget. During this annual event, the Government presents an estimate of its expenditures and revenues for the upcoming year. These datasets cover the 2018¹¹, 2019¹², and 2020¹³ budgets and consist of over 6,000 online posts from newswires and social networking services. To our knowledge, this is the only user-generated content Government Budget dataset that is available for Opinion Mining.

Section 2.6 lists several research gaps from the existing literature within the Social

¹¹<https://finance.gov.mt/en/The-Budget/Pages/The-Budget-2018.aspx>

¹²<https://finance.gov.mt/en/The-Budget/Pages/The-Budget-2019-G5J3D1.aspx>

¹³<https://www.finance.gov.mt/en/The-Budget/Pages/The-Budget-2020-GD-9691.aspx>

Opinion Mining research area. In fact, the three datasets presented here focus on the following ones:

- Gathering of social data from more than one data source, namely social networking services and newswires;
- Enabling multiple techniques to be explored for classification purposes;
- Collection of a new social dataset for a real-world application area which contains bilingual data (Maltese and English); and
- Annotations of **five** social opinion dimensions within the data (subjectivity, sentiment polarity, emotion, irony, and sarcasm).

A first version of this social dataset was published in 2019 [Cortis and Davis, 2019], which consists of social opinions for the 2018 budget. This volume has been increased with the user-generated data of the 2018, 2019, and 2020 budgets and further enhanced with the five different social opinion dimensions mentioned above, and in terms of the negation, topic, and language annotation types.

3.4.1 Generation Process

This section describes the multi-stage process used for building each of the three natural language Malta Government Budget datasets (2018, 2019, 2020), namely the methods employed for data collection, the annotation process of each dataset, and the data quality measures carried out to consolidate the final datasets.

3.4.1.1 Data Collection

The datasets were collected from the following data sources: **Newswires** - Times of Malta¹⁴, MaltaToday¹⁵, The Malta Independent¹⁶; and **Social networking services** - Twitter. Similar to [Cortis and Davis, 2019], the data source selection was based on citizens' preference for online news in Malta, with the Times of Malta and MaltaToday being

¹⁴<https://www.timesofmalta.com/>

¹⁵<https://www.maltatoday.com.mt/>

¹⁶<https://www.independent.com.mt/>

the top two, followed by The Malta Independent in fourth [Martin, 2020]. On the other hand, Twitter is actively used for Maltese politics, especially during each annual Government Budget. This is reflected in [Kemp, 2019], where the total advertising audience on Twitter in Malta amounts to over 60,000 monthly active users.

3.4.1.1.1 Newswires Table 3.7 presents the following newswires’ information for each respective budget (**row 1**): initial number of comments collected from Times of Malta (**row 2**), MaltaToday (**row 4**), The Malta Independent (**row 6**), the total number of comments left for each newswire after removing images or the ones deleted by the editor/comment owner (**row 3, row 5, row 7**), and the overall total number of comments (**row 8**).

The online news articles selected from each newswire for each budget year contained content in one of the following categories:

1. overview of the upcoming budget, published either on the day prior to the budget announcement or on the day of the budget, a few hours before the announcement;
2. near to real-time live updates of the budget measures being presented for the upcoming year; and
3. overview or feedback on the presented budget, published after the budget finishes, on the same day, or the following day.

1	Newswire comments	Budget 2018	Budget 2019	Budget 2020
2	Times of Malta (Initial)	253	354	275
3	Times of Malta (Total)	249	350	270
4	MaltaToday (Initial)	178	296	349
5	MaltaToday (Total)	175	280	306
6	The Malta Independent (Initial)	46	10	39
7	The Malta Independent (Total)	45	9	39
8	Overall (Total)	469	639	615

Table 3.7: Details of Newswires data for each dataset

These articles enable citizens to post their opinions and/or reactions on the budget and the content published in the said articles. Therefore, the articles that produced most comments, in terms of volume from citizens, were selected from each newswire. It is

important to note that the majority of these user comments are similar in nature to online posts published on social networking services, such as Facebook.

All of these comments were manually extracted for building the datasets, in order to annotate them in terms of the different social opinion dimensions mentioned in Section 3.4. Moreover, four online articles from each newswire were chosen for each budget to ensure a diverse sample of online posts¹⁷. These reflect the opinion of the general public with respect to the budget as a whole. Therefore, budget domain specific articles, e.g., an article focusing only on Technology budget measures, were omitted from the ones selected, with priority given to the ones reviewing the budget at large and ones that listed or gave an overview of all the budget measures for each domain.

A total of 1,800 comments were collected from the selected newswires for the 2018, 2019, and 2020 budgets. The ones that resulted in deleted comments (by the respective newswire or comment owner) or comments that consisted of images only were removed, leaving a total of 1,723 online posts for annotation purposes.

3.4.1.1.2 Social networking services As for online posts from Twitter (tweets), the ones that contained the following hashtags and/or keywords were extracted for each of the three budgets: “maltabudgetYY”, “malta budget YYYY”, “maltabudget YYYY”, “malta budgetYYYY”, “maltabudgetYYYY”, “malta YYYY budget”, and “YYYY budget malta”, with “YY”/“YYYY” referring to the respective budget year “18”/“2018”, “19”/“2019”, and “20”/“2020”. The chosen keywords were based on the manual identification of the most common keywords used in the content of tweets relevant to the Malta Budget.

Table 3.8 presents the following Twitter data information for each respective budget (**row 1**): date range of the data collection period for each budget where the date of the first and last tweet were determined through a manual search using the Twitter Advanced Search feature (**row 2**), the total amount of tweets initially collected using the seven hashtags and/or keywords (**row 3**), the total amount of tweets remaining after removing duplicate (based on exact content) tweets and retweets (**row 4**), and the official budget hashtag used by the Government of Malta (**row 5**).

¹⁷“Online posts” is the general term used within this thesis to refer for both comments and tweets.

1	Budget	2018	2019	2020
2	Data collection dates	28/08/2017-05/06/2018	20/07/2018-23/04/2019	01/09/2019-25/03/2020
3	Tweets collected (Initial)	4,168	4,682	4,904
4	Tweets remaining (Total)	1,673	1,677	1,314
5	Official budget hashtag	maltabudget18	maltabudget19	maltabudget20

Table 3.8: Details of Twitter data for each dataset

A total of 13,754 tweets were collected from Twitter for the 2018, 2019, and 2020 budgets. Any duplicate tweets and retweets (based on exact content) were removed, leaving a total of 4,664 tweets for annotation purposes.

The Twitter Premium Search API¹⁸ was used via the TwitterAPI Python library¹⁹ (used to access the Twitter API) to collect the tweets related to the three budgets, in particular the full-archive data endpoint. No online posts from Facebook were collected given that access to the Public Feed API is restricted and users cannot apply for it.

3.4.1.2 Sampling Strategy

A random sampling strategy was used to gather the data. Four online articles for each newswire were chosen (in total twelve articles), specifically ones that had the highest number of user-generated comments. As for Twitter, all data made available within the limits of the respective Twitter API was gathered, therefore equating a significant representative sample of the population.

3.4.1.3 Annotation

All the online posts collected from the newswires and social networking services were presented to three raters. In terms of expertise, all of the raters were proficient in Malta’s two official languages (Maltese and English), with two raters being computer science graduates and working in the Technology domain and one rater being a business and management graduate and working in the Human Resources domain.

All the raters were given a lecture on Opinion Mining, whereas annotation guidelines (See Appendix A for the Malta Government Budget Annotation Guidelines) were provided

¹⁸<https://developer.twitter.com/en/docs/twitter-api/premium/search-api/api-reference/premium-search>

¹⁹<https://www.github.com/geduldig/TwitterAPI>

to support them during the annotation process. These guidelines were piloted twice during the annotation process –primarily after twenty-five (25) annotations, and then after a hundred (100) annotations–, following clarification and feedback with the raters. Each rater took approximately 120 hours to annotate the three datasets, therefore the total estimated annotator time is 360 hours. A fourth rater, a computational linguist from academia, consolidated the annotation values to create the three final datasets. The annotation process discussed above follows the Model (Model and Guidelines) - Annotate - Model (Evaluate) - Annotate (Revise) cycle defined in [Pustejovsky and Stubbs, 2012].

Each online post is annotated with the following information (annotation types):

1. **Subjectivity:** binary value, with 1 referring to subjective posts and 0 referring to objective posts;
2. **Sentiment Polarity:** categorical value (3-levels) for the sentiment polarity of the online post (negative, neutral, positive);
3. **Emotion:** categorical value for the emotion of the online post based on Plutchik’s [Plutchik, 1980] eight primary emotions (joy, sadness, fear, anger, anticipation, surprise, disgust, trust);
4. **Irony:** binary value, with 1 referring to irony in online posts;
5. **Sarcasm:** binary value, with 1 referring to sarcasm in online posts;
6. **Negation:** binary value, with 1 referring to negated online posts²⁰;
7. **Off-topic:** binary value, with 1 referring to off-topic online posts that are political but not related to the budget; and
8. **Language:** numerical value, with 0 referring to online posts in English, 1 referring to posts in Maltese, 2 referring to Maltese-English (Maltenglish) code-switched²¹ posts, and 3 referring to posts in other languages.

The following is an example of an online post and the annotations for each type:

²⁰A negated post refers to the opposite of what is conveyed due to certain grammatical operations, such as ‘not’ (English) and ‘mhux’ (‘not’ in Maltese).

²¹Code-switching is a linguistic phenomenon that occurs when two or more languages are used in a single sentence or discourse.

Online post	“Online Maltese language spellchecker to be commissioned #maltabudget20”
Annotation Types	Subjectivity: 0; Sentiment Polarity: Positive; Emotion: Joy; Irony: 0; Sarcasm: 0; Negation: 0; Off-topic: 0; Language: 0

3.4.1.4 Quality

To ensure that the final datasets provided are of good quality, some basic pre-processing was carried out on the source data collected, whereas inter-rater reliability was calculated to determine that the level of agreement between the raters’ annotations for each annotation type.

3.4.1.4.1 Pre-processing Basic pre-processing was carried out on the data collected as discussed in Section 3.4.1.1. Any deleted comments (by the respective newswire or comment owner) or comments that consisted of images only were removed from the newswires data, whereas any duplicate tweets and retweets were removed from the Twitter data. Moreover, any HTML tags and line breaks were also removed from the collected tweets.

3.4.1.4.2 Inter-rater Reliability The quality of the three datasets for each annotation type (described in Section 3.4.1.3 above), is evaluated through inter-rater reliability, that is, the level of agreement between the raters’ annotations. The percent agreement (% Agree) is primarily calculated on the annotations performed by the three raters, which basic measure is calculated for two different levels, annotations agreed by all of the three raters (% Agree - 3 raters) and annotations agreed by two raters (% Agree - 2 raters). Two de facto statistical measurements, Fleiss’ kappa [Fleiss and Cohen, 1973] and Krippendorff’s Alpha [Krippendorff, 2011] have also been calculated. Fleiss’ kappa takes chance agreement into consideration, which is commonly used for categorical variables, whereas Krippendorff’s Alpha is used for content analysis to identify the agreement between raters and can apply to incomplete or missing data, any number of raters, any number of measurement level (nominal, ordinal, interval, ratio, etc.), and small and large sample sizes alike. Therefore, both measures are applicable when three or more raters perform the annotations and are used to measure the degree of agreement in classification over agreement

that is expected when raters randomly assign class labels i.e., by chance.

Tables 3.9, 3.10, and 3.11 provide the inter-rater reliability agreement scores of the 2018, 2019, and 2020 Malta Government Budgets respectively, for each annotation type.

Annotation Type	% Agree - 3 raters	% Agree - 2 raters	Fleiss' kappa	Krippendorff's Alpha
Subjectivity	0.9841	0.0159	0.9776	0.9776
Sentiment Polarity	0.8978	0.1022	0.8721	0.8721
Emotion	0.4599	0.5401	0.5160	0.5001
Sarcasm	0.9804	0.0196	0.7626	0.7625
Irony	0.9818	0.0182	0.8256	0.8256
Negation	0.9300	0.0700	0.7539	0.7537
Off-topic	0.9370	0.0630	0.8227	0.8226
Language	1	0	1	1

Table 3.9: Malta Government Budget 2018 - Inter-rater reliability measures for each annotation type

Annotation Type	% Agree - 3 raters	% Agree - 2 raters	Fleiss' kappa	Krippendorff's Alpha
Subjectivity	1	0	1	1
Sentiment Polarity	0.7323	0.2677	0.7155	0.7151
Emotion	0.3804	0.6196	0.4269	0.4155
Sarcasm	0.9996	0.0004	0.9950	0.9950
Irony	0.9417	0.0583	0.6397	0.6394
Negation	0.9275	0.0725	0.6361	0.6353
Off-topic	0.9154	0.0846	0.8263	0.8263
Language	0.9175	0.0825	0.8714	0.8714

Table 3.10: Malta Government Budget 2019 - Inter-rater reliability measures for each annotation type

Annotation Type	% Agree - 3 raters	% Agree - 2 raters	Fleiss' kappa	Krippendorff's Alpha
Subjectivity	1	0	1	1
Sentiment Polarity	0.7351	0.2649	0.7131	0.7128
Emotion	0.4795	0.5205	0.5212	0.5159
Sarcasm	0.9990	0.0010	0.9827	0.9827
Irony	0.9326	0.0674	0.6167	0.6150
Negation	0.9559	0.0441	0.8833	0.8833
Off-topic	0.9020	0.0980	0.7881	0.7880
Language	0.9984	0.0016	0.9969	0.9969

Table 3.11: Malta Government Budget 2020 - Inter-rater reliability measures for each annotation type

Interpretation of the reliability results listed in these tables differs between measures. All result values range from 0 to 1, where 0 signifies a perfect disagreement and 1 a perfect agreement for all measures. The % Agree is straightforward and the results simply

provide an overview of the annotations that were in agreement by all three and two raters respectively (which signifies the majority result from three raters). On the other hand, Fleiss' kappa results generally can be interpreted according to the classification guidelines by [Landis and Koch, 1977] for categorical data. Such results are interpreted as follows: less than 0 - poor agreement, 0.0 to 0.20 - slight agreement, 0.21 to 0.40 - fair agreement, 0.41 to 0.60 - moderate agreement, 0.61 to 0.80 - substantial agreement, and 0.81 to 1.0 - almost perfect agreement. Lastly, Krippendorff's Alpha results are generally interpreted as follows: a value of 0.80 or higher constitutes a marker of good reliability, whereas results within the 0.667 to 0.80 range allow for tentative conclusions to be drawn [Krippendorff, 2018]. It is worth noting that these guidelines are more strict than the ones drawn up by [Landis and Koch, 1977]. Therefore, one has to interpret these results in accordance to the particular hypothesis that is being tested and the validity requirements established on the research results.

An almost perfect agreement was achieved across the three datasets for *subjectivity* and *language* annotations, whereas a substantial/almost perfect agreement was achieved for *sentiment polarity*, *sarcasm*, and *off-topic* annotations. The *emotion* annotation was consistent across, with a moderate agreement. Lastly, the *irony* and *negation* annotations produced substantial to almost perfect agreements.

The moderate and contrasting results across datasets highlight the challenge behind these annotations tasks, especially when determining the *emotion*, *irony*, and *negation*. In fact, the % Agree - 2 raters of online posts from newswires is higher than that of online posts from social networking services (in this case Twitter), due to user-generated content in newswires being lengthier and hence more difficult to annotate. This is the opposite in the case of % Agree - 3 raters of online posts from social networking services, which agreement is higher than its equivalent for newswires. Emotions are very subjective and can differ from one person to another, therefore can be annotated in an inconsistent manner [Mohammad and Turney, 2013]. Also, people tend to confuse sarcasm for irony and vice-versa, and sometimes find their interpretation difficult [Van Hee, 2017]. Irony proved to be more challenging to annotate than sarcasm, probably due to irony being a more sophisticated form of communication and the different types of irony categories, such

as verbal and situational [Reyes et al., 2013]. More statistics, remarks, and observations for each annotation type are discussed in Section 3.4.2.

3.4.1.5 Consolidation

A computational linguist (fourth rater mentioned in Section 3.4.1.3), consolidated the annotations to create the three final datasets. In cases where at a minimum two out of three raters agreed on the annotation, this was selected as being final. However, in cases of non-agreement between the three raters, the computational linguist discussed the results with the three raters and selected the most appropriate annotation value after an agreement was reached. This was only necessary for annotations containing categorical values, namely *sentiment polarity*, *emotion*, and *language*.

3.4.1.6 Topic Annotation

Each online post within this dataset was annotated with the main topics mentioned based on the Government ministries and entities²². The topic annotation process was automated using a Python script where a keyword-based approach was adopted. These results were then manually corrected to improve the topic annotations. Table 3.12 provides an overview of the number of online posts for each annotated topic.

The online posts for four main Government domains, namely **Environment**, **Taxation**, **Transport**, and **Pension**, were further annotated for their sub-topics. The following is the list of sub-topics annotated for each respective domain, with the number of annotated online posts listed in brackets:

- **Environment:** biodiversity (1), bottles (1), car batteries (2), carbon (1), concrete (1), bottle refund (3), carbon (1), culture change (1), development (4), electric cars (4), green corners (1), hydrogen cars (1), machinery (5), nature (1), park (1), plastic (1), pollution (1), recycling (1), environmental sustainability (5), trees (2), urban garden (2), waste (3), waste technology (1), zero emission vehicles (1);

²²<https://www.gov.mt/en/Government/Government%20of%20Malta/Ministries%20and%20Entities/Pages/default.aspx>

Topic Annotation	Number of Online Posts
Agriculture	7
Business	112
Climate	73
Culture	82
Economy	132
Education	46
Employment	51
Energy	47
Environment	73
Equality	3
Finance	73
Fisheries	8
Governance	6
Healthcare	11
Heritage	14
Infrastructure	50
Innovation	20
Justice	40
Legal	22
Pensions	104
Public Sector	13
Research	31
Science	3
Sport	38
Sustainable Development	2
Taxation	567
Trade	23
Transport	136
Technology	48
Tourism	15
Water	44

Table 3.12: Details of online posts for each topic annotation

- **Taxation:** alcohol (5), bread (3), cars (7), cigarettes (10), congestion (1), construction (6), dividends (1), doctorate (2), emission (1), exemption (2), foreign tax (1), fuel (12), grants (7), home loan (1), indirect tax (1), income tax (32), married rate (1), masters (2), milk (5), music (2), overtime (4), part-time (3), pension (12), photovoltaic batteries (1), property (6), rebate (3), refund (6), rent (3), rental (2), self-employed (1), shares (1), single rate (1), sports (1), start-ups (1), tax credit (2), tax evasion (3), utility bills (9), VAT (7), wages (1);
- **Transport:** allowance (1), alternative methods (1), bicycle (6), civil service (3), commercial vehicles (1), elderly (3), ferry transport (1), Gozo link (2), health treat-

ment (1), imports (1), motorcycle (1), pedelec (1), pensioners (1), people with disability (1), private cars (1), public transport (15), school transport (12), students (8), sustainability (2), traffic (3), traffic pollution (1), youths (8); and

- **Pension:** armed forces (1), pension increase (22), pension schemes (1), private pension (1).

3.4.2 Statistics and Discussion

The three datasets consist of 6,387 online posts in total. The distribution of the 2018, 2019, and 2020 datasets’ annotations for the information (annotation types) discussed in Section 3.4.1.3 are presented in the following sub-sections. Moreover, any remarks and observations made during the annotation, quality, consolidation, and analysis processes are also discussed.

3.4.2.1 Subjectivity

3.4.2.1.1 Descriptive Statistics Table 3.13 presents the distribution of subjectivity annotations of the online posts for the 2018, 2019, and 2020 budget datasets. The majority of the posts in 2019 and 2020 datasets are *subjective*, with the ones in 2018 being mostly *objective*.

Subjectivity	Budget 2018	Budget 2019	Budget 2020
Subjective	38.66%	58.59%	58.32%
Objective	61.34%	41.41%	41.68%

Table 3.13: Distribution of subjectivity annotations

3.4.2.1.2 Remarks and Observations

- Objective online posts can imply a sentiment polarity and emotion, since they can represent desirable or undesirable facts in certain specific domains or contexts [Liu, 2015], such as the Socio-Economic and Government domains in the context of the Government Budget. This opposes a general misconception within the Opinion Mining research area, that objective text does not have any sentiment polarity by definition.

- Online posts that had a budget measure written (objective) followed by a subjective hashtag (see Example 1) were classified as being *objective*, since the emphasis was on the budget measure and not the subjective hashtag #WeNowLook2TheFuture.

In first legislature Government managed to reduce debt from 70% of GDP, to 57.6%.
#MaltaBudget18 #WeNowLook2theFUTURE (**Example 1**)

- Online posts that had a reference to the budget through a subjective hashtag, such as #WeNowLook2TheFuture (see Example 2), or a personal opinion followed by a subjective hashtag (see Example 3), were classified as *subjective*, since both instances emphasised the user’s opinion.

#MaltaBudget18 #WeNowLook2TheFUTURE #Malta (**Example 2**)

A budget with a true socialist heart - grazzi @JosephMuscat_JM #MaltaBudget18
#WeNowLook2TheFuture (**Example 3**)

- The majority of online posts from Twitter (tweets) are *objective*, with their text referencing the budget measures being read out by the Minister for Finance.
- Even though retweets of objective budget measure online posts can indicate a show of support/statement of approval, hence opinion of the entity retweeting, they were still classified as being *objective*, since they were posted for information sharing purposes rather than for expressing their opinion through additional text (see Example 4).

RT @MaltaGov: Robotic surgery to be introduced in oncology #maltabudget20
(**Example 4**)

3.4.2.2 Sentiment Polarity

3.4.2.2.1 Descriptive Statistics Table 3.14 presents the distribution of sentiment polarity annotations of the online posts for the 2018, 2019, and 2020 budget datasets. All three datasets provide a high number of *positive* posts, with the *negative* ones showing an increase in each subsequent budget.

Sentiment Polarity	Budget 2018	Budget 2019	Budget 2020
Positive	63.21%	49.96%	53.86%
Neutral	20.82%	30.87%	23.02%
Negative	15.97%	19.17%	23.12%

Table 3.14: Distribution of sentiment polarity annotations

3.4.2.2.2 Remarks and Observations

- Certain online posts contained multiple sentiment polarities, such as a *positive* polarity on the current budget or budget measure and a *negative* polarity on the future anticipated long-term effect of the said budget or budget measure. In these cases, the sentiment polarity is annotated in relation to the current budget or budget measure. Such instances highlight the importance of aspect-based Opinion Mining and the challenges faced when determining an overall sentiment polarity to certain online posts given that multiple budget measures can have a different sentiment polarity (refer to Section 3.4.2.11 for a more in-depth discussion on aspect-based Opinion Mining).
- The sentiment polarity annotated does not always reflect the sentiment towards a particular budget or budget measure, due to it being compared to previous budgets or budget measures. For example, certain online posts have a *negative* sentiment due to the current budget being compared with the 2011 budget which had introduced several new taxes and did not achieve certain targets, such as Gross Domestic Product (GDP) growth.
- Similarly, certain polarities were aimed at a particular person or person’s reaction and not at the budget itself. Example 5 is of a *negative* sentiment due to a lack of recognition for the budget by the opposition leader.

Inkredibbli! @adriandeliapn jibqa’ jsostni li dan hu Gvern bla pjan... ma tgħallem xejn. [Maltese] / Incredible! @adriandeliapn keeps insisting that this Government has no plan... he didn’t learn anything. [English] #maltabudget19 @JosephMuscat_JM (**Example 5**)

- Certain online posts from newswires have a *negative* sentiment aimed at the writer or opinion of the previous online post and not the budget itself, in which cases the sentiment towards the budget would be the opposite, that is, *positive* (see Example 6).

<p>L-Unions kollha jghejdu li hu tajjeb u int u xi erba ohra tghejdu li kien hazin . Min ihobb jeqred, jeqred jibqa . [Maltese] / All the Unions say that it is good and you and a few others say that it was bad . Who enjoys grumbling, keeps grumbling . [English] (Example 6)</p>
--

- Even though some online posts are classified as having a *negative* sentiment, this does not mean that the Government is being directly criticised but merely, the overall impression of the budget measure in question is not good and certain proposals and suggestions are being made in the subsequent online posts. For example, two online posts from the 2020 budget discuss the *negative* environmental impact of a budget measure that offers a grant to cover part of the cost of buying a battery storage system for owners of photovoltaic panels. These two online posts are actually providing alternative solutions to this measure, such as offering different feed-in tariffs, that can leave a *positive* environmental and economic impact.
- Online posts of a sarcastic or ironic nature result in changing and/or influencing the overall sentiment conveyed, therefore their text cannot be taken at face value.

3.4.2.3 Emotion

3.4.2.3.1 Descriptive Statistics Table 3.15 presents the distribution of emotion annotations of the online posts for the 2018, 2019, and 2020 budget datasets. All three datasets produce a high number of posts conveying *joy* and *anticipation* emotions.

Emotion	Budget 2018	Budget 2019	Budget 2020
Joy	43.42%	34.46%	47.07%
Trust	8.59%	5.44%	2.75%
Fear	0.61%	1.30%	1.50%
Surprise	1.96%	3.58%	2.70%
Sadness	2.15%	4.23%	5.24%
Disgust	8.31%	6.48%	8.81%
Anger	6.12%	4.88%	6.48%
Anticipation	28.85%	39.64%	25.45%

Table 3.15: Distribution of emotion annotations

3.4.2.3.2 Remarks and Observations

- The annotation of the *joy* emotion for certain online posts does not always reflect the typical joyous nature as usually expressed by a person through the use of certain

special characters (e.g., exclamation mark) or emoticons (e.g., smiley face). However, in context of the Government Budget domain, this emotion category is the closest towards annotating one of a positive nature, such as the announcement of a positive budget measure (see Example 7). On the other hand, the *trust* emotion was used for posts expressing support to the Government.

#MaltaBudget18 Live — Taskforce set up to focus on implementation of #Blockchain National Strategy in #Malta (**Example 7**)

- In the context of objective online posts (mostly tweets), emotions are somewhat different than those for subjective ones, which generally reflect the person’s emotions. The *joy* emotion, for example, would be conveying concrete support for the budget measures in question and/or through foreseeing a better quality of life to the citizens and residents of Malta.
- The *anticipation* emotion was used a lot in objective online posts that provided either links publishing budget updates (see Example 8), or updates on the current budget and/or budget measures announced (see Example 9).

#maltabudget2019 #taxes #maltaindependent <https://t.cozdjCZxleYw> (**Example 8**)

More investment expected in Artificial Intelligence and Internet of Things #MaltaBudget20 (**Example 9**)

- Throughout the 2018, 2019, and 2020 budget datasets, there were a number of online posts where the three annotators chose different emotion classifications, due to numerous emotions being expressed (e.g., *fear*, *anger*, and *sadness*) or the emotions being of a similar nature within the spectrum (e.g., *fear* and *surprise*).
- In certain instances the annotators found it challenging to select one of Plutchik’s eight primary emotions, which highlights the complexity of such a task and the identification of the appropriate emotion category, which can easily differ from one person to another. Our claim is supported by [Susanto et al., 2020] where the authors highlight that “emotions are still a rather mysterious subject to study”. This is reflected by a lack of universal emotion categorisation model and hence why numerous emotion classifications have been published in literature over the years. In

addition, detecting emotion in text can be difficult for humans due to the personal context of individuals which can influence emotion interpretation, thus resulting in a low level of inter-rater agreement [Canales Zaragoza, 2018]. Moreover, words used in different senses can lead to different emotions, hence making the emotion annotation more challenging [Mohammad and Turney, 2013]. This claim is also supported in [Devillers et al., 2005], who mention that categorisation and annotation of real-life emotions is a big challenge given that they are context-dependent and also highly person-dependent, whereas unambiguous emotions are only possible in a small portion of any real corpus. Therefore, the nature of relevant emotion data is too infrequent to provide adequate support for consistent annotation and modelling through fine-grained emotion labels.

3.4.2.4 Sarcasm

3.4.2.4.1 Descriptive Statistics Table 3.16 presents the distribution of sarcasm annotations of the online posts for the 2018, 2019, and 2020 budget datasets. The number of *sarcastic* posts is more or less consistent for the three datasets, which number diminishes from the 2018 till the 2020 budget.

Sarcasm	Budget 2018	Budget 2019	Budget 2020
Sarcastic	3.17%	2.98%	2.07%
Not Sarcastic	96.78%	97.02%	97.93%

Table 3.16: Distribution of sarcasm annotations

3.4.2.4.2 Remarks and Observations

- Certain sarcastic online posts still keep their original sentiment polarity, e.g., *negative* (see Example 10).

Said like the true monkey that you are (**Example 10**)

- Other sarcastic online posts have a particular sentiment polarity even though in reality they convey an opposite one. Example 11 has a *positive* sentiment polarity due to its sarcastic nature and use of “face with tears of joy” emoticons (also present in original text), however, it conveys a *negative* one towards the referenced person.

Thanks for the advice pycho Joe!!! (**Example 11**)

3.4.2.5 Irony

3.4.2.5.1 Descriptive Statistics Table 3.17 presents the distribution of irony annotations of the online posts for the 2018, 2019, and 2020 budget datasets. The number of *ironic* posts slightly increased in the 2019 and 2020, when compared to 2018.

Irony	Budget 2018	Budget 2019	Budget 2020
Ironic	3.78%	5.87%	5.81%
Not Ironic	96.22%	94.13%	94.19%

Table 3.17: Distribution of irony annotations

3.4.2.5.2 Remarks and Observations

- Certain terms that are usually used to express a *positive* sentiment polarity, such as “thanks”, “kind”, “hope”, and “entertaining”, have sometimes been used in ironic online posts, even though they are conveying the opposite meaning of what is being said towards the particular entity, such as a person (see Example 12).

thanks for the kind words Albie. (**Example 12**)

- The ironic nature of certain posts convey a *negative* sentiment polarity, however express a positive emotion, such as *joy*. Example 13 is in reality praising the existing Government for the positive budget measures being announced (e.g., incentives) and referring to previous Governments in a negative sentiment based on past budget measures (e.g., additional taxes).

How boring, he keeps using the same words - give, giving, we give. We are so accustomed, for 27 years, to hearing the words - pay, taxes, tariffs. sacrifices etc
(**Example 13**)

3.4.2.6 Negation

3.4.2.6.1 Descriptive Statistics Table 3.18 presents the distribution of negation annotations of the online posts for the 2018, 2019, and 2020 budget datasets. The 2020 budget dataset produced the highest number of negations.

Negation	Budget 2018	Budget 2019	Budget 2020
Negated	12.65%	8.72%	14.05%
Not Negated	87.35%	91.28%	85.95%

Table 3.18: Distribution of negation annotations

3.4.2.6.2 Remarks and Observations

- Any negations within an online post that were not related to the budget or a particular budget measure announced by the Government for the respective year, were ignored. Example 14 contains a word (“never”) that indicates a negation, however, it was ignored (in line with the annotation guidelines), since only the first sentence was related to the 2019 budget.

Yeah sure, a rise of 2 euros a week just to cover the extra cost of bread and milk, As for the other extra costs, we will tackle that in our next budget. Accepting more than 40,000 economic immigrants can never improve the way of life of the Maltese workers, JMO. (**Example 14**)

3.4.2.7 Off-topic

3.4.2.7.1 Descriptive Statistics Table 3.19 presents the distribution of off-topic annotations of the online posts for the 2018, 2019, and 2020 budget datasets. A substantial percentage of posts in each dataset are off-topic, especially the one from 2019.

Off-topic	Budget 2018	Budget 2019	Budget 2020
Off-topic	12.65%	19.73%	17.73%
On-topic	87.35%	80.27%	82.27%

Table 3.19: Distribution of off-topic annotations

3.4.2.7.2 Remarks and Observations

- Online posts from newswires have a tendency to end up being classified as being off-topic due to several reasons, such as reference to previous Government administrations, measures e.g., pensions schemes introduced in the past, context of a previous online post misunderstood or its reply being ironic and not within context of Maltese politics, direct reference to a political figure after providing budget feedback e.g., opposition leader at the time, comparison of non-budget matters with other ju-

risdictions e.g., United Kingdom, and reference to current situations in the country e.g., scandals.

- Even though the text does not directly refer to any particular budget topic/measure, certain online posts were correctly annotated as on-topic since they would still indirectly refer to certain budget topics/measures. In Example 15 the text is referring to the cost of bread and milk which is related to the cost of the living allowance (COLA)²³, even though this is not specifically mentioned within the text.

Milk and bread in Malta cost nothing. (**Example 15**)

3.4.2.8 Language

3.4.2.8.1 Descriptive Statistics Table 3.20 presents the distribution of language annotations of the online posts for the 2018, 2019, and 2020 budget datasets. The majority of the posts in each dataset are in English, with the Maltese language (either as the primary language or as a secondary language) used in around a quarter of the posts.

Language	Budget 2018	Budget 2019	Budget 2020
English	71.52%	71.55%	79.99%
Maltese	4.34%	6.22%	3.21%
Maltese-English	23.20%	21.24%	15.97%
Other	0.93%	0.99%	0.83%

Table 3.20: Distribution of language annotations

- A total of 20 online posts from the 2018 budget dataset were classified as being written in other languages, namely: 15 in English-Italian, 1 in Italian, 1 in Maltese-Italian, 1 in English-French, 1 in Spanish-English, and 1 in Maltese-English-Italian.
- A total of 23 online posts from the 2019 budget dataset were classified as being written in other languages, namely: 12 in English-Italian, 3 consisted of links only, 2 in Dutch, 2 in English-French, 1 in Maltese-Italian, 1 in English-Spanish, 1 in English-Swedish, and 1 consisted only of one emoticon.
- A total of 16 online posts from the 2020 budget dataset were classified as being written in other languages, namely: 5 in English-Italian, 4 in Maltese-Italian, 1 in

²³<https://www.gemma.gov.mt/cost-of-living-increase/>

English-Spanish, 1 in English-French, 1 in English-Japanese, 1 in Maltese-English-Italian, 1 in Italian, 1 in Japanese, and 1 consisted only of emoticons.

3.4.2.8.2 Remarks and Observations

- Most of the Maltese-English code-switched online posts result in the majority of the terms being in Maltese, with only a few words written in English. Some common occurrences are: “budget” (English) instead of “baġit” (Maltese), and “euro” (English) instead of “ewro” (Maltese). However, there were still some cases where it was the opposite, that is, English being the primary language and Maltese the secondary language.
- Several online posts were written in Maltese, however, they used a hashtag (tweets) written in English (e.g., #maltabudget20), hence were classified as being code-switched. This was a common occurrence across the three datasets obtained from Twitter.
- Loan words such as “cappuccino” are not of a Maltese origin, however they have been incorporated within the Maltese language (and others worldwide, such as English), therefore, they were not classified as being non-Maltese.
- Certain terms are well accepted in Maltese e.g., “amen” (Christian word spelling), however, there is a Maltese translation of this word “ammen”²⁴ and more so a Maltese version “hekk ikun”²⁵. Therefore, given that this term is used worldwide, in principle classification in Maltese or English are both considered as being correct.
- The words “pastizz” (singular) or “pastizzi” (plural) refers to a traditional Maltese savoury pastry. Both words are in Maltese and have an English translation depending on the flavour (cheese cakes/pea cakes). Given that the words are widely used by the general public irrespective of the language, there were some instances where the annotators did not always recognise the words as being in Maltese and therefore classified the language as English, in cases where all the other text was in English.

²⁴<https://www.independent.com.mt/articles/2015-09-01/blogs-opinions/Chiselling-the-Maltese-Language-6736141379>

²⁵<https://www.timesofmalta.com/articles/view/Amen-written-in-Maltese.377498>

- Acronyms/slang words referring to English phrases, such as LOL (laughing out loud), were treated in their original language during annotation. Therefore, online posts in Maltese containing such terms, were classified as Maltese-English.
- Hashtags such as “#MaltaSuccess”, could have been meant to be in Maltese. However, due no Maltese characters used, the word “suċċess” (Maltese) might have been written as “success” (English). In this case, the words were classified as being in English.
- The level of inter-rater reliability agreement between Maltese and Maltese-English might be a bit lower than expected. This is due to the fact that certain terms, such as “man” within Maltese text and “pastizzi” within English text, have been embedded in our language and day-to-day vocabulary for such a long time, that they may have seemed natural for the annotators.

3.4.2.9 Data Sources

3.4.2.9.1 Descriptive Statistics Table 3.21 presents the distribution of all annotations from the online posts for the consolidated 2018, 2019, and 2020 budget datasets, for each data source, namely Times of Malta, MaltaToday, The Malta Independent, and Twitter.

3.4.2.9.2 Remarks and Observations Online posts from newswires tend to be of a more *negative* sentiment polarity to those from social networking services, such as Twitter, which are mostly of a *positive* sentiment polarity. The same applies for emotions, with the ones of a *positive* nature, such as *joy* and *trust*, being mostly present in Twitter, as opposed to the ones of a *negative* nature, such as *fear*, *sadness*, *disgust*, and *anger*, conveyed in online posts from newswires. A high number of online posts from Twitter (tweets) are *objective* due to them being about budget measures (factual). In terms of sarcasm and/or irony, online posts of this nature are mostly found in newswires. Moreover, a large portion of tweets were carried out by the members of the Cabinet of Malta, therefore may not provide a true reflection of the general population. A similar observation was made by Mellon and Prosser in their political science study [Mellon and Prosser, 2017]. However,

these online posts are still relevant since Twitter is an open social media platform that can be used by the general public, which social media data provides several opportunities for studying public opinion [Mellon and Prosser, 2017].

Data source	Times of Malta	MaltaToday	The Malta Independent	Twitter
<i>Total</i>	869	761	93	4664
Subjectivity				
Subjective	864	759	93	875
Objective	5	2	0	3789
Sentiment Polarity				
Positive	164	141	10	3235
Neutral	188	114	17	1286
Negative	517	506	66	143
Emotion				
Joy	62	93	5	2476
Trust	74	42	4	243
Fear	22	29	5	16
Surprise	84	53	8	32
Sadness	110	65	10	60
Disgust	204	228	24	42
Anger	166	148	28	27
Anticipation	147	103	9	1768
Sarcasm				
Sarcastic	59	105	7	6
Not Sarcastic	810	656	86	4658
Irony				
Ironic	160	125	12	32
Not Ironic	709	636	81	4632
Negation				
Negated	273	211	33	227
Not Negated	596	550	60	4437
Off-topic				
Off-topic	429	504	24	113
On-topic	440	257	69	4551
Language				
English	586	470	58	3618
Maltese	117	134	9	39
Maltese-English	156	145	25	971
Other	10	12	1	36

Table 3.21: Distribution of annotations by data source

3.4.2.10 Online Posts

3.4.2.10.1 Descriptive Statistics Tables 3.22, 3.23, 3.24, and 3.25 present statistics on the online posts for each data source, in terms of maximum, minimum, and average characters, and words within posts for each of the three datasets. Moreover, an analysis was carried out on emoticons/emojis in terms of total online posts containing at least one, overall total number for each dataset, and the highest and lowest number of emoticons/emojis present in a post for each dataset.

Data source	Budget 2018	Budget 2019	Budget 2020
Characters - Average	169.79	185.94	220.51
Characters - Maximum	1176	1851	1576
Characters - Minimum	5	4	10
Words - Average	29.38	31.67	38.60
Words - Maximum	173	324	280
Words - Minimum	1	1	2
Emoticons/Emojis - Total posts	4	4	0
Emoticons/Emojis - Overall total	5	5	0
Emoticons/Emojis - Highest number	2	2	0
Emoticons/Emojis - Lowest number	1	1	0

Table 3.22: Online posts statistics - Times of Malta

Data source	Budget 2018	Budget 2019	Budget 2020
Characters - Average	139.33	198.75	181.98
Characters - Maximum	748	2204	1350
Characters - Minimum	8	2	4
Words - Average	23.89	34.29	31.60
Words - Maximum	121	398	238
Words - Minimum	1	1	1
Emoticons/Emojis - Total posts	6	30	8
Emoticons/Emojis - Overall total	9	145	46
Emoticons/Emojis - Highest number	3	9	30
Emoticons/Emojis - Lowest number	1	1	1

Table 3.23: Online posts statistics - MaltaToday

Data source	Budget 2018	Budget 2019	Budget 2020
Characters - Average	339.39	275.60	181.03
Characters - Maximum	2037	610	694
Characters - Minimum	15	13	31
Words - Average	57.39	50	30.90
Words - Maximum	368	124	108
Words - Minimum	2	2	5
Emoticons/Emojis - Total posts	1	0	0
Emoticons/Emojis - Overall total	1	0	0
Emoticons/Emojis - Highest number	1	0	0
Emoticons/Emojis - Lowest number	1	0	0

Table 3.24: Online posts statistics - The Malta Independent

Data source	Budget 2018	Budget 2019	Budget 2020
Characters - Average	108.87	117.56	134.23
Characters - Maximum	295	352	318
Characters - Minimum	23	14	34
Words - Average	13.91	14.62	17.48
Words - Maximum	49	49	49
Words - Minimum	1	1	1
Emoticons/Emojis - Total posts	55	31	100
Emoticons/Emojis - Overall total	88	60	296
Emoticons/Emojis - Highest number	5	10	14
Emoticons/Emojis - Lowest number	1	1	1

Table 3.25: Online posts statistics - Twitter

3.4.2.10.2 Remarks and Observations Online posts within newswires data sources tend to be much longer than ones made on social networking services, such as Twitter.

This is evident from the statistics presented, with the largest post from the three datasets containing 2204 characters / 398 words (Budget 2019 - MaltaToday), whereas the largest post from Twitter consisted of 352 characters / 49 words (Budget 2019). It is worth noting that the maximum number of text content of a tweet can contain up to 280 characters (updated from 140 characters in November 2017). However, content returned by the TwitterAPI Python library includes certain links (e.g., of images embedded in a tweet) and certain character entity references (e.g., *&* for the *&* character), which content is made available for the end-users to decide on whether further pre-processing is needed, depending on their application. Moreover, the use of emoticons/emojis in online posts within social networking services (such as Twitter), is usually higher than those made in newswires. However, an online post from a newswire (Budget 2020 - MaltaToday) contained the highest number of emoticons/emojis (30).

3.4.2.11 Overall Remarks and Observations

- **Implicit vs. explicit opinions:** Certain online posts express a particular sentiment polarity and emotion, in view of the opinion expressed by someone else. Hereby, the users making such a post implicitly approve of the budget even though their post does not explicitly express it or vice-versa. Example 16 conveys a *negative* sentiment polarity and *sadness* emotion for the view expressed by Adrian Delia and not the respective budget.

Opposition leader @adriandeliapn is as inept as they come. His analysis of #maltabudget18 is totally out of sync with people. **(Example 16)**

- **Sarcasm vs. Irony:** It is important to clarify that an online post can only be annotated as being either sarcastic or ironic and not both.
- **Aspect-based Opinion Mining:** In certain cases, online posts are long in nature and contain opinions on multiple budget measures and/or Government entities.

@adriandeliapn said that #maltabudget20 has no measures for women. **Less tax for everyone, climate change measures, free transport for youths and elderly, minimum wage for severely disabled, higher pensions.** This budget is for men and women. Stop putting people in isolated boxes **(Example 17)**

In Example 17, multiple budget measures (highlighted) and other aspects were mentioned. These have a different sentiment polarity (*positive* for each measure and *negative* for Adrian Delia) and emotion (*joy* for each measure and *anger* for Adrian Delia). This shows why aspect-based Opinion Mining is important and the benefits of having such an approach that can interpret opinions in an accurate manner based on each aspect and/or entity.

- **Links** within online posts were not followed during the annotation process, with the classifications being based on the existing text only.
- The **Government Budget** context within the **Socio-Economic** and **Government** domains, is a complex topic of choice and this can be seen from the classification of certain on-topic/off-topic online posts which are not always straightforward to determine, especially given that user-generated content can be within the Government context but not within the context of the specific budget.

3.4.3 Data Reuse

The three multidimensional and multilingual datasets adhere to the Findability, Accessibility, Interoperability, and Reuse (FAIR) principles as follows:

- **Findable**: publicly available through the Zenodo²⁶ open-access repository;
- **Accessible**: through the Digital Object Identifier (DOI)²⁷ assigned by Zenodo;
- **Interoperable**: data available in a structured, open and machine-readable format, as comma separated values (CSV) files; and
- **Reusable**: published under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license for non-commercial use.

The three datasets do not contain any sensitive data, since they only include published public user-generated content. The identity of the users has been protected, where no usernames have been provided with respect to online posts collected from newswires.

²⁶<https://www.zenodo.org/>

²⁷<https://www.doi.org/10.5281/zenodo.4650232>

As for Twitter, the Developer Agreement and Policy²⁸ has been observed for all the data gathered. Therefore, only the Twitter IDs and respective annotation types shall be distributed, which data can only be used for non-commercial research purposes.

3.4.4 Applicability

The three datasets provide a valuable resource for developing Opinion Mining tools that gather Political and Socio-Economic insights from user-generated content in Malta's two official languages, Maltese and English. These can be used by the Government of Malta for policy formulation, policy-making, decision-making, and decision-taking. Moreover, their use can support similar initiatives in other countries (e.g., Irish Government Budget), studies in the Socio-Economic and Government domains, and other application areas, such as Politics, Finance, Marketing, Advertising, Sales, and Education.

Furthermore, these quality datasets are valuable for multiple research applications, namely:

- Tools and resources for low-resourced languages, such as Maltese;
- NLP for social media content in Maltese and English;
- NLP approaches for the analysis and processing of mixed-language online user-generated content, with a focus on code-switching in Maltese-English;
- Opinion Mining on monolingual (English/Maltese) and code-switched online user-generated content;
- Aspect-based Opinion Mining for multiple social opinion dimensions;
- Fine-grained opinion search and summarisation; and
- Subjectivity detection, sentiment analysis, emotion analysis, sarcasm detection, and irony detection (as separate research areas or otherwise) in multiple application areas.

²⁸<https://developer.twitter.com/en/developer-terms/agreement-and-policy>

3.4.5 Impact

Governments are increasingly using Political and Socio-Economic online user-generated content created across social media platforms and other websites to get a better grasp of the citizens' perceptions and needs, and society's problems at large. This resulted in the development of several information and communication tools and technologies, with the most critical ones being in the Opinion Mining area [Charalabidis et al., 2015]. According to Eurostat (2019 statistics), 13% of individuals living in Malta post opinions on civic or political issues via websites, such as blogs and social networks. Consequently, these datasets of user-generated content provide a voice to the citizens who use social media platforms to make their opinions known and/or provide feedback about any particular measure announced by the Government, whether it is tax related, industry specific, or any other social initiative.

In terms of impact, the annotated datasets of social opinions for the Malta Government Budget have the potential of being used for initiatives by the Government to capture the public opinion and perception. These valuable insights can be evaluated and taken in consideration for revision of measures and/or any bills presented and discussed in Parliament. In fact, Malta's Minister for Justice emphasised on the importance of personal opinions that are expressed on social media and blogs, and mentioned that these are a very important source of information for the Government when carrying out certain initiatives and processes, such as the rule-of-law reforms [Xuereb, 2020].

Moreover, the datasets were annotated for Malta's two official languages, Maltese (Malti), a Semitic language written in the Latin script which is also the national language, and English. The recent European Language Equality report on the Maltese language mentions that assessing the current state of the Maltese language within the digital sphere is hard to assess for several reasons, such as the lack of reporting on the difference between local internet users and local users of the Maltese language [Rosner and Borg, 2022]. However, the multidimensional and multilingual dataset presented in this section has been mentioned by Rosner and Borg as the only exception given that the online post language results have been reported in terms of the English, Maltese, and Maltese-English languages. The languages of this dataset align with Malta's Strategy and Vision for Ar-

tificial Intelligence [Schembri, 2019], where the country is investing in the development of Maltese language resources and tools. This work shall also counter the threat of “digital extinction” for the Maltese language, which has low technological support available in comparison with other European languages [Rosner et al., 2012]. Such datasets will encourage other researchers working on low-resourced languages to create similar datasets to safeguard their languages from a technological standpoint.

3.5 Conclusion

This chapter determines whether Social Opinion Mining can influence multiple application areas, namely Politics, Finance, and Government. In this regard, three datasets spanning the application areas mentioned, namely Politics targeting Referendums, Finance targeting Stock Trading, and Socio-Economic and Government targeting Government Budgets, were generated. Given that all the three datasets cater for one common social opinion dimension in terms of sentiment polarity, the datasets can be evaluated in terms of the influence of Social Opinion Mining in multiple application areas. However, this influence could not be evaluated for multiple social opinion dimensions. Due to this research gap, we proceeded with the research direction of catering for multiple social opinion dimensions that target multilingual data. Moreover, even though a number of social datasets exist for conducting Social Opinion Mining research in multiple application areas, current literature shows that there is still a lack of datasets that cater for multidimensional and multilingual social opinion data for Social Opinion Mining purposes. The collected and annotated Socio-Economic and Government dataset, specifically for Malta’s annual Government Budget, is made up of multilingual data gathered from newswires and social networking services, which annotations cater for multiple social opinion dimensions, namely subjectivity, sentiment polarity, emotion, irony, and sarcasm. Moreover, this dataset in itself caters for multiple domains which are handled within the Malta Government Budget, such as Transport, Environment, and Taxation. In this regard, Social Opinion Mining can be applied in several domains and hence influence multiple application areas. Therefore, this dataset caters for the research gaps of multidimensional and multilingual social opinion data. The content of this chapter answers **Research Question 2** and in part both

Research Questions 3 and 5. The next chapter targets how multidimensional and multilingual social opinion data, namely the Socio-Economic and Government dataset, can be used for the development of social opinion classification models.

Chapter 4

Multidimensional Opinion Classification

4.1 Introduction

This chapter presents multidimensional Social Opinion Mining on the novel multidimensional and multilingual social opinion dataset presented in Section 3.4, which consists of user-generated content gathered from newswires and social networking services in three different languages: **English** – a high-resourced language, **Maltese** – a low-resourced language, and **Maltese-English** – a code-switched language. This dataset is validated through the development of several baseline classification models using traditional machine learning techniques and state-of-the-art deep neural network models. The former techniques are still very much used with success for classification purposes, whereas usage of the latter models is on the increase especially for studies focused on adaptation to new domains and/or for knowledge transfer from one language to another. Validation is important since the annotated linguistic dataset shall be used to train the said machine learning and deep learning algorithms to automatically classify unseen data for the Social Opinion Mining dimension in question e.g., sentiment polarity [Pustejovsky and Stubbs, 2012]. This chapter also presents the relevant social opinion classification literature that use techniques based on machine learning algorithms, deep learning algorithms, and neural language models. This builds upon the Social Opinion Mining analysis presented in

Section 2.3.

The following baseline models are developed:

1. Machine learning-based classification models for **subjectivity**, **sentiment polarity**, **emotion**, **irony**, and **sarcasm**, at a **monolingual** level using user-generated content in **English**;
2. Machine learning-based classification models for **subjectivity**, **sentiment polarity**, **emotion**, **irony**, and **sarcasm**, at a **code-switched** level using user-generated content in **Maltese-English** and **Maltese**; and
3. Neural-based classification models for **subjectivity**, **sentiment polarity**, **emotion**, **irony**, and **sarcasm**, at a **multilingual** level using user-generated content in **English**, **Maltese**, and **Maltese-English**.

Sections 4.2-4.6 of this chapter are in part based on the following publications:

- **Keith Cortis**, Kanishk Verma, Brian Davis. “Fine-tuning Neural Language Models for Multidimensional Opinion Mining of English-Maltese Social Data.” In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), pp. 309-314. 2021.
- **Keith Cortis**, Brian Davis. “Baseline English-Maltese Classification Models for Subjectivity Detection, Sentiment Analysis, Emotion Analysis, Sarcasm Detection, and Irony Detection.” In Proceedings of the LREC 2022 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages (SIGUL 2022), pp. 161-168. 2022.

4.2 Social Opinion Classification Literature

This section presents the relevant literature for social opinion classification in terms of machine learning and deep learning algorithms (Section 4.2.1) and neural language models (Section 4.2.2).

4.2.1 Machine Learning and Deep Learning-based Algorithms

Studies focusing on text classification tasks, such as sentiment analysis, at a binary (two classes) and/or multi-class (more than two classes) level generally use (traditional) machine learning and deep learning algorithms for building their baseline models. The systematic review carried out on Social Opinion Mining (refer to Section 2.2) analysed a large number of studies that make use of social data, such as user-generated content from social media platforms, and identified techniques used for carrying out classification tasks in this research area. In terms of traditional supervised learning algorithms, the most common ones used for baseline, experimentation, evaluation, and/or comparison purposes are:

- Naïve Bayes (NB) [Lewis, 1998]: is a simple learning algorithm that uses Bayes rule and a strong assumption that the attributes are conditionally independent given the class. For classification purposes two variants are normally used [McCallum et al., 1998], the Multinomial NB (MNB) model for discrete counts i.e., the number of times a given term (word or token) appears in a document, whereas the Multivariate Bernoulli NB (MBNB) model is based on binary data where every token in a feature vector of a document is classified with the value of 0 or 1.
- Support Vector Machine (SVM) [Cortes and Vapnik, 1995]: is a method that looks at the given data and sorts it in two categories i.e., binary classification. If multi-class classification is required, the Support Vector Classification (SVC)¹, Nu-Support SVC (NuSVC)², or LinearSVC³ algorithms are usually applied, where the “one-against-one” approach is implemented for SVC and NuSVC, whereas the “one-vs-the-rest” multi-class strategy is implemented for LinearSVC.
- Logistic Regression (LR) [McCullagh, 1984]: is a simple and efficient method for binary and linear classification that models the probability of a discrete outcome given an input variable.

¹<https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC>

²<https://scikit-learn.org/stable/modules/generated/sklearn.svm.NuSVC.html#sklearn.svm.NuSVC>

³<https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html#sklearn.svm.LinearSVC>

- Maximum Entropy (MaxEnt) [Yu et al., 2011]: is a generalisation of LR for multi-class scenarios.
- Decision Tree (DT) [Quinlan, 1986]: is a supervised learning method used for classification tasks, where the trees are constructed starting with the root of the tree and proceed down to its leaves.
- Random Forest (RF) [Breiman, 2001]: is a supervised learning algorithm –which can be used for both classification and regression tasks– that creates a forest (which is an ensemble of DTs) and makes it somehow random.

The choice of traditional supervised learning algorithms selected is supported by other Opinion Mining reviews, such as [Ravi and Ravi, 2015], [Hemmatian and Sohrabi, 2019], [Carvalho and Plastino, 2021], [Ligthart et al., 2021]. Even though recent advances in Opinion Mining has seen an increase in the use of deep learning approaches, such as the Transformer model architecture [Vaswani et al., 2017], traditional machine learning algorithms are still very much used to carry out Opinion Mining classification tasks, with good results obtained especially on small datasets [Ligthart et al., 2021].

Several high-quality datasets for Social Opinion Mining are available for research purposes as part of shared evaluation tasks, such as SemEval, and/or through open access repositories, such as Zenodo. Teams submitting their systems in the SemEval sentiment analysis task on code-mixed tweets [Patwa et al., 2020] used the following techniques: traditional machine learning algorithms such as NB, LR, RF, and SVM; word embeddings such as word2vec [Mikolov et al., 2013], GloVe [Pennington et al., 2014], and fastText [Joulin et al., 2016]; and deep learning algorithms such as RNN, CNN, and BERT. In [Gupta et al., 2017], several machine learning (SVM best performer) and deep learning algorithms are used as baselines for contextual emotion detection on tweets. In [Mohammad et al., 2018], the authors trained a baseline SVM system using just word unigrams as features for their SemEval-2018 Task 1 focused on affect (emotion-related categories e.g., joy, fear, valence, and arousal) in tweets. Similarly, in [Van Hee et al., 2018a] an SVM TF-IDF word unigram baseline was implemented and it outperformed several techniques and overall performed very well for this task focused on irony detection in English tweets.

Supervised learning algorithms such as SVM, were also very popular in the SemEval task focused on sentiment analysis in Twitter [Rosenthal et al., 2017]. Participants in the SemEval-2017 Task 5 focusing on fine-grained sentiment analysis on financial microblogs and news [Cortis et al., 2017] i.e., the dataset presented in Section 3.3, made use of lexicon-based, machine learning, deep learning, and hybrid techniques. The machine learning and deep learning techniques used algorithms, such as NB, SVM, LR, RF, ANNs, CNN, and RNN. The authors in [Pontiki et al., 2016] trained an SVM classifier as a baseline for sentiment polarity detection, as part of SemEval-2016 Task 5 focused on aspect-based sentiment analysis.

A supervised learning approach using SVM was used in [Kothari et al., 2013] for subjectivity classification of news articles’ comments and tweets. In [Appidi et al., 2020], the authors used machine learning algorithms, such as SVM, to carry out emotion classification experiments on an annotated corpus of code-switched Kannada-English tweets. The authors in [Bansal et al., 2020] used SVM and RF for training baseline models to show how code-switching patterns can be used to improve several downstream NLP applications. In [Mamta et al., 2020], the authors also implemented baseline models for sentiment analysis using machine learning and deep learning algorithms, such as SVM and CNN. Similarly, the authors in [Yimam et al., 2020] built several baseline models for Amharic sentiment analysis from social media text using supervised learning algorithms, such as SVM and LR.

4.2.2 Neural Language Models

In [Nguyen et al., 2020], the authors developed the first large-scale pre-trained language model BERTweet for English tweets, which outperforms its baselines. Experiments were conducted on three NLP tasks: POS tagging, NER, and text classification specifically for sentiment analysis and irony detection. For the latter task, the authors used the 3-class sentiment analysis dataset from SemEval-2017 Task 4A [Rosenthal et al., 2017] and the 2-class irony detection dataset from the SemEval-2018 Task 3A [Van Hee et al., 2018b]. The authors in [Croce et al., 2020] propose GAN-BERT which extends the fine-tuning of architectures similar to BERT [Devlin et al., 2018], using unlabelled data in a generative

adversarial setting. Experimental results show that around 50-100 annotated examples can still produce good performance in sentence classification tasks. Results are confirmed for sentiment analysis over the SST-5 dataset [Socher et al., 2013] containing 5-class sentiment polarity categories. Babanejad et al. propose two novel deep neural network models for sarcasm detection by including affective and contextual features in the extended BERT architecture [Babanejad et al., 2020].

Certain studies focused on low-resourced languages, with [Fei and Li, 2020] investigating cross-lingual sentiment classification where the low-resource language does not have any labels or parallel corpus, [Grießhaber et al., 2020] exploring the reduction of trainable model parameters for fine-tuning a model with a small amount of data, [Koto et al., 2020] releasing a new pre-trained language model for Indonesian which was evaluated on several tasks such as sentiment analysis, and [Yimam et al., 2020] using RoBERTa [Liu et al., 2019] –a replication of BERT developed by Facebook– for exploring Amharic sentiment analysis from social media text. In [Demszky et al., 2020], the authors conducted transfer learning experiments on existing emotion benchmarks to show that the GoEmotions dataset of fine-grained emotions generalises across domains and taxonomies. The authors demonstrated that if little target domain labelled data is available, this dataset can be used as a baseline for emotion understanding. Similarly, the XED multilingual dataset for emotion detection catering for a total of 32 languages has been evaluated using language-specific BERT models [Öhman et al., 2020]. Lastly, [Makarencov and Rokach, 2020] explored several off-the-shelf BERT models, where they show that the complexity and computational cost of BERT does not provide a guarantee for an improved predictive performance for classification tasks. This is especially relevant in cases where small domain-specific datasets are used, which datasets are also imbalanced due to the minority class being under-represented.

The studies above demonstrate that the mostly used neural language model is BERT, whereas BERTweet is more applicable for NLP tasks that are based on social data.

4.3 Baseline Classification Models

This section discusses the baseline classification models generated for each social opinion dimension using machine learning algorithms and neural language models.

4.3.1 Dataset

The dataset of multidimensional and multilingual social opinions for Malta’s Annual Government Budget presented in Section 3.4, is used for the development of baseline classification models using traditional machine learning techniques and state-of-the-art deep neural network models. In terms of languages, the majority of the online posts were in English (74.09%) with most of the rest being in Maltese-English and Maltese (24.99%). It is important to note that the online posts in Maltese-English and Maltese have been merged together due to the low amount of online posts in Maltese only.

Dataset	All		English		Maltese-English and Maltese	
	Count	Percentage	Count	Percentage	Count	Percentage
Subjectivity						
Subjective (1)	2591	40.57%	1713	36.20%	852	53.38%
Objective (0)	3796	59.43%	3019	63.80%	744	46.62%
Sentiment Polarity						
Negative (0)	1232	19.29%	775	16.38%	441	27.63%
Neutral (1)	1605	25.13%	1355	28.63%	219	13.72%
Positive (2)	3550	55.58%	2602	54.99%	936	58.65%
Emotion						
Joy (0)	2636	41.27%	1976	41.76%	648	40.60%
Trust (1)	363	5.68%	219	4.63%	144	9.02%
Fear (2)	72	1.13%	61	1.29%	11	0.69%
Surprise (3)	177	2.77%	116	2.45%	60	3.76%
Sadness (4)	245	3.84%	176	3.72%	67	4.20%
Disgust (5)	498	7.80%	275	5.81%	216	13.53%
Anger (6)	369	5.78%	238	5.03%	127	7.96%
Anticipation (7)	2027	31.74%	1671	35.31%	323	20.24%
Sarcasm						
Sarcastic (1)	177	2.77%	101	2.13%	74	4.64%
Not Sarcastic (0)	6210	97.23%	4631	97.87%	1522	95.36%
Irony						
Ironic (1)	329	5.15%	189	3.99%	136	8.52%
Not Ironic (0)	6058	94.85%	4543	96.01%	1460	91.48%
Language						
English (0)	4732	74.09%	4732	100%		
Maltese (1)	299	4.68%			299	18.73%
Maltese-English (2)	1297	20.31%			1297	81.27%
Other (3)	59	0.92%				

Table 4.1: Class distribution for each annotation per dataset

Table 4.1 presents the overall class distribution of online posts for each social opinion

dimension and the language annotation. Statistics are provided for the entire dataset (columns 2 and 3), the subset of online posts in English (columns 4 and 5), and subset of online posts in Maltese-English and Maltese (columns 6 and 7).

4.3.2 Machine Learning-based Models

As discussed in Section 2.3.3.2, Section 4.2.1, and in [Cortis and Davis, 2021b], the NB, SVM, LR, DT, and RF supervised machine learning algorithms are the ones mostly used with success for Social Opinion Mining research, hence why they were chosen for the baseline classification models generated.

In this regard, baseline classification models for *subjectivity* (binary), *sentiment polarity* (multi-class), *emotion* (multi-class), *sarcasm* (binary), and *irony* (binary) were built at a monolingual level for English (using the English dataset) and at a code-switched level for Maltese-English (using the Maltese-English and Maltese dataset), using the following eight supervised learning algorithms:

- **NB**: MBNB –classifier suitable for discrete data and is designed for binary/boolean features [scikit learn, a], and Complement NB (CNB) –designed to correct “severe assumptions” made by the standard MNB classifier and suited for imbalanced datasets [scikit learn, b];
- **SVM**: SVC –C-SVC implementation based on libsvm (a library for SVM) [scikit learn, h], NuSVC –similar to SVC however it can control the number of support vectors [scikit learn, f], and Linear SVC –similar to SVC however it has more flexibility and supports both dense and sparse input [scikit learn, d];
- **LR**: a probabilistic classifier also known as logit or MaxEnt [scikit learn, e];
- **DT**: an optimised version of the Classification and Regression Trees (CART) algorithm [scikit learn, c]; and
- **RF**: an ensemble of decision tree algorithms [scikit learn, g].

4.3.3 Neural-based Models

With regard to neural-based models, BERT has been very successful when fine-tuned for certain NLP tasks, such as sentiment analysis, using social data as training data. Sections 2.3.10 and 4.2.2 provide an overview of Social Opinion Mining research that use BERT and is supported by other relevant studies [Polignano et al., 2019, Yang et al., 2019, Nozza et al., 2020, Kokab et al., 2022]. In particular, the BERTweet model has already been pre-trained on tweets, which data is very similar in nature to the dataset discussed in Section 4.3.1 which is used for the development of baseline classification models.

Based on the reasons mentioned above, the following state-of-the-art deep neural network models have been chosen for experimentation purposes, where they have been fine-tuned for *subjectivity* (binary), *sentiment polarity* (multi-class), *emotion* (multi-class), *sarcasm* (binary), and *irony* (binary) classification, at a multilingual level using the entire dataset:

- **BERT** [Devlin et al., 2018]: A pre-trained model on BookCorpus and English Wikipedia. The BERT-Base uncased, 12-layer, 768-hidden, 12-heads, 110M parameters model is used.
- **DistilBert** [Sanh et al., 2019]: A distilled version of the BERT model which is smaller and faster than BERT and is pre-trained on the data. The uncased model which has 40% less parameters than BERT-Base uncased is used.
- **BERTweet** [Nguyen et al., 2020]: A large-scale language model pre-trained for English tweets based on the RoBERTa [Liu et al., 2019] pre-training procedure using the same model configuration as BERT-Base. Both *bertweet-base* (base) and *bertweet-covid19-base-uncased* (covid-19) models with 135M parameters each are used. The former model is trained on 845M English cased tweets, whereas the latter model is trained on 23M COVID-19 English uncased tweets. The uncased model does not make a difference between a cased and uncased word e.g., *maltese* and *Maltese*.

4.4 Experiments

This section discusses the baseline classification models experiments carried out for each social opinion dimension using machine learning techniques and neural language models.

4.4.1 Machine Learning-based Models

All experiments have been carried out in the Python programming language using Jupyter Notebook⁴ on a machine with an Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz 1.99 GHz processor and 8.00 GB (7.88 GB usable) installed memory (RAM).

4.4.1.1 Pre-processing

Pre-processing on the online posts used for building the baseline models was carried out, using the following NLP tasks of a syntactic nature:

- **Data cleaning:** Removal of any numbers, HyperText Markup Language (HTML)/ Extensible Markup Language (XML) tags, special characters, and whitespaces;
- **Tokenisation:** text composed of string of words or sentences split into tokens, in terms of alphabetic and non-alphabetic characters, using the NLTK [Bird et al., 2009] word punctuation tokeniser;
- **Stemming:** removes suffices or prefixes used with a word to reduce inflectional forms to a common base form, using NLTK's implementation of the Porter stemming algorithm⁵; and
- **Conversion of textual data into numerical representations:** TF-IDF statistical measure (using the scikit-learn TfidfVectorizer function⁶) used to evaluate the word relevance and importance in online posts and hence represent the online posts into a feature vector for training a classifier using any algorithm discussed in Section 4.3.2.

⁴<https://jupyter.org/>

⁵<https://tartarus.org/martin/PorterStemmer/>

⁶https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

During the pre-processing steps mentioned above, data cleaning removes noise that does not bring any value from the online posts, tokenisation of the online posts is carried out so that the tokens can be used to convert a string (text) into numeric data for use by the machine learning models, stemming normalises the online posts in a uniform structure to identify word variations that do not bring any value (i.e., remove redundancy) when training machine learning models for predictions, and vectorisation turns tokens into meaningful numbers i.e., count the number of times each term appears in each online post where the relevant TF-IDF weights are applied for each online post. Pre-processing is also important given that the dataset used for carrying out the experiments is imbalanced (more information on this in Section 4.4.2.1). Given that machine learning classifiers are more sensitive in detecting the majority class and less sensitive in detecting the minority class, it is important to handle this issue for imbalanced data to ensure that the classification output is not biased towards the majority class [Padurariu and Breaban, 2019].

4.4.1.2 Model Generation

Given that the dataset used for building these baseline models is relatively small in terms of data volume, no chunk of data was omitted for model generation. Therefore, cross-validation provided a better modelling approach for small datasets, as opposed to the traditional training-validation-test set split. Stratified 10-fold cross-validation is applied on the entire dataset being used for model generation and evaluation. This cross-validation technique is used since the ratio between the target classes is preserved (keeps the same percentage of samples for each class) as is in the full dataset. Moreover, it is also adequate for imbalanced datasets such as the one being used, as reflected in Table 4.1. It is important to note that this technique just shuffles and splits the dataset once into 10 folds. Therefore, the test sets used for validating the trained model (on $k - 1$ of the folds used as training data) do not overlap between any of the 10 splits. Lastly, the model itself is trained 10 times, with the weights and any biases being reset with each new model. This cross-validation procedure was applied for each baseline model built using the supervised learning algorithms discussed in Section 4.3.2.

Baseline classification models for **subjectivity, sentiment polarity, emotion, sar-**

casm, and **irony**, were built on i) the subset of English online posts and ii) the subset of Maltese-English and Maltese online posts, which results are reported in Section 4.5.1. The scikit-learn machine learning library was used for building these baseline models.

4.4.2 Neural-based Models

All experiments have been carried out in the Python programming language using Jupyter Notebook on Google Colaboratory⁷ using a Tesla K80/Tesla T4/Tesla P100-PCIE-16GB GPU.

4.4.2.1 Handling Imbalanced Data

As reflected in Table 4.1, the dataset used is imbalanced. Data belonging to the majority of the social opinion dimensions is not evenly distributed among their respective classes, with the emotion, irony, and sarcasm dimensions being mostly visible. Most machine learning and deep learning classification algorithms, such as LR and DTs, are not capable of handling imbalanced classes and therefore end up being biased towards majority classes [Padurariu and Breaban, 2019], with classes having less data points being ignored and/or considered as noise. As a result, trained classification models are not able to predict the correct outcome of unseen data, since there is not enough data for the model to learn from. There are several re-sampling techniques [Cateni et al., 2014, More, 2016] for treating the problem of an imbalanced dataset. In the initial experiments carried out the problem of class imbalance is not addressed. Nor do we examine whether it influences the classification tasks carried out and if so, which ones. Given that good results and performance have been obtained for certain NLP tasks when existing transformer-based models were fine-tuned with a small amount of data (see discussion in Section 4.2.2 for more information), the entire dataset discussed in Section 4.3.1 is used for these experiments. In this regard, the dataset was divided in a training set of 70%, validation set of 20%, and a test set of 10%. The scikit-learn `train_test_split` function is used to split the sets in a random state.

⁷<https://colab.research.google.com/>

4.4.2.2 Model Generation

The experiments were carried out using the Hugging Face [Wolf et al., 2019] state-of-the-art Transformer library for Pytorch and TensorFlow 2.0. This tool provides general-purpose architectures, such as BERT, RoBERTa, and DistilBert for NLP tasks such as sentiment analysis, where over 32+ pre-trained models are available in 100+ languages.

The following hyperparameters are used:

- Optimiser and learning rate scheduler: batch size - 32, Adam [Kingma and Ba, 2014] learning rate - $2e-5$, number of epochs - 4, epsilon parameter - $1e-8$;
- Method of choosing values and criterion used: Manual tuning based on training and validation loss, learning rates of $5e-5$, $3e-5$, $2e-5$, and maximum sentence length of 96, 128, and 256 tokens; and
- Fine-tuning classification layer: Rectified Linear Unit (ReLU) [Agarap, 2018].

The hyperparameters chosen above are based on existing literature [Devlin et al., 2018, Sun et al., 2019] and initial experiments carried out. Baseline classification models for **subjectivity**, **sentiment polarity**, **emotion**, **sarcasm**, and **irony**, were built on the entire dataset of English, Maltese, and Maltese-English online posts, which results are reported in Section 4.5.2.

4.5 Results and Discussion

This section presents the results of the experiments carried out for each social opinion dimension using machine learning techniques and neural language models.

4.5.1 Machine Learning-based Models

Results of the baseline classification models for **subjectivity**, **sentiment polarity**, **emotion**, **sarcasm**, and **irony**, built on i) the subset of English online posts and ii) the subset of Maltese-English and Maltese online posts (mentioned in Section 4.4.1.2) are presented and discussed in this section. Table 4.2 displays results obtained on the subset of online posts in English, whereas Table 4.3 displays results obtained on the subset of online posts

in Maltese-English and Maltese (merged together due to the low amount of online posts in Maltese only). Note that results for both tables have been rounded to three decimal places, where the full evaluation results are available in Appendix B. The following evaluation metrics were used to measure the classification performance of the models generated for each social opinion dimension:

- **Precision weighted** [Allen et al., 1955]: Precision is the ratio of true positives over the number of true positives plus the number of false positives. It is a measure of result relevancy in terms of accurate results. The weighted score calculates the precision score for each label with their average being weighted by support, that is, the number of true instances for each label. This metric caters for label imbalance.
- **Recall weighted** [Allen et al., 1955]: Recall is the ratio of true positives over the number of true positives plus the number of false negatives. It is a measure of how many truly relevant results are returned. The weighted score calculates the recall score for each label with their average being weighted by support, that is, the number of true instances for each label. This metric caters for label imbalance.
- **F1 score weighted** [Chinchor, 1992]: F1 score is the weighted average of precision and recall. The weighted score calculates the F1 score for each label with their average being weighted by support, that is, the number of true instances for each label. This metric caters for label imbalance.
- **Balanced accuracy** [Brodersen et al., 2010]: This metric is defined as the average of recall scores obtained per class. This metric is used for imbalanced binary and multi-class classification.

Tables 4.2 and 4.3 present the weighted average precision (precision weighted), weighted average recall (recall weighted), weighted average F1 score (F1 score weighted), and balanced accuracy results obtained for all eight supervised learning algorithms using the stratified 10-fold cross-validation technique.

With respect to the English data (see Table 4.2), the LR algorithm obtained the best precision weighted, recall weighted, and F1 score weighted results for the subjectivity classification model. The CNB algorithm produced the best precision weighted, recall

weighted, and F1 score weighted results for the sentiment polarity and emotion classification models. The MBNB obtained the best precision weighted results for both the irony and sarcasm classification models. Moreover, the RF obtained the best recall weighted and F1 score weighted results for the irony classifier, whereas NuSVC fared best for the sarcasm classifier in terms of recall weighted and F1 score weighted. When considering the balanced accuracy, the CNB algorithm produced the best results for all the social opinion dimensions.

Opinion Dimension	LR	Linear SVC	Nu SVC	SVC	MBNB	CNB	DT	RF
Subjectivity								
Precision weighted	0.899	0.893	0.898	0.407	0.872	0.898	0.850	0.886
Recall weighted	0.891	0.884	0.886	0.638	0.855	0.888	0.844	0.880
F1 score weighted	0.884	0.880	0.876	0.497	0.840	0.884	0.837	0.872
Balanced accuracy	0.867	0.867	0.855	0.50	0.812	0.874	0.825	0.853
Sentiment Polarity								
Precision weighted	0.795	0.778	0.799	0.302	0.796	0.810	0.735	0.782
Recall weighted	0.783	0.772	0.784	0.550	0.777	0.786	0.730	0.771
F1 score weighted	0.773	0.767	0.777	0.390	0.777	0.783	0.730	0.762
Balanced accuracy	0.723	0.717	0.739	0.333	0.727	0.767	0.684	0.712
Emotion								
Precision weighted	0.548	0.567	0.560	0.174	0.551	0.622	0.532	0.536
Recall weighted	0.604	0.596	0.60	0.418	0.598	0.605	0.542	0.587
F1 score weighted	0.559	0.573	0.566	0.246	0.559	0.598	0.529	0.539
Balanced accuracy	0.248	0.283	0.268	0.125	0.249	0.319	0.273	0.227
Irony								
Precision weighted	0.921	0.928	0.922	0.922	0.943	0.935	0.941	0.926
Recall weighted	0.960	0.957	0.959	0.960	0.907	0.923	0.936	0.960
F1 score weighted	0.940	0.940	0.940	0.940	0.917	0.928	0.933	0.941
Balanced accuracy	0.50	0.511	0.50	0.50	0.536	0.562	0.560	0.50
Sarcasm								
Precision weighted	0.958	0.961	0.961	0.958	0.971	0.964	0.966	0.960
Recall weighted	0.979	0.977	0.979	0.979	0.950	0.946	0.966	0.979
F1 score weighted	0.968	0.968	0.969	0.968	0.957	0.954	0.962	0.968
Balanced accuracy	0.50	0.508	0.509	0.50	0.558	0.566	0.540	0.50

Table 4.2: Classification model results - English dataset

As for the results on the Maltese-English and Maltese data (see Table 4.3), the CNB algorithm fared best in terms of F1 score weighted for the subjectivity, emotion (same as for English data), and irony classification models. The same algorithm produced the best recall weighted result for the subjectivity classification model and the best precision weighted results for the emotion, irony, and sarcasm classification models. The LinearSVC algorithm produced the best precision weighted and F1 score weighted result for the sentiment polarity classifier. The RF algorithm produced the best precision weighted result for subjectivity and the best recall weighted results for the sentiment polarity and emo-

tion classifiers. Moreover, the LR and SVC algorithms obtained the best and same recall weighted results for both the irony and sarcasm classification models, whereas the same algorithms produced the best and same F1 score weighted results for sarcasm. Similar to the results obtained on the English data, the CNB algorithm produced the best balanced accuracy results for subjectivity, irony, and sarcasm. On the other hand, LinearSVC obtained the best balanced accuracy results for both sentiment polarity and emotion.

Opinion Dimension	LR	Linear SVC	Nu SVC	SVC	MBNB	CNB	DT	RF
Subjectivity								
Precision weighted	0.862	0.861	0.867	0.285	0.858	0.887	0.856	0.891
Recall weighted	0.844	0.844	0.849	0.534	0.791	0.860	0.829	0.850
F1 score weighted	0.840	0.841	0.846	0.372	0.773	0.855	0.840	0.848
Balanced accuracy	0.844	0.843	0.848	0.50	0.803	0.864	0.838	0.854
Sentiment Polarity								
Precision weighted	0.712	0.754	0.751	0.344	0.597	0.733	0.738	0.745
Recall weighted	0.733	0.760	0.752	0.586	0.659	0.745	0.728	0.760
F1 score weighted	0.689	0.740	0.725	0.434	0.593	0.725	0.714	0.728
Balanced accuracy	0.562	0.638	0.619	0.333	0.450	0.613	0.601	0.622
Emotion								
Precision weighted	0.387	0.431	0.378	0.165	0.306	0.460	0.397	0.428
Recall weighted	0.466	0.462	0.436	0.406	0.438	0.478	0.414	0.491
F1 score weighted	0.377	0.427	0.376	0.234	0.314	0.433	0.403	0.423
Balanced accuracy	0.205	0.275	0.241	0.125	0.163	0.255	0.235	0.264
Irony								
Precision weighted	0.837	0.869	0.837	0.837	0.853	0.899	0.874	0.837
Recall weighted	0.915	0.914	0.912	0.915	0.868	0.886	0.882	0.914
F1 score weighted	0.874	0.879	0.873	0.874	0.856	0.884	0.873	0.875
Balanced accuracy	0.50	0.519	0.499	0.500	0.508	0.611	0.563	0.503
Sarcasm								
Precision weighted	0.909	0.909	0.909	0.909	0.920	0.923	0.911	0.909
Recall weighted	0.954	0.951	0.953	0.954	0.930	0.901	0.927	0.953
F1 score weighted	0.931	0.930	0.931	0.931	0.921	0.910	0.917	0.931
Balanced accuracy	0.50	0.499	0.50	0.50	0.507	0.530	0.503	0.50

Table 4.3: Classification model results - Maltese-English and Maltese dataset

The following are some observations on the results on why some models perform better for certain dimensions and/or different languages:

- The CNB algorithm obtained good performance for all languages and handled the imbalanced classes better than the other algorithms.
- Results obtained for the subjectivity and sentiment polarity classifiers are very promising for the English subset and Maltese-English and Maltese subset, even though the latter subset only amounts to 1,596 online posts and the classes are not evenly balanced (for both subsets).

- Further evaluation using online posts unseen by the trained models is needed on the emotion, irony, and sarcasm classifiers to ensure that they are not biased towards the majority classes [Padurariu and Breaban, 2019], due to small amount of online posts available for the minority classes. Re-sampling techniques [Cateni et al., 2014, More, 2016] such as over-sampling and under-sampling can be used for handling such imbalances.

4.5.2 Neural-based Models

Results per classification model for each social opinion dimension are presented in Table 4.4 and further discussed below. The full evaluation results are available in Appendix C.

Five evaluation metrics were used to measure the classification performance of the fine-tuned models:

- **Precision weighted** [Allen et al., 1955]: Precision is the ratio of true positives over the number of true positives plus the number of false positives. It is a measure of result relevancy in terms of accurate results. The weighted score calculates the precision score for each label with their average being weighted by support, that is, the number of true instances for each label. This metric caters for label imbalance.
- **Recall weighted** [Allen et al., 1955]: Recall is the ratio of true positives over the number of true positives plus the number of false negatives. It is a measure of how many truly relevant results are returned. The weighted score calculates the recall score for each label with their average being weighted by support, that is, the number of true instances for each label. This metric caters for label imbalance.
- **F1 score weighted** [Chinchor, 1992]: F1 score is the weighted average of precision and recall. The weighted score calculates the F1 score for each label with their average being weighted by support, that is, the number of true instances for each label. This metric caters for label imbalance.
- **Area Under the Curve Receiver Operating Characteristics (AUC ROC)** [Hanley and McNeil, 1982]: Score shows the model's true positive rate against the false positive rate and can indicate how well (score of 1) a model can distinguish

between classes⁸.

- **Matthews correlation coefficient (MCC)** [Matthews, 1975]: Measures quality of binary and multi-class classifications by taking into account true and false positives and negatives and provides a balanced measure for imbalanced classes.

	Precision weighted	Recall weighted	F1 score weighted	AUC ROC	MCC
Subjectivity					
BERT	0.930	0.930	0.930	0.983	0.864
DistilBERT	0.930	0.930	0.930	0.980	0.851
BERTweet (base)	0.930	0.930	0.930	0.970	0.857
BERTweet (covid19)	0.950	0.950	0.940	0.975	0.887
Sentiment Polarity					
BERT	0.850	0.850	0.850	1 - 0.945	0.748
DistilBERT	0.830	0.830	0.830		0.710
BERTweet (base)	0.870	0.860	0.860	0 - 0.961	0.772
BERTweet (covid19)	0.870	0.870	0.870	2 - 0.964	0.781
Emotion					
BERT	0.670	0.650	0.60	3 - 0.935 4 - 0.847 5 - 0.914 6 - 0.894	0.495
DistilBERT	0.660	0.640	0.60	0 - 0.913 1 - 0.821 7 - 0.882	0.484
BERTweet (base)	0.680	0.630	0.580	2 - 0.862	0.478
BERTweet (covid19)	0.760	0.650	0.590		0.501
Irony					
BERT	0.930	0.950	0.930	0.883	0.179
DistilBERT	0.950	0.950	0.930	0.896	0.240
BERTweet (base)	0.950	0.950	0.920	0.862	0
BERTweet (covid19)	0.950	0.950	0.920	0.887	0
Sarcasm					
BERT	0.950	0.960	0.960	0.858	0.073
DistilBERT	0.970	0.970	0.960	0.879	0.265
BERTweet (base)	0.970	0.970	0.960	0.873	0
BERTweet (covid19)	0.970	0.970	0.960	0.792	0

Table 4.4: Evaluation results of all the pre-trained models for each social opinion dimension

The following is an overview of the results and some observations:

- **Subjectivity:** For the BERT and DistilBERT models, the training and validation loss converged in epoch 2, whereas both BERTweet models converged in epoch 3. The BERTweet covid19-base-uncased fine-tuned model produced the best performance overall.
- **Sentiment Polarity:** The fine-tuned BERT and DistilBERT models converged in

⁸For the sentiment polarity and emotion multi-class models only the maximum AUC ROC score for each respective class is displayed.

epoch 3, whereas both BERTweet models converged in epoch 4. The BERTweet covid19-base-uncased fine-tuned model also produced the best performance overall.

- **Emotion:** The fine-tuned BERT and DistilBERT models converged in epoch 4, whereas both BERTweet models did not converge by epoch 4 albeit close. An additional experiment showed convergence in epoch 6. In terms of performance, both BERT and DistilBERT fared best overall.
- **Sarcasm:** All fine-tuned models performed similarly in terms of F1 score, with DistilBERT performing best overall. The BERTweet covid19-base-uncased model did not converge in epoch 4 albeit close.
- **Irony:** DistilBERT produced the best results overall, which model converged in epoch 3.
- **Language:** It is interesting to see English-based fine-tuned models adapt to non-English text. This Maltese-English and Maltese subset amounts to only a quarter of the dataset (1596 online posts). Initial results obtained are promising for building language models that are capable of handling code-switched data, which is common practice in countries like Malta. More in-depth experiments and qualitative analysis shall be beneficial to measure the adaptability of the English-based fine-tuned models to code-switched languages, such as Maltese-English.
- **Domain:** A Socio-Economic and Government dataset (domain specific) has been used, with only 16.75% of the data being off-topic. The results obtained in the preliminary work carried out demonstrate that fine-tuning models to new domains is possible when using deep neural network models.
- The DistilBERT model took less time for training and validation for all five classifiers.
- Even though the dataset is imbalanced, the *subjectivity* and *sentiment polarity* models produced good results. However, certain re-sampling techniques shall help increase the performance of the *sarcasm* and *irony* fine-tuned models, which class distribution is very unbalanced as reflected by the MCC. The same also applies to the emotion model for certain classes, such as fear, surprise, sadness, and anger.

- Several researchers recommend only 2-4 epochs of training for fine-tuning BERT on a particular NLP task. However, certain multi-class classification tasks with a large number of classes, such as the emotion 8-class classification fine-tuned model, might require more than 4 epochs when certain models such as BERTweet are fine-tuned.
- Given that the dataset used contains a mix of newswire comments and tweets, the maximum sentence length in the dataset used is 867. Therefore, more experiments should be carried out using a higher maximum sentence length than the 128 tokens used. However, the high computation power needed for training such deep learning models should be taken in consideration to reduce the carbon footprint in terms of finance and the environmental [Strubell et al., 2019].

4.6 Impact

Even though the baseline classification results presented are a work-in-progress, the work by [Xia et al., 2020] encourages researchers to provide multilingual benchmarks which can be further used, evaluated, and adapted for low-resourced languages, such as Maltese. Classification models capable of understanding English and Maltese data, both being Malta's official languages, can be used by governments for policy formulation, policy-making, decision-making, and decision-taking. Multidimensional Social Opinion Mining carried out on user-generated content shall provide a nuanced voice to the citizens and residents of Malta on certain topics and issues of a national importance, with the aim of leaving a positive impact on society at large. Research applications for the developed classification models include opinion summarisation and fine-grained opinionated search of each social opinion dimension.

4.7 Conclusion

This chapter discusses baseline classification models that were developed using traditional machine learning techniques and state-of-the-art deep neural network models. These were developed to validate the dataset of multidimensional and multilingual social opinions for Malta's Annual Government Budget which was presented in Section 3.4. Different baseline

classification models were developed for each social opinion dimension, namely as subjectivity, sentiment polarity, emotion, irony, and sarcasm, and for different language levels, namely monolingual for English –a high-resourced language, code-switched for Maltese –a low-resourced language and Maltese-English –a code-switched language, and multilingual for the English, Maltese, and Maltese-English languages. In this regard, the supervised machine learning algorithms and deep neural network models used are discussed together with the results obtained and some observations made. Overall, the CNB supervised machine learning algorithm produced the best balanced accuracy results for the majority of the classification models. As for the neural language models, the following fine-tuned models produced the best performance overall: BERTweet covid19-base-uncased for subjectivity and sentiment polarity, BERT and DistilBERT models for emotion, and DistilBERT for both sarcasm and irony. The content of this chapter answers **Research Questions 3 and 5**. The next chapter discusses whether fine-grained opinion search and summarisation at an aspect-based and multidimensional level provides added-value for a real-world use case, namely the Malta Government Budget.

Chapter 5

Multidimensional Opinion Search and Summarisation

5.1 Introduction

This chapter focuses on **fine-grained opinion search and summarisation** at a multidimensional level for a real-world use case, namely the Malta Government Budget. The main **objective** of this research study is to evaluate whether a multidimensional Social Opinion Mining approach at an aspect-based (or feature-based) level affects fine-grained opinion search and summarisation and whether it can be applied to the Government domain. Multidimensional opinion summarisation provides a nuanced voice to the general public i.e., citizens and residents of Malta, on their opinions about a topic that is related to the annual Malta Government Budget. The following are the main contributions presented in this chapter:

- Opinion summarisation literature focuses on opinion summaries based on the sentiment polarity opinion dimension only and not on multiple opinion dimensions. The multidimensional opinion summarisation work presented here is based on **multiple opinion dimensions**, namely *subjectivity*, *sentiment polarity*, *emotion*, *irony*, and *sarcasm*. This is the first attempt to explore fine-grained opinion summaries in a multidimensional context.
- An **Opinion Search and Summarisation system** developed, in order to de-

termine whether multidimensional Social Opinion Mining at an aspect-based level benefits users engaged in fine-grained opinion search and summarisation. In this regard, **three** variations of the system based on different social opinion dimensions, namely i) subjectivity and sentiment polarity (**System 1**); ii) subjectivity, sentiment polarity, and emotion (**System 2**); and iii) subjectivity, sentiment polarity, emotion, and sarcasm or irony (**System 3**), have been evaluated for the purposes of this study.

- Evaluation of opinion summarisation is usually carried out using automatic metrics against a ground-truth or reference summaries. A proper **human evaluation** comprising of an *extrinsic* evaluation, *usability* evaluation, and *subject matter expert* evaluation was carried out to evaluate whether the multidimensional opinion summarisation work presented can help in policy formulation, policy-making, decision-making, and decision-taking. Therefore, automatic metrics are not suitable for such a purpose.
- The **5-point Mean Opinion Score** [Strejil et al., 2016] has been used to *extrinsically* evaluate the quality of the three Opinion Search and Summarisation systems mentioned above. This continuous scale has been used in a different application than its traditional use as a measure of media quality. Therefore, this work goes beyond the state-of-the-art and has been adopted in an innovative and real-world setting to evaluate aspect-based opinion summaries returned by each of the three aforementioned systems (**System 1**, **System 2**, **System 3**), in terms of usefulness and informativeness.
- Four different measures, namely *summary usefulness*, *source usefulness*, *online post usefulness*, and *summary informativeness*, based on three extrinsic quality measures were used to evaluate the extrinsic quality of the three Opinion Search and Summarisation systems mentioned above (**System 1**, **System 2**, **System 3**). In this regard, an **additional extrinsic quality measure** was introduced to be able to cater for the measurement of online post usefulness that is presented as part of each aspect-based opinion summary returned by each of the three systems.

- The **System Usability Scale** [Brooke et al., 1996] has been used to evaluate the *usability* of the generic Opinion Search and Summarisation system developed.

The human evaluation consisting of an extrinsic evaluation, usability evaluation, and a subject matter expert evaluation was carried out to determine whether an Opinion Search and Summarisation system can be used for such a specific purpose, i.e., to help in any decision-making and/or policy-making processes. This human evaluation is the appropriate method to measure the impact and quality of the generated opinion summaries, since automatic metrics are not suitable for such a purpose [van der Lee et al., 2021]. The main reasons being that automatic metrics are not informative in terms of observed differences by humans and when the goal is to assess real-world use cases, the evaluations should address the usefulness of the system for which a human evaluation is the gold standard [van der Lee et al., 2021]. The opinion summaries are generated for the Government of Malta by the developed Opinion Search and Summarisation system, where each opinion summary is composed of different social opinion dimensions, namely *subjectivity*, *sentiment polarity*, *emotion*, *irony*, and *sarcasm*.

This chapter is in part based on the following publication:

- **Keith Cortis**, Brian Davis. “Social Opinion Search and Summarisation: A Multi-dimensional Approach.” (*under review*). 2023.

5.2 Opinion Summarisation Literature

The Opinion Mining research area has several sub-topics, such as sentiment classification, aspect-based Opinion Mining, opinion spam detection, opinion search, and opinion summarisation. This section further explores the latter two sub-topics which build on the Social Opinion Mining literature analysis presented in Chapter 2. Given that the opinion summarisation work presented in this chapter focuses on aspect-based opinion summarisation of social data, this section initially presents some background on automatic summarisation (refer to Section 5.2.1) followed by related work on opinion summarisation of social data (refer to Section 5.2.2), literature on aspect-based (or feature-based) opinion summarisation (refer to Section 5.2.3), and how the proposed work advances over

aspect-based opinion summarisation of social data (refer to Section 5.2.4).

5.2.1 Automatic Summarisation

Automatic summarisation focuses on developing automatic methods to produce a summary output based on text [Orăsan, 2019]. It is a challenging NLP task which can be carried out using a number of approaches such as statistical and hybrid. This research area has benefitted from related work in the areas of information retrieval and question answering and has been evaluated at both an intrinsic and extrinsic level. Intrinsic evaluation focuses on the extent that a system meets its objectives, whereas extrinsic evaluation focuses to the system's functional effectiveness in context [Jones, 2007]. The increasing interest in the research areas of Opinion Mining and Sentiment Analysis has resulted in the need for opinion summaries (also referred to as sentiment-based summaries) [Orăsan, 2019]. The research area of opinion summarisation builds upon several related areas, such as subjectivity detection, sentiment classification, emotion classification, and text summarisation [Kim et al., 2011]. In this section the most important research areas for the work presented in this thesis are discussed, namely text summarisation (Section 5.2.1.1) and opinion summarisation (Section 5.2.1.2).

5.2.1.1 Text Summarisation

Automated text summarisation is the process of producing a textual summary from one or more texts, where certain information from the original text(s) are provided [Hovy and Lin, 1999]. There are two types of automatic summarisation methods, namely *extractive summarisation* and *abstractive summarisation*. Extractive summaries are made up of text segments like sentences from the original text. On the other hand, abstractive summaries do not use text segments from the input data, which approach analyses the original text to generate sentences [Kim et al., 2011].

5.2.1.2 Opinion Summarisation

Opinion summarisation is the process of mining opinionated text from data sources, such as blogs and review articles, to produce a short and easy-to-read summary of a large num-

ber of opinions [Kim et al., 2011]. Initial applications carried out opinion summarisation from textual consumer reviews of products as made available on dedicated review sites [Hu and Liu, 2006]. There are different opinion summarisation techniques, namely *aspect-based* and *non-aspect-based*. An aspect-based opinion summarisation technique usually consists of the following main processes: aspect identification (identification of the main topics), sentiment prediction (determination of the sentiment polarity of the text about the identified topics), and summary presentation (aggregation of sentiment polarity ratings and presentation of opinion summaries). This shall be discussed further in Section 5.2.3. The following are non-aspect-based opinion summarisation techniques, some of which can be combined together: simple opinion classification summarisation, textual summary, entity-based summary, and visualisation [Kim et al., 2011].

Moreover, it is important to mention that opinion summarisation outputs can be textual-based, non-textual-based, or a combination of both. Liu mentions that even though an opinion summary can be in many forms (e.g., structured summary or textual summary), the main components of a summary should contain opinions on different entities and their aspects and should also have a quantitative perspective [Liu, 2012]. In fact, opinion summarisation is different than text summarisation, where the primary aim is to find the main features e.g., product features, that have been commented within the text being evaluated, e.g., multiple online product reviews. An opinion summary is usually based on sentiment classification results extracted from the text being evaluated, where the sentiment polarity is determined as being positive, negative, or neutral [Hu and Liu, 2006]. This is also supported by [Balahur et al., 2009, Balahur et al., 2012] who demonstrated that opinion summarisation is different than content-based summarisation.

5.2.2 Opinion Summarisation of Social Data

Opinion summarisation carried out on social data brings several challenges, such as unstructured data, noisy data, emoticons, sarcasm, irony, non-dictionary-standard words, numerous mistakes (spelling, grammar, punctuations, capitalisation), text in multiple languages other than English, handling of diverse and conflicting opinions, dealing with change of opinions over time, lack of available datasets, and large-scale data to process

[Pecar, 2018, Moussa et al., 2018]. Even though social data has numerous challenges that should be catered for as previously mentioned, the main focus of this research is on opinion summarisation of social data. The main reason behind this specific focus on social data is due to the potential of mining people’s social opinions for further use in multiple application areas, such as the Government domain. Initial opinion summarisation of social data focused on news and blog articles. In [Ku et al., 2006], text-based brief or detailed summaries categorised by sentiment polarity and topic are provided. Moreover, the same work provided a time series graph-based summary as part of their opinion tracking system. In [Potthast and Becker, 2010], the authors introduced the OPINIONCLOUD technology to summarise and visualise opinions expressed as Web comments from YouTube videos and Flickr images. As visualised in Figure 5.1, these opinion summaries consisted of the percentages of positive and negative terms from all non-neutral terms, together with a tag cloud of the respective words based on their occurrence frequency. The original comments are made available by clicking on a word within the list.

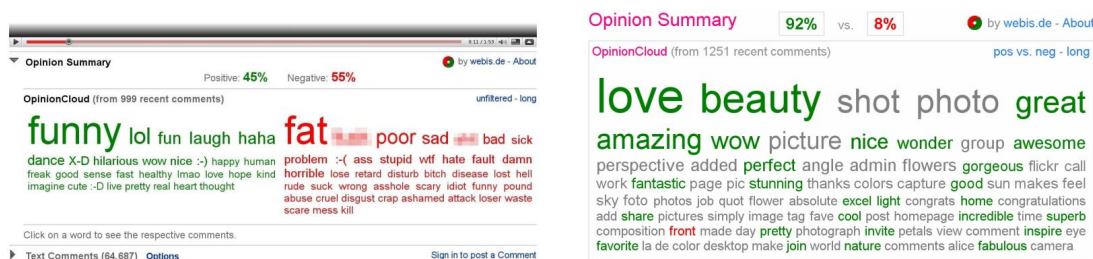


Figure 5.1: OPINIONCLOUD opinion summary as extracted from [Potthast and Becker, 2010]

The authors in [Balahur et al., 2012] present different approaches to perform opinion summarisation from blogs and reviews, with the focus being on sentiment polarity in terms of positive and negative opinions about a particular topic. Two summaries are produced for each blog thread, one for positive posts and another for negative posts, where sentences for the respective sentiment polarity are passed to the summariser [Balahur et al., 2009, Balahur et al., 2012]. Such summarisation approaches can be based on sentiment classification, sentiment intensity, or on topic-sentiment analysis and semantic information. Moreover, this work identified how opinions can be summarised in a way that they can be used in real-life applications, such as Marketing and decision-making [Balahur et al.,

2012]. The rise of social media platforms, such as Facebook and Twitter, have resulted in numerous studies focusing on opinion summarisation on social data, such as tweets [Meng et al., 2012, Bahrainian and Dengel, 2013, Trupthi et al., 2017, Nguyen et al., 2018, Ali et al., 2020].

5.2.3 Aspect-based Opinion Summarisation

Aspect-based (or feature-based) opinion summarisation is different in nature from traditional text summarisation, which process involves three steps: (i) identification of the main aspects of a particular entity (e.g., product) that have been commented on by users; (ii) identification of the opinion in terms of sentiment polarity (positive, negative, neutral) in each comment; and (iii) summarisation of the opinion results. The opinion summary for each aspect produces statistical values for the number of positive and negative opinions, where a link to the respective sentences and/or comments are also provided for each sentiment polarity value [Hu and Liu, 2004]. Figure 5.2 provides an illustration of a feature-based summary of a particular digital camera, named *Digital_camera_1* based on customer reviews.

```

Digital_camera_1:
  Feature: picture quality
    Positive: 253
              <individual review sentences>
    Negative: 6
              <individual review sentences>
  Feature: size
    Positive: 134
              <individual review sentences>
    Negative: 10
              <individual review sentences>
  ...

```

Figure 5.2: Example of a feature-based summary as extracted from [Hu and Liu, 2004]

The “Opinion Observer” opinion analysis and comparison system, presented in Figure 5.3, contains a visualisation component where the visualisation is highly tailored to the use case in question i.e., aspect-based opinion comparison of three cell phones from three different brands. Here the textual reviews are displayed together with the total number of positive and negative opinions or the percent of positive and negative opinions for each feature [Liu et al., 2005].

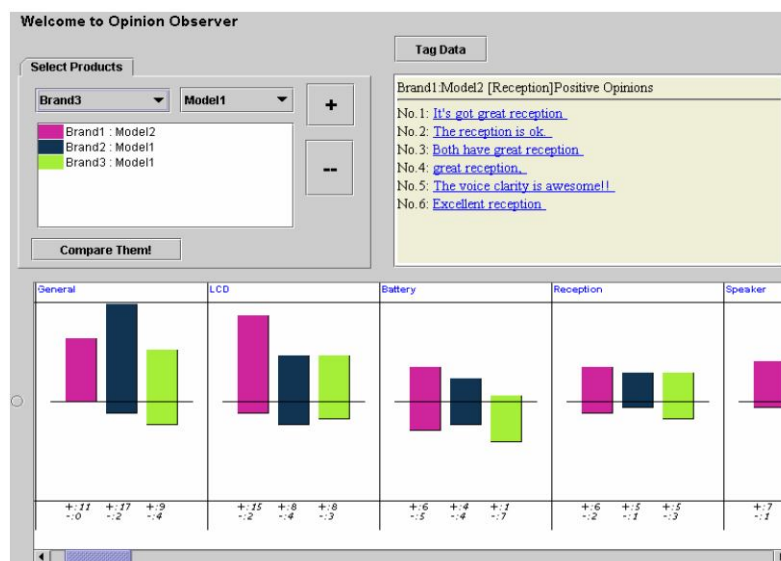


Figure 5.3: Opinion Observer’s comparison screen as extracted from [Liu et al., 2005]

In [Meng et al., 2012], an entity-centric topic-oriented opinion summarisation framework is presented. Each opinion summary includes three key elements, namely the topic, sentiment polarity, and insight, where the most relevant tweets are selected as the final opinion summary for each topic. The quality of the opinion summaries was evaluated on a set of opinion summaries generated from tweets and also in a real-life dataset of tweets. Bahrainian and Dengel present a solution for target-oriented (aspect-based) sentiment summarisation of tweets, where the opinion summary as represented in Figure 5.4 consists of the percentage for each sentiment polarity value and a textual summary of the review sentences [Bahrainian and Dengel, 2013].

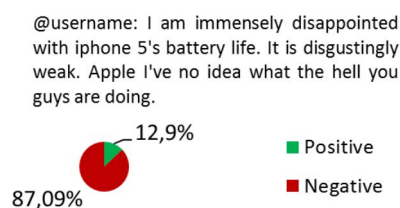


Figure 5.4: A sample opinion summary generated the sentiment summarisation system as extracted from [Bahrainian and Dengel, 2013]

The opinion summarisation framework in [Mukherjee et al., 2020] generates an opinion score for each sentence based on three criteria: readability, sentiment strength, and aspect-relevance. The aspect-based opinion summaries presented are personalised and use an

unsupervised extractive summarisation framework with the aim of minimising redundant information and maximising the collective salience of the relevant subset of sentences. The Aspect Discovery for OPinion Summarisation (ADOPS) methodology proposed by [López et al., 2021] generates structured and explainable opinion summaries about a single entity which are built on human interpretable rules. The aim behind the explainable aspect-based opinion summaries generated by the ADOPS methodology is to apply other data mining techniques and automatically generate reports on the opinion state of a single entity. The research presented in [Condori and Pardo, 2017] provides several extractive (see Figure 5.5) and abstractive (see Figure 5.6) methods for carrying out aspect-based opinion summarisation. Both methods are evaluated where the results indicate that traditional summarisation approaches are not suitable for the generation of opinion summaries.

```

Aspect: Samsung Smart TV
  Positive Sentences: 16
  - The best tv nowadays!
  Negative Sentences: 11
  - Its screen burned 1 month after the warranty.
Aspect: Price
  Positive Sentences: 0
  Negative Sentences: 2
  - What I did not like: Price too high
Aspect: Durability
  Positive Sentences: 0
  Negative Sentences: 2
  - Durability 0
Aspect: Camera
  Positive Sentences: 1
  - The camera with motion sensor works well and impresses who do not know it.
  Negative Sentences: 1
  - What I did not like: The camera could have horizontal movement of vision
  and not only just vertical.
Aspect: Image Quality
  Positive Sentences: 1
  - Excellent TV with great image quality and features.
  Negative Sentences: 0

```

Figure 5.5: An example of an extractive opinion summary as extracted from [Condori and Pardo, 2017]

```

Most reviews of the Samsung Smart TV are favorable. On the other hand, in relation to the price, users did not like it because it is high. In contrast, regarding the design, it was rated as excellent as it is sophisticated and modernity. Furthermore, with respect to the camera, they expressed controversial opinions on this feature because it works well and impresses who do not know it, but it could have horizontal movement of vision and not only just vertical. In addition, with respect to image quality, the reviews were excellent because it is great.

```

Figure 5.6: An example of an abstractive opinion summary as extracted from [Condori and Pardo, 2017]

Hayashi et al. proposed WikiAsp, a large-scale dataset based on Wikipedia, for multi-domain aspect-based opinion summarisation based on the model presented in Figure 5.7, with the aim of advancing research of open-domain aspect-based summarisation [Hayashi et al., 2021].

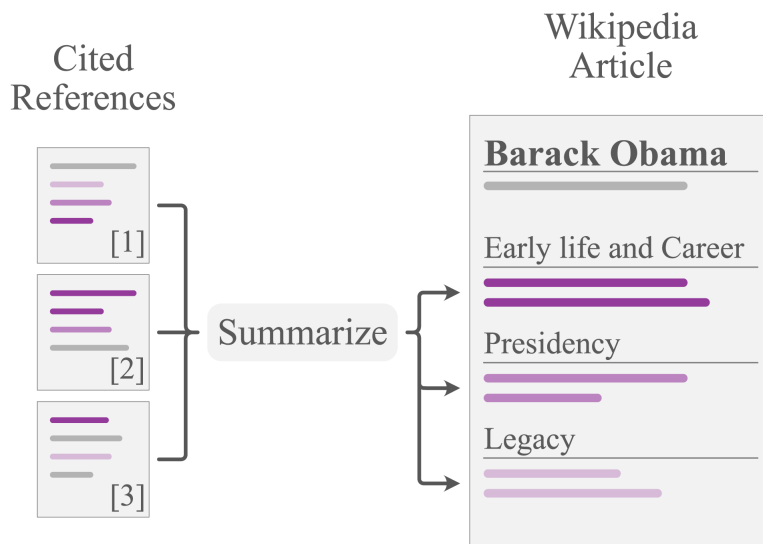


Figure 5.7: WikiAsp aspect-based summarisation model as extracted from [Hayashi et al., 2021]

In terms of visualisation, aspect-based opinion summaries can be represented using other different visualisations, namely bar charts [Liu et al., 2005, Abulaish et al., 2009], pie charts [Balage Filho et al., 2012, Bahrainian and Dengel, 2013, Kamal, 2015], radar plots [Soulis et al., 2013], and tree maps [Shamim et al., 2021]. The work in [Eirinaki et al., 2012] provides textual results for each sentiment polarity value and feature ratings of the queries input. Moreover, [Shamim et al., 2015] carried out an evaluation of opinion visualisations representing opinions in terms of descriptive statistics, with the top ones being bar charts, glowing bars, tree maps, line graphs, and pie charts. Similarly, [Kucher et al., 2018] shows that different visualisations, such as basic pie and bar charts, and extensive visual analytics systems, are used based on the use case and complexity of the datasets. Lastly, the authors in [Moussa et al., 2018] conclude that visualisation-based summarisation is very useful and it brings many advantages, such as visually appealing, easy to read and comprehend, concise representation of data, and ability to perform a quick analysis of data.

5.2.4 Advancing over Aspect-based Opinion Summarisation of Social Data

The opinion summaries of all of the studies presented in Sections 5.2.2 and 5.2.3 are based on the sentiment polarity opinion dimension only. The work presented in this chapter focuses on a niche research area that has been identified, that of **aspect-based multidimensional opinion summarisation**. The aspect-based multidimensional opinion summarisation approach presented involves the following steps: (i) identification of the main topics and sub-topics (referred as aspects) which are catered for within the Malta Government Budget (refer to Section 3.4.1.6) for each online post; (ii) identification of the opinion for each online post in terms of subjectivity, sentiment polarity, emotion, irony, and sarcasm; and (iii) summarisation of the opinion results. In this regard, we **advance** and **improve** upon existing studies by exploring and providing opinion summaries for **five** social opinion dimensions, namely, *subjectivity*, *sentiment polarity*, *emotion*, *irony*, and *sarcasm*.

From a conceptual perspective, the aspect-based summaries developed in our work are in line with the ones proposed by [Hu and Liu, 2004] and [Liu et al., 2005] for products sold online. The work in [Liu et al., 2005] shows that **opinion summary visualisations** should be **tailored** to the needs of the use case in question and should not be tied to a particular visualisation, e.g., bar chart. In fact, this approach was adopted in our work, specifically for the Malta Government Budget real-world use case. Here the aim is not to perform any comparison of products but to show the opinion summaries of the different budget aspects, that is, the main Government topics and sub-topics as aligned with the domains handled by the Ministries e.g., Taxation, Healthcare, Environment, etc.

In terms of **opinion summary visualisation**, pie charts were selected to display the descriptive statistics for each social opinion dimension where percentages for each dimension value are displayed upon mouse hover. This choice of **pie charts** is also supported by [Shamim et al., 2015], who mention that pie charts are common, simple, effective, and easy-to-use visualisations. It is important to note that each generated aspect-based opinion summary is composed of statistical pie charts for each of the five social opinion dimensions, the corresponding textual online posts, respective social opinion dimension

values for each online post, and the data source and year of each online post.

The proposed work focusing on aspect-based multidimensional opinion summarisation is cutting-edge when compared to existing literature. Therefore, it advances over existing state-of-the-art with respect to aspect-based opinion summarisation and provides the basis for further multidimensional opinion search and summarisation research.

5.3 Opinion Search and Summarisation System

This section discusses the Opinion Search and Summarisation system developed for the purposes of this research in order to determine whether multidimensional Opinion Mining at an aspect-based level provides benefits for fine-grained opinion search and summarisation.

5.3.1 Overview

A web-based Opinion Search and Summarisation system was developed for the purposes of this study in order to determine whether **multidimensional** Opinion Mining at an **aspect-based** level provides benefits for **fine-grained opinion search and summarisation**. The main **motivation** behind this study is to provide a nuanced voice to the general public on topics that are of national importance. In this respect, this system was evaluated for a **real-world** use case, namely the annual Malta Government Budget. The **goal** is to determine whether the aspect-based multidimensional approach provides added benefits for decision-making and/or policy-making purposes, within context of the Government.

This system offers the following functionality:

1. **Searching** for rich **fine-grained social opinions** of different aspects (at a topic or sub-topic level) within social data provided, either via faceted navigation or natural language; and
2. **Visualisation** of returned results in terms of an **aspect-based opinion summary**.

5.3.2 Dataset

The opinion Search and Summarisation system leverages the novel multidimensional and multilingual social opinion dataset, presented in Section 3.4. This Socio-Economic and Government dataset for Malta’s annual Government Budget comprises social data from the 2018, 2019, and 2020 budgets, contains 6,387 online posts collected from three Maltese online newswires and Twitter, and was annotated for five social opinion dimensions.

It is important to mention that this system is pre-loaded with the mentioned dataset which was annotated and consolidated manually as discussed in Sections 3.4.1.3 and 3.4.1.5. Therefore, in order for the system to work with any newly annotated data (manually, semi-automatically, and/or automatically) targeting the annual Government Budget, this would need to be primarily pre-loaded in the existing system prior to be used for opinion search and summarisation purposes.

5.3.3 Technologies

In terms of technologies, this system was developed using Microsoft Azure¹ cloud computing services, where the Azure Cognitive Search² service was used for the search functionality and the Azure SQL Database³ for the online posts storage. The faceted navigation was implemented to facilitate search for the end-user. This feature provides end-users with the ability to explore and find any information needed by filtering or navigating with the help of pre-determined facets [Tunkelang, 2009, Ganesan and Zhai, 2012]. Moreover, faceted search provides end-users with more effective information-seeking support than conventional search approaches [Tunkelang, 2009]. In fact, facets were added for all the main Government Budget topics and sub-topics, and social opinion dimensions.

A full-text search feature was also implemented for purposes of conducting a search in natural language. Figure 5.8 shows an opinion summary using the full-text search feature. The counts displayed next to each facet indicate the number of matches for each Government Budget topic and sub-topic, and each social opinion dimension.

¹<https://azure.microsoft.com/en-us/>

²<https://azure.microsoft.com/en-us/services/search/>

³<https://azure.microsoft.com/en-us/products/azure-sql/database/>



Figure 5.8: Opinion Summary Visualisation about the “Innovation” topic using the full-text search feature

5.3.4 Context behind Visualisations of Opinion Summaries

The aspect-based opinion summaries generated by the Opinion Search and Summarisation system are visualised through pie charts for each social opinion dimension. They provide a basic and easy-to-understand visualisation of the mentioned opinions posed by the general public on a particular topic and sub-topic for certain domains. These can be used by the Government of Malta for further social, economic, and/or political considerations. The example provided in Figure 5.9 shows an opinion summary about the “Transport” topic and “Bicycle” sub-topic using the faceted search feature.

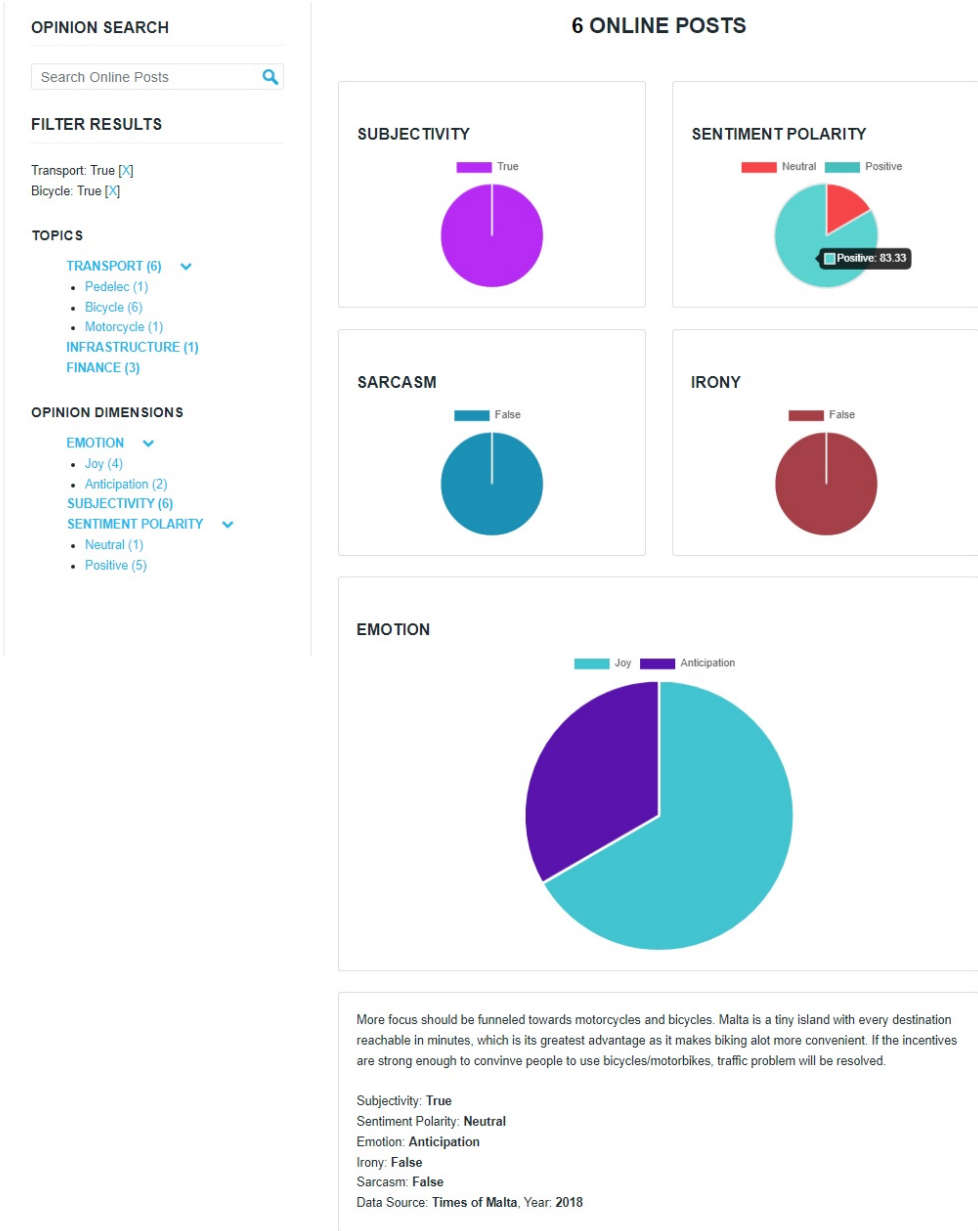


Figure 5.9: Opinion Summary Visualisation about the “Transport” topic and “Bicycle” sub-topic using the faceted search feature

The opinion summary visualisation is based on a total of 6 online posts. In summary, this shows that all the online posts focused on the mentioned topic and respective sub-topic consist of the following:

- **Subjectivity:** All of the online posts are *subjective* (100%);

- **Sentiment Polarity:** 83.33% of the online posts are *positive* and 16.67% are *neutral*;
- **Emotion:** 66.67% of the online posts convey a *joy* emotion and 33.33% convey an *anticipation* emotion;
- **Sarcasm:** None of the online posts are *sarcastic*; and
- **Irony:** None of the online posts are *ironic*.

5.4 Human Evaluation Setup

The main focus of the human evaluation carried out is to determine whether the Opinion Search and Summarisation system can be used for a **specific purpose**, in our case for **policy formulation, policy-making, decision-making, and decision-taking**, within context of the **Government**. The opinion summaries provide a nuanced-voice to the general public i.e., citizens and residents of Malta, on their opinions about a topic that is related to the annual Malta Government Budget. The reason why a human evaluation was chosen for this study as opposed to a metrics-based one is due to latter using automatic metrics e.g., ROUGE [Lin and Hovy, 2003] and BLEU [Papineni et al., 2002], to compare system-generated textual summaries with human-generated reference texts [Graham, 2015], which is not in line with the main objective of this work. Given that this work is focused on evaluating aspect-based opinion summaries for the purpose mentioned above, a human evaluation is the only way how their impact and quality can be measured, with automatic metrics not being suitable such a purpose [van der Lee et al., 2021].

This human evaluation consisted of an **extrinsic** (also called task-based) evaluation, **usability** evaluation, and **subject matter expert** evaluation. An extrinsic evaluation assesses a summary based on a task and measures whether it was helpful to the human performing this specific task [Lloret et al., 2018]. This differs from intrinsic evaluations where the quality of the summarisation output is measured on the said summary whereas the source text is not considered [Iskender et al., 2021]. Extrinsic evaluations are time-, resource-, and cost-intensive [Reiter and Belz, 2009, Gatt and Krahmer, 2018, Pecar, 2018], especially when developing a system which can be tested in a real-world scenario

[van der Lee et al., 2021]. However, [Reiter and Belz, 2009] argue that extrinsic evaluations are more **useful** than intrinsic ones, and if applied in a **real-world setting**, they are more useful than a controlled one. On the other hand, a usability evaluation assesses a system’s usability and user experience before its wide adoption in a real-world setting [Nielsen, 2012]. This consolidates our decision towards opting for an extrinsic evaluation and usability evaluation of the Opinion Search and Summarisation system, which has been evaluated in a real-world scenario, namely the Malta Government Budget.

An **extrinsic evaluation** (discussed in detail in Section 5.4.5) was carried out to determine whether a **fine-grained multidimensional opinion summary** provides any **added benefit** when compared to an opinion summary based on sentiment polarity only (one dimensional). Participants were asked to carry out a number of tasks using the Opinion Search and Summarisation system discussed in Section 5.3. A **usability evaluation** (discussed in detail in Section 5.4.6) was also carried out to evaluate the **generic usability** of the Opinion Search and Summarisation system. Lastly, a **subject matter expert** evaluation (discussed in detail in Section 5.4.7) was carried out to evaluate the applicability of multidimensional opinion summarisation within the Government domain and to determine whether there are any correlations with the results obtained from the other evaluations.

5.4.1 Methodology

The methodology behind the human evaluation consisted of the following:

- **Ethics Application:** Submitted to the university Research Ethics Committee (REC) for review due to the study involving human participants. Once the REC issued approval, the human evaluation could commence (See Appendix D for the DCU Research Ethics Committee Application).
- **Plain Language Statement:** Written information provided to the participants, outlining the nature of their involvement in the study and an invitation for their participation (See Appendix D for the Plain Language Statement).
- **Informed Consent Form:** Document requiring participants to indicate their con-

sent to participate in the study and provide their signature (See Appendix D for the Informed Consent Form).

- **Pre-Evaluation Questionnaire:** Asks each participant for background information (3 questions) and on their knowledge (6 questions) within the study’s research area i.e., Opinion Mining/Sentiment Analysis and the applied real-world use case i.e., the Malta Government Budget (See Appendix E for the Pre-Evaluation Questionnaire). The knowledge questionnaire was scored by assigning each answer a value from 0 to 2 and dividing the total by 12 to obtain a score between 0-100.

- **Opinion Search and Summarisation system - Extrinsic Evaluation and Usability Evaluation:** The following were carried out for these evaluations:
 - An overview of the system to be used for evaluation (See Appendix F);
 - Information about the three systems to be evaluated for the purposes of the study (See Appendix G);
 - An opinion search task for each of the three systems (See Appendix H for all the topic-based tasks carried out by the participants of each group as discussed in Section 5.4.4);
 - A questionnaire on the opinion summarisation results (obtained from the opinion search task) for each system evaluated (System 1, System 2, System 3) based on extrinsic quality measures where the 5-point Mean Opinion Score (MOS) [Streijl et al., 2016] is used for each question (See Appendix I for the Opinion Summarisation Results’ Questionnaire);
 - A generic system usability evaluation based on the System Usability Scale (SUS) [Brooke et al., 1996] (See Appendix J for the SUS Questionnaire); and
 - Other generic questions for any comments, feedback, or suggestions (See Appendix K for the other questions).

- **Opinion Search and Summarisation system - Subject Matter Expert Evaluation:** An interview process that was carried out with a Subject Matter Expert to evaluate the three systems (See Appendix L for the list of interview questions).

5.4.2 Participants' Sample

In total, 18 participants with different professional backgrounds, levels of education, and working sector were recruited to carry out the human evaluation, namely the extrinsic evaluation and usability evaluation. In terms of other requirements, all participants had to be over 16 years of age, have basic digital skills (e.g., familiar with using a personal computer, browser, etc.), able to understand the English and Maltese languages (since data used for evaluation contains data in both languages), and potentially have basic knowledge of the annual Malta Government Budget.

The selection process was undertaken to ensure that the sample is highly representative of the general population that might potentially use such an Opinion Search and Summarisation system for decision-making and/or policy-making purposes. Therefore, it was important that such a sample contained a diverse set of participants that are both ICT professionals and otherwise. Moreover, these requirements and criteria ensured that a purposeful sample of participants were identified and selected based on certain principles, such as knowledge in a domain of interest, availability, willingness to participate, and have an ability to communicate their experiences and opinions in an expressive and reflective manner [Palinkas et al., 2015]. In addition, the sample size of participants (18) satisfies the requirements of at least 12-14 participants to obtain reasonably reliable usability results [Tullis and Stetson, 2004].

5.4.3 Systems' Overview

In total, the following three systems have been evaluated for the purposes of this study:

1. **System 1** - Opinion Search and Summarisation based on *sentiment polarity* (**One Dimensional**);
2. **System 2** - Opinion Search and Summarisation based on *sentiment polarity* and *emotion* (**Two Dimensional**); and
3. **System 3** - Opinion Search and Summarisation based on *sentiment polarity*, *emotion*, and *sarcasm* or *irony* (**Three Dimensional**).

The Opinion Search tasks defined for Systems 1-3 (see Section 5.4.4 below for more information) are based on *subjective* online posts i.e., the ones that reflect an opinion by the end-user about the Malta Government Budget. For this reason, *objective* online posts have not been considered for this study.

5.4.4 System Evaluation Strategy

The 18 participants were split in 6 groups, with each consisting of 3 people. Each participant group evaluated three Opinion Search and Summarisation systems, one focusing on *sentiment polarity* (System 1), one on *sentiment polarity* and *emotion* (System 2), and one on *sentiment polarity*, *emotion*, and *sarcasm* or *irony* (System 3). All participants were asked to complete a set of tasks for each system. The following is an overview of the steps that each participant had to carry out for each system:

- Access the Opinion Search and Summarisation system;
- Filter the online posts by the “subjectivity” social opinion dimension;
- Click on a Government Budget topic and respective sub-topic;
- View the pie chart (opinion visualisation) of the social opinion dimension in question depending on the system being evaluated e.g., sentiment polarity, and examine the percentage of online posts broken down by the social opinion dimension values e.g., *positive*, *negative*, by hovering the mouse on the pie chart; and
- Filter the online post by one of the social opinion dimension values e.g., *positive* sentiment polarity in the left side bar of the system, where the search results will return with online posts having a *positive* sentiment polarity only.

In total three main Government Budget topics, namely **Transport**, **Environment**, and **Taxation**, were chosen for the Opinion Search tasks defined for evaluation by Systems 1-3. These topics were arbitrarily chosen based on their importance at a national level and the information offered in terms of diversity regarding social opinion dimensions i.e., having different statistics which cover all aspects of the different social opinion dimensions evaluated, namely sentiment polarity, emotion, irony, and sarcasm. For example, topic A

produces more search results with a positive sentiment polarity, topic B produces more search results with a negative sentiment polarity, etc.

The diversity in the topics chosen shows that the interpreted results are topic-agnostic and will show different aspects that are output from the opinion summary visualisations. It is important to note that two different groups repeated the tasks based on each topic, hence a fair balance is provided with the variety of topics chosen for defining the tasks to be evaluated. Moreover, the main focus was on evaluating the different systems and not the topic, therefore the approach taken disassociates the topic from the systems.

Table 5.1 presents all the system combinations (6 in total) carried out by each group. Each group evaluated the three systems (Systems 1-3) in varying orders, resulting in 6 permutations overall (See Appendix H for all the topic-based tasks carried out by the participants of each group). This system variation process was selected in order to minimise any bias from using one system before another. The tasks for Systems 1-3 were based on the same topic to ensure a same level playing field is provided for each system, since varying the topic between the systems would result in an unfair comparison due to the change in tasks, activities, and returned opinion summary results.

Group	System Order	Topic-based Tasks
1	1-2-3	Transport
2	1-3-2	Environment
3	2-1-3	Taxation
4	2-3-1	Transport
5	3-1-2	Environment
6	3-2-1	Taxation

Table 5.1: Groups of participants by system and topic

The human evaluation was initially piloted with 3 participants to test the robustness of the three systems. However, they were not used for the human evaluation. From this pilot, the participants provided feedback on some information to be included to the system overview to be used for evaluation, suggested some improvements to the tasks that each participant was asked to carry out for each system in cases where the defined steps were not clear, provided feedback on the system usability, and suggested some minor changes to the pre-evaluation questionnaire.

5.4.5 Extrinsic Evaluation

An extrinsic evaluation was carried out for each of the three systems, where all participants were asked to carry out a set of tasks for each system. Following completion of the said tasks, the extrinsic quality for each system was evaluated using the following four measures:

1. **Summary Usefulness** (defined as “content responsiveness” in [Conroy and Dang, 2008]): This measure provides a score for each aspect-based opinion summary. Therefore, it indicates whether the information content representing the aspect-based opinion summary provided the information needed on the particular topic or sub-topic that was searched upon.
2. **Source Usefulness** (defined as “relevance assessment” in [Mani et al., 2002]): This measure provides a score for each aspect-based opinion summary and the corresponding data sources of the online posts from which the summary was generated, in our case being online newswires (Times of Malta, MaltaToday, The Malta Independent) and/or social networking services (Twitter). Therefore, it indicates whether the aspect-based opinion summary accurately reflects the relevance or irrelevance of the corresponding data sources.
3. **Online Post Usefulness** (building on “post usefulness” as defined in [Iskender et al., 2021]): This measure provides a score for each aspect-based opinion summary and the corresponding online posts from which the summary was generated, in our case being newswires online comments and/or tweets (similar in nature to forum posts considered in [Iskender et al., 2021]). Therefore, it indicates whether the provision of online posts were deemed relevant to the aspect-based opinion summary.
4. **Summary Informativeness** (defined as “informativeness” in [Mani et al., 2002]): This measure provides a score to determine whether the extent of information provided for each aspect-based opinion summary, i.e., statistical pie charts and corresponding online posts in terms of actual text, respective social opinion dimension values, and data source, is enough and preserves the key information used to generate the summary.

These four measures are based on the three extrinsic quality measures introduced in [Iskender et al., 2021], namely “summary usefulness”, “post usefulness”, and “summary informativeness”. In our case, these three measures were not enough to extrinsically evaluate the aspect-based opinion summaries, since both the data source (e.g., Twitter, online newswires) and respective online post are provided as part of each opinion summary. Therefore, an **additional extrinsic quality measure** was introduced to be able to cater for the measurement of both the source and online post usefulness. The latter measure was also **adapted** from [Iskender et al., 2021] in order to be more representative of the online nature of the social data used in this study.

The evaluations for each measure was carried out using the 5-point MOS with the labels “Excellent” (5 points), “Good” (4 points), “Fair” (3 points), “Poor” (2 points), and “Bad” (1 point). This continuous scale was used for a subjective quality assessment where opinions about the performance of a system are useful. Even though the MOS is the “de-facto” metric used to quantify perceived media quality [Streijl et al., 2016], it has been applied in human extrinsic evaluation research for text summarisation by [Iskender et al., 2021] to assess human evaluation reliability. Therefore, for the purposes of our extrinsic evaluation, the MOS calculation for each measure above provides a numerical measure of the human-judged overall quality of the respective opinion summaries returned from the opinion search tasks (i.e., opinion search and summarisation experience) carried out by each participant for each of the three systems. The MOS for each of the four measures above is calculated by finding the average of the human scores obtained (i.e., points based on the selected label e.g., 5 points for “Excellent”) for each system evaluated.

5.4.6 Usability Evaluation

All participants were asked to fill in a system usability questionnaire to evaluate the generic usability of the Opinion Search and Summarisation system which is used for purposes of policy formulation, policy-making, decision-making, and decision-taking. This questionnaire is based on the SUS, a reliable and robust tool, used for measuring the usability of products and services, such as software, websites, and applications. This consists of a ten item questionnaire with each containing five response options, i.e., a 5-point scale ranging

from “strongly agree” to “strongly disagree”, and produces a score between 0-100. This scale provides a global view of usability subjective assessments [Brooke et al., 1996].

5.4.7 Subject Matter Expert Evaluation

In terms of selection, besides the requirements specified in Section 5.4.2, the interviewed Subject Matter Expert had to be a high-ranking Government official, such as Chief Information Officer, Chief Technology Officer, or policy-maker. The following is an overview of the interview process that was carried out with the Subject Matter Expert:

- Interview time duration was expected to be between 30-45 minutes.
- Background information and general information on the Opinion Search and Summarisation system was initially provided (based on information provided in Sections 5.3 and 5.4.3). This included a brief overview of the Malta Government Budget dataset (based on the information provided in Section 3.4) which was used to generate the opinion summaries.
- A live demonstration of the Opinion Search and Summarisation system was carried out. In this regard, the online posts were filtered by the “subjectivity” social opinion dimension (same as was done for the extrinsic evaluation), where the faceted search feature was initially used on the “Technology” topic, followed by the full-text search feature with the “Technology” keyword. The opinion summaries were filtered using any of the other social opinion dimensions offered i.e., sentiment polarity, emotion, irony, and sarcasm, to show the potential of this system.
- The interview questions listed in Appendix L were asked to the Subject Matter Expert after the live demonstration was completed.

5.5 Results and Analysis

This section discusses the results and analysis of the human evaluation presented in Section 5.4. These include information on the background and knowledge of the 18 participants (also referred to as evaluators below) in relation to the Opinion Mining research area and

Malta Government Budget use case, and results obtained from the extrinsic evaluation, usability evaluation, and subject matter expert evaluation of the Opinion Search and Summarisation system presented in Section 5.3.

5.5.1 Participants' Background

In terms of the participants' professional background, 7 work in the private sector, 7 work in the public sector, 3 work in the public service, whereas 1 is self-employed. With regard to their working sector, there was a good diversity between people working in ICT (7 in total) and others working in other sectors, namely Healthcare (4), Aviation (1), Construction and Manufacturing (1), E-commerce (1), Education (1), International Development (1), Retail and Beauty (1), and Transportation (1). Table 5.2 provides a list of the highest education of the participants based on the European Qualifications Framework (EQF)⁴. This shows that the evaluation was carried out by participants with a wide range of background, all of who can be potential users of such an Opinion Search and Summarisation system either as a consumer, or as a policy-maker and/or decision-maker.

EQF Level	Qualification	Number of Participants
8	Doctorate Degree	1
7	Master's Degree / Postgraduate Diploma / Postgraduate Certificate	8
6	Bachelor's Degree	4
5	Undergraduate Diploma / Undergraduate Certificate	3
4	Matriculation Certificate / Advanced Level / Intermediate Level	1
1	General Education School Leaving Certificate	1

Table 5.2: Highest education of the participants based on the EQF

5.5.2 Participants' Knowledge

All participants were tested for their degree of knowledge with respect to the research area of "Opinion Mining" or "Sentiment Analysis", and Malta's annual Government Budget at both a generic and in-depth level. The average score across all participants resulted in

⁴<https://europa.eu/europass/en/description-eight-eqf-levels>

54.63 (out of 100) which indicates that the level of knowledge in this research area and applied real-world use case domain is at a medium level. Out of 18 participants, 7 obtained a score of 66.67 or higher (highly knowledgeable), 5 obtained a score between 50.00 and 66.66 (medium knowledgeable), whereas 6 obtained a score less than 50.00 (not very knowledgeable). In this respect, the participants had limited knowledge in the research area of this study, however they were familiar with Malta's annual Government budget, and most of them follow the main updates of the annual Malta Government Budget.

5.5.3 Extrinsic Evaluation Outcome

All three systems discussed in Section 5.4.3 have been evaluated, where the extrinsic quality for each was measured in terms of **summary usefulness** (SUU), **source usefulness** (SOU), **online post usefulness** (OPU), and **summary informativeness** (SUI). The context of these measures was to determine whether the results of all three systems reflect the summary of opinions of the general public if they were to use the system for **policy formulation, policy-making, decision-making, and decision-taking** (original purpose of evaluation). Table 5.3 provides the MOS results for each measure, where they are interpreted as follows:

- 1.0 to 2.6 - not recommended;
- 2.6 to 3.1 - nearly all participants are dissatisfied;
- 3.1 to 3.6 - many participants are dissatisfied;
- 3.6 to 4.0 - some participants are dissatisfied;
- 4.0 to 4.3 - participants are satisfied; and
- 4.3 to 5.0 - participants are very satisfied.

The results for each measure in each sub-section below are discussed together with some insights that have been observed and interpreted.

System	SUU	SOU	OPU	SUI
1	3.94	4.61	4.56	4.33
2	4.22	4.50	4.61	4.28
3	4.72	4.72	4.72	4.72

Table 5.3: Mean Opinion Score (MOS) results for the summary usefulness (SUU), source usefulness (SOU), online post usefulness (OPU), and summary informativeness (SUI) extrinsic quality measures

5.5.3.1 Summary Usefulness

MOS results indicate that the participants were **very satisfied** with the aspect-based opinion summary based on sentiment polarity, emotion, and sarcasm or irony (System 3) in terms of content, and hence favour this output as opposed to the opinion summary provided by the other systems. Results show that some participants were not satisfied with an opinion summary based on sentiment polarity (System 1) only and found it less useful for the original purpose of this evaluation.

The evaluators favoured having more context (three dimensions) whilst evaluating the aspect-based opinion summary (MOS - 4.72). This indicates that for purposes such as decision-making and policy-making, more context-rich opinion summaries can be more effective towards achieving the relevant goals, since such information truly reflects the general public opinion in a more comprehensive manner.

5.5.3.2 Source Usefulness

MOS results indicate that the participants were **very satisfied** with the name of the data sources of the online posts for the aspect-based opinion summary provided. Even though the participants favoured System 3 (MOS - 4.72), the participants were very satisfied with this aspect in all three systems, which highlights the importance of data provenance for opinion summarisation.

Taking into consideration the nature of the data sources used in this evaluation, it is worth noting that the evaluators might experience bias towards or against specific sources. For instance, a specific newswire might lean towards certain political beliefs, therefore when it is used as one of the sources within an opinion summary, the evaluators might have an increased, or decreased feeling of trust in the opinions expressed within the summary based on their political affiliation. Therefore, whilst the source usefulness is

indeed confirmed as being quite high in this evaluation for all three systems, the resulting bias in the interpretation of the opinion summaries might be reflected in the purpose, such as decision-making and policy-making.

5.5.3.3 Online Post Usefulness

MOS results indicate that the participants were **very satisfied** with the provision of corresponding online posts for the aspect-based opinion summary provided. Similar to the source usefulness measure, even though the participants favoured System 3 (MOS - 4.72), the participants were very satisfied with this aspect in all three systems and found the online posts useful to better understand the respective opinion summary. Moreover, the additional metadata provided with each online post, namely the respective social opinion dimension values, was deemed as being very useful.

The minor change between the results obtained by the evaluators for each system might suggest that the provision of more fine-grained online post results might be useful. For instance, it might be useful to provide the evaluators with the specific textual extract/s from the respective online post for each expressed social opinion dimension. This would provide the evaluators with more explainable results for each expressed social opinion dimension, thereby providing them with more insight, and in turn facilitating certain processes such as decision-making and policy-making.

5.5.3.4 Summary Informativeness

MOS results indicate that the participants **favoured** the extent of information for the aspect-based opinion summary provided by System 3 i.e., based on sentiment polarity, emotion, and sarcasm or irony (MOS - 4.72), followed by System 1 i.e., based on sentiment polarity (MOS - 4.33), and System 2 i.e., based on sentiment polarity and emotion (MOS - 4.28), when used for purposes such as policy-making and/or decision-making. It is interesting to note that results for Systems 1 and 2 were very closely rated, thereby indicating that the participants felt that the addition of the emotion opinion dimension did not make much of a difference in terms of informativeness.

These results indicate that the evaluators were able to easily process the provided

information within the opinion summaries. This suggests that there was no information overload in the opinion summaries as presented to the end-user, and a balance was successfully struck between providing the required information, whilst still not being overwhelming.

5.5.4 Usability Evaluation Outcome

The **generic usability** of the Opinion Search and Summarisation system as a whole for purposes of **policy formulation, policy-making, decision-making, and decision-taking**, was evaluated using the SUS. The following is an overview of the SUS scores obtained from the 18 participants:

- *Average SUS score* resulted in 86.53 which resolves to the highest rating (Grade A - score above 80.3) indicating that all participants loved the system and shall recommend it to their peers;
- *Highest SUS score* was that of 100 (Grade A - score above 80.3); and
- *Lowest SUS score* was 72.5 (Grade B - score between 68 to 80.3) which is well above the standard average SUS score of 68 [Brooke, 2013].

5.5.5 Subject Matter Expert Evaluation Outcome

The interviewed Subject Matter Expert is a Chief Information Officer for one of the Government of Malta ministries, where the interview session lasted for around 1 hour. In terms of professional background, the Subject Matter Expert works in the public service within the ICT working sector and possesses a Master's Degree/Postgraduate Diploma/Postgraduate Certificate (EQF Level 7). In terms of knowledge, the Subject Matter Expert has high knowledge in the research area of "Opinion Mining" or "Sentiment Analysis" and has worked on a Sentiment Analysis application at a proof-of-concept level for the Government. Moreover, the Subject Matter Expert is highly knowledgeable in Malta's annual Government Budget.

The following is an overview of the feedback and insights gathered from the interview with the Subject Matter Expert:

- System 2 (i.e., based on sentiment polarity and emotion) best reflects the summary of opinions of the general public for purposes of policy formulation, policy-making, decision-making, and decision-taking. In terms of irony and sarcasm, it is good to know about them but they would not give much value for policy drafting, policy writing, and certain decision-making processes.
- As a potential end-user of the system the full-text search is preferred for obtaining the relevant aspect-based opinion summaries. The reason being that through this functionality one can use any keyword needed (even if it's not provided within the list of facets) and it is simpler to find what one is searching for.
- The following is a list of potential end-users who would benefit from using the opinion search and summarisation system:
 - **Quality of Service Directorate**⁵: The system can be integrated with a central system where citizens and residents of Malta can submit their complaints. This can help them better identify certain emotions, such as anger and fear, for certain public service aspects and therefore improve the public service quality.
 - **Permanent Secretaries**⁶: This system would be a “brilliant” way for each Permanent Secretary within the public service to identify what people are saying about a particular government measure/initiative e.g., education service launched, that is related to their Ministry.
 - **Public Consultations Online**⁷: Currently the Government offers a website where citizens and/or residents of Malta can write their feedback on public consultations on certain topics of national interest e.g., Climate Action Bill. However, such a system can complement what the citizens and/or residents of Malta write on social sources (e.g., Facebook, Twitter) about certain topics of national interest that are related to the public consultations.
 - **Principal Permanent Secretary**⁸: A system of this nature can be used by

⁵<https://publicservice.gov.mt/en/people/Pages/QualityandStandardsDirectorate/QualityStandardsDirectorate.aspx>

⁶<https://publicservice.gov.mt/en/Pages/Leadership/PermanentSecretaries.aspx>

⁷www.konsultazzjoni.gov.mt

⁸<https://publicservice.gov.mt/en/Pages/Leadership/LeadershipStructure.aspx>

the Head of the Public Service to obtain the citizen and resident opinions about public services which can therefore be used to continuously improve the public service.

- **Servizz.gov**⁹: The agency responsible for the online guide to the Government services can use this system to track citizen and/or resident sentiment on any online Government service and relevant Government keywords e.g., “public service”, “servizz pubbliku” (‘public service’ in Maltese), “servizz.gov”, as expressed on social sources (e.g., Facebook, Twitter, online newspapers commenting sections).
 - **Office of the Prime Minister**¹⁰: Team working on budget measures in preparation for each annual Government Budget.
- In terms of system functionality it was suggested to consider including the user profile for each online post as part of each corresponding online post for the aspect-based opinion summary provided. This would help in providing user context, particularly for any strongly biased opinions that are part of a generated opinion summary about a particular Government Budget topic or sub-topic.

5.5.6 Feedback and Suggestions

The participants carrying out the extrinsic evaluation and usability evaluation were asked if they have any comments on any of the systems and whether they have any other feedback or suggestions. The following is an overview of the main comments, feedback, and suggestions:

- Determining which system to use depends on the end-user. For generic use by common users such as the general public, System 2 will suffice since it provides enough informative responses and can be enough for decision-making purposes. Similarly, this correlates with the feedback provided by the Subject Matter Expert with regard to usage of this system for policy drafting, policy writing, and decision-making processes. However, System 3 might be better suited if the end-user is the state or

⁹<https://www.servizz.gov.mt/>

¹⁰<https://opm.gov.mt/>

a private/public company, since more fine-grained information shall result in better **informed decisions**.

- Such an Opinion Search and Summarisation system is very useful for personnel working on **policies, economic measures, and social measures**. Moreover, the system is good for government entities to gather general public opinion on the annual budget and is very useful to analyse the budget data. Lastly, given that a good condensed overview of the general public opinion is provided, such a system can be an asset to various bodies and entities of a different nature. This also correlates with the feedback provided by the Subject Matter Expert.
- In terms of **usability**, many participants commented that the system is user friendly.
- The **sarcasm** and **irony** social opinion dimensions might not be much needed for policy-making, except for the instances where sarcasm changes the emotion value (e.g., an online post which is marked as ‘joy’ is actually negative due to the sarcasm dimension). This also correlates with the feedback provided by the Subject Matter Expert.
- Analysing data for **sarcasm** and/or **irony** could reduce the efficiency of the opinion summaries, especially since a more detailed breakdown of the **sentiment polarity** can be provided via **emotions**.
- Clarification on the actual meaning of **emotion** values. For example, the ‘anticipation’ emotion since this could be subjectively interpreted and it can mean ‘excitement’ for one person or it can mean ‘anxiety’ for another. Hence, it is important that the emotions are defined for people who might use the system and might not be very familiar with the emotional vocabulary.
- It would have been beneficial if the tasks provided resulted in opinion summaries that have **mixed emotions**, e.g., angry or disgust, and trust or joy.
- Providing too much information can be more time consuming to the end-user. However it depends on the scenario applied and on the time one can spend to analyse results.

- The **data sources** are more beneficial when the opinion summary provides the **emotion**, and **sarcasm** or **irony** social opinion dimensions (System 2 and System 3). Moreover, showing the data source can be both of value and otherwise, since it can also lead to biased decisions by the person who is using this system such that the data can be interpreted in a certain way just because of one’s personal opinion.
- Providing some indication of how the social opinion dimensions were extracted would enable the system to be more “explainable”, i.e., providing reference to the specific part of the online posts (spans) corresponding to the relevant opinion dimensions identified.
- The following is some **functionality** that can be added: filter results by data source; filter results by clicking on the pie chart sections; filter by year and plotting a time series to portray how the social opinion dimensions develop; provide a numeric value (0-10) for each dimension value to show its strength e.g., disgust emotion (8); make the system accessible for the visually impaired; add a word cloud for each social opinion dimension to make it easier to the end-user to filter through the online posts; and make the system customisable for specific end-users (e.g., government employees, general public) or target groups (e.g., youth, elderly).
- The following are some **potential improvements** that can be added: pie chart sections should be displayed clockwise sorted by the highest percentage first; pie chart colours are consistent e.g., positive always green and negative always red; and topics and sub-topics ordered in alphabetical order.
- A good **use case** that the system can be used for is to gauge general public opinion on the **COVID-19** regulation announcements imposing restrictions over time, or on the COVID-19 vaccine booster.

5.6 Impact

The Government of Malta can use **social opinion summarisation** results about certain national aspects (e.g., Environment domain, Social Security benefits) for **policy formu-**

lation, policy-making, decision-making, and decision-taking. These results provide a nuanced voice to citizens and residents who use social media platforms to make their opinions known and/or provide feedback about anything, in this case Malta’s annual Government Budget. User-generated content has always been invaluable for certain needs, such as improving an entity’s service/perception and tracking citizen opinion.

In order to better demonstrate the applicability of our contribution towards opinion search and summarisation, we here provide a discussion on a real-world use case, and provide comments on the theoretical outcome of the application of the Opinion Search and Summarisation system. This provides a more concrete view on the impact of this contribution within the Government domain. We here take the example of Malta’s annual Government Budget of 2022, where the Minister for Finance announced that grants to purchase electric vehicles were increased over the previous grant available. This was part of a policy drive to increase the use of electric and plug-in hybrid vehicles in a bid to reduce vehicle pollution and reach carbon neutrality by 2050¹¹.

In our use case, the Government of Malta leverages the Opinion Search and Summarisation system presented in Section 5.3 for any decision-making and/or policy-making processes similar to the budget measures mentioned above. The Government of Malta uses this system to analyse the feedback provided by the general public on the topic of “vehicle pollution” from the Socio-Economic and Government dataset comprising social data from the Malta Government Budgets of 2018, 2019, and 2020. The overall feedback by the general public towards this topic is *negative* in terms of sentiment polarity and several emotions conveyed were of a negative nature i.e., *disgust*, *anger*, and *sadness* (see opinion summaries visualised in Figure 5.10). In retrospect, the Government of Malta could have used such results and data-based insights from citizen and resident feedback analysis to pro-actively make certain decisions –similar to the measures announced in the Government Budget of 2022– in earlier budgets and/or concretely take certain data-driven decisions on matters of national importance in a more timely manner. In terms of national impact, these opinion summarisation results and insights can help increase the trust between the citizens and residents of Malta and the Government given that certain deci-

¹¹<https://www.independent.com.mt/articles/2021-10-11/local-news/Budget-2022-Free-public-transport-to-all-Maltese-residents-as-from-October-2022-6736237463>

sions are taken based on public opinion. Moreover, it can enable better communication and interaction between Government policy-makers and decision-makers and citizens and residents of Malta.

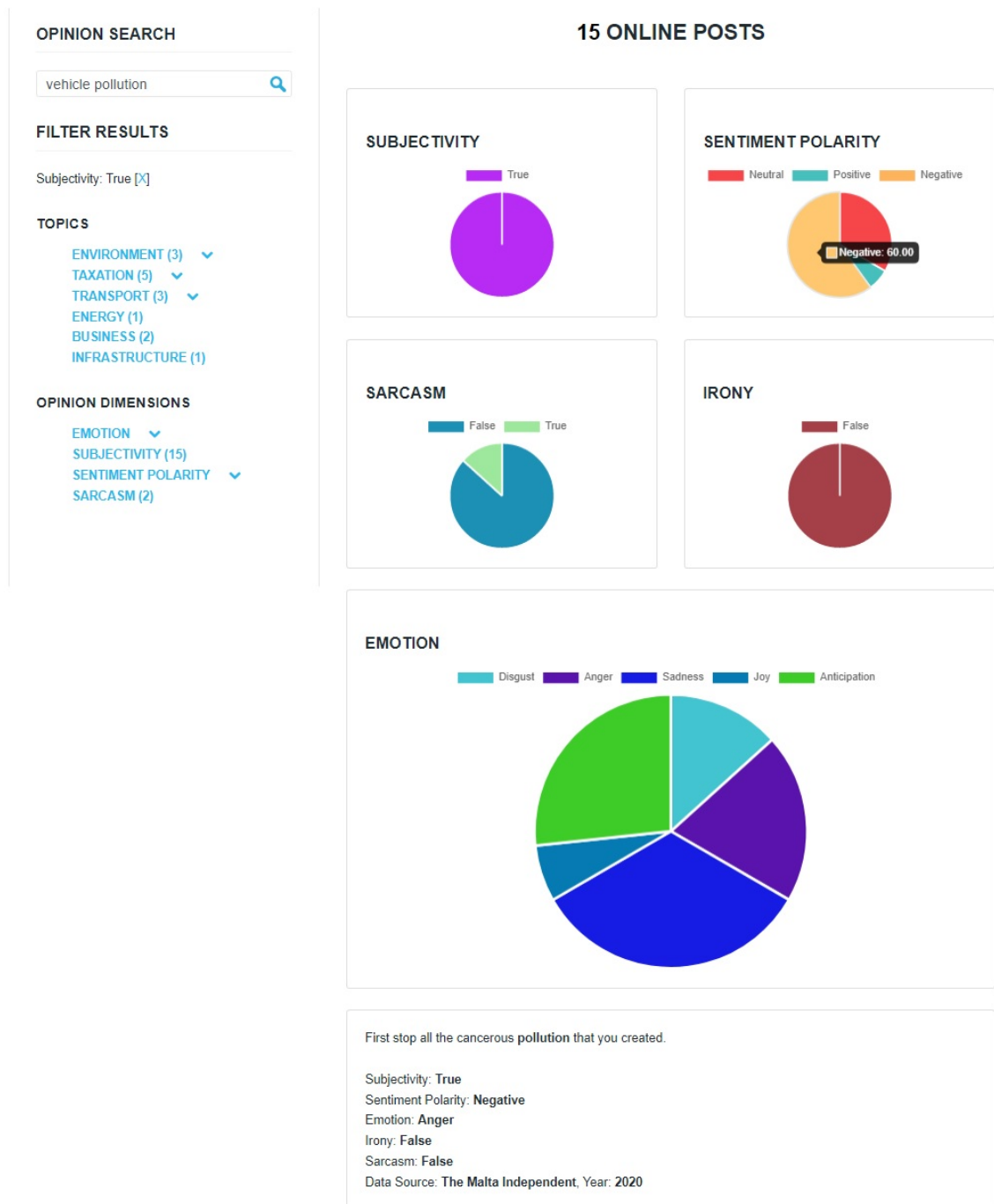


Figure 5.10: Opinion Summary Visualisation - real-world use case

5.7 Conclusion

This chapter discusses multidimensional opinion search and summarisation for a real-world use case in the Government domain, namely the annual Malta Government Budget. This use case was chosen since Government Budgets focus on a different application area than the ones traditionally applied to and in itself it caters for multiple domains, such as Transport, Environment, and Taxation. All of this shows that Social Opinion Mining can be applied in multiple domains. In this respect, a human evaluation was carried out to determine whether fine-grained opinion search and summarisation at an aspect-based and multidimensional level can be used for a specific purpose, in our case for policy formulation, policy-making, decision-making, and decision-taking, within context of the Government domain. This human evaluation consisted of an extrinsic evaluation, usability evaluation, and a subject matter expert evaluation. Results obtained from the human evaluation carried out using the Opinion Search and Summarisation system indicated that a more fine-grained aspect-based opinion summary based on the combined dimensions of *subjectivity*, *sentiment polarity*, *emotion*, and *sarcasm* or *irony* is more informative and more useful than an opinion summary based on sentiment polarity only. This highlights our contribution over current state-of-the-art opinion summarisation research which focuses on sentiment polarity only. In addition, such opinion summaries provide a nuanced voice to the general public i.e., citizens and residents of Malta, on their opinions about a topic that is related to the annual Malta Government Budget. This chapter answers **Research Questions 4 and 5**. The next chapter presents the conclusions and future directions of this thesis.

Chapter 6

Conclusion

6.1 Conclusions

Social media popularity and importance is on the increase due to people using it for various types of social interaction across multiple channels. This thesis focuses on the evolving research area of Social Opinion Mining, tasked with the identification of multiple opinion dimensions, such as *subjectivity*, *sentiment polarity*, *emotion*, *affect*, *irony*, and *sarcasm*, from user-generated content represented across multiple social media platforms and in various media formats, like textual (e.g., online posts, newswires' comments), visual (e.g., images, videos), and audio. Mining of people's social opinions from social sources, such as social media platforms and newswires commenting sections is a valuable business asset that can be utilised in many ways, ranging from marketing strategies to product/service improvement [Cambria et al., 2013], and in multiple domains, such as Finance, Politics, Sports, Marketing, Advertising, and Sales [Ravi and Ravi, 2015].

It is important to mention that there is a lack of research focusing on multidimensional Social Opinion Mining which is a niche area within Opinion Mining research. The majority of the literature focuses on subjectivity detection and sentiment analysis with emotion analysis increasing in popularity, whereas studies that are of a multidimensional nature handle at most two social opinion dimensions, such as subjectivity and sentiment polarity, or sentiment polarity and emotion. In fact, there is a deficit of quality datasets in this research area, especially datasets that are multilingual, multidimensional, and mul-

tilingual multidimensional. Moreover, research on sarcasm detection and irony detection is still evolving and hence is not in a mature state. As a result of all of these points, there is a lack of research that supports multiple languages or targets multilingual data on multiple Opinion Mining tasks, such as opinion classification, aspect-based Opinion Mining, opinion search, and opinion summarisation that take into consideration multidimensional social opinions. Regarding the applicability of Social Opinion Mining, most of the research targets common areas and domains, such as Finance, Marketing, Advertising, Sales, and Politics. Other research in non-traditional application areas and domains, such as Government, can help in determining whether Social Opinion Mining can influence multiple application areas making it one of the most active research areas within the field of NLP.

Through Social Opinion Mining, natural language can be understood in terms of the different opinion dimensions as expressed by humans. Unfortunately, current practice in Social Opinion Mining mostly limits to the use of a single opinion dimension (i.e., sentiment polarity) which results in a limited and simplified view of social opinions. In fact, handling multiple social opinion dimensions contribute to the formulation of a more context-rich and fine-grained representation of the real-life complexity of human opinions which is very important within this research area. In addition, they provide a nuanced voice to the general public who use social media platforms and online newswires commenting sections to make their opinions known and/or provide feedback about anything. Multidimensional Social Opinion Mining has the potential for researchers to better model and understand human beliefs and their behaviour. This contributes to the advancement of several real-world scenarios, such as adaptive customer online service [Yadollahi et al., 2017], tracking of overall customer satisfaction for a product or service [Zhao et al., 2019], and detection of changes in customer opinion towards a brand, product, or service [Geetha et al., 2017]. Moreover, it impacts entities utilising Opinion Mining results towards effective decision-making at a strategic level. In turn, this will contribute towards the evolution of Artificial Intelligence. This is especially the case for Governments who should be more citizen-centric and therefore should better understand public opinions on national topics so that they can address their needs [AI4PublicPolicy, 2022]. The use of Artificial Intelligence, such as Social Opinion Mining, shall benefit the public sector in terms of policy delivery,

decision-making processes, public services, and citizen-government interaction, which will in turn increase citizens' trust in public sector activities [Commission et al., 2022]. All of this provided the motivation to pursue further research in this area with the aim of showing the potential of Social Opinion Mining so that people's social opinions are taken in consideration in real-world applications and are utilised to their full potential in certain domains.

In this thesis the research was aimed towards answering the following research question, as proposed in Section 1.3:

How do multiple Social Opinion Mining dimensions enable a better understanding of the complexity of an aspect-level social opinion?

Through this research question we explored whether an aspect-level social opinion targeting multiple dimensions namely *subjectivity*, *sentiment polarity*, *emotion*, *sarcasm*, and *irony*, produces more accurate opinions with regard to reflecting the real-life complexity than those based on a single dimension (e.g., sentiment polarity only). This has been evaluated for a real-world use case, the Malta Government Budget. This human evaluation determines whether the multidimensional opinion summarisation results provide added-value to potential end-users, such as policy-makers and decision-takers, thereby providing a nuanced voice to the general public on their social opinions on topics of a national importance. In this regard, results obtained indicate that a more fine-grained aspect-based opinion summary based on the combined dimensions of *subjectivity*, *sentiment polarity*, *emotion*, and *irony* or *sarcasm* is more informative and more useful than an opinion summary based on sentiment polarity alone. This is especially relevant if the end-user is the state or a private or public company, where more fine-grained information shall result in better informed decisions. However, based on the human evaluation carried out, the level of granularity needed also depends on the real-world scenario that aspect-based opinion summarisation is applied to. This includes generic use by common users e.g., the general public, where an aspect-based opinion summary based on sentiment polarity and emotion will suffice since it provides enough informative responses and can be enough for decision-making purposes.

To summarise the relevant contributions of this thesis, we revisit each research question presented in Section 1.3.

Research Question 1: *What are the existing Opinion Mining approaches that make use of social data and what are the different dimensions of Social Opinion Mining used?*

Several Opinion Mining approaches have explored the use of user-generated content in multiple application areas, such as Politics, Marketing, Advertising, Sales, Technology, and Finance. Through this research question we identified, analysed, and evaluated existing Opinion Mining approaches that make use of social data in terms of social media platforms, techniques, social datasets, language, modality, tools and technologies, NLP tasks including aspect-based Social Opinion Mining, and other aspects identified within literature. The different Social Opinion Mining dimensions, namely subjectivity, sentiment polarity, emotion, affect, irony, sarcasm, and mood were identified from the analysed Opinion Mining solutions, together with their main application areas. The in-depth analysis also provided us with a clear direction on the social opinion dimensions that are under-represented in literature, whether existing work caters for multiple dimensions such as subjectivity, sentiment polarity, and emotion, and if current approaches accurately portray the real-life complexity of human opinions. The latest research developments, advancements, and current literature gaps within the Social Opinion Mining research area have also been identified. In fact, the main results outline that there is a need for aspect-based Social Opinion Mining studies that cater for multiple social opinion dimensions, research focused on code-switched languages (e.g., Maltese-English) and less-resourced languages (e.g., Maltese), collection of new social datasets from more than one data source for a real-world application area that target bilingual/multilingual data, which data is annotated for multiple opinion dimensions, and the need to develop Social Opinion Mining classification techniques based on machine learning and deep learning approaches. The direction of this thesis and the contributions within are based on these identified research gaps.

Research Question 2: *How does Social Opinion Mining influence multiple applica-*

tion areas?

To answer this question, we collected and annotated three social datasets in the following domains: i) Politics targeting the Brexit Referendum from Twitter; ii) Finance targeting Stock Trading from Twitter, StockTwits, and News Statements and Headlines; and iii) Socio-Economic and Government targeting Government Budgets from the Times of Malta, MaltaToday, The Malta Independent (all newswires), and Twitter. In terms of Social Opinion Mining dimensions, the Political dataset focuses on sentiment in terms of polarity and strength, the Financial dataset focuses on sentiment in terms of score, and the Socio-Economic and Government dataset focuses on subjectivity, sentiment polarity, emotion, irony, and sarcasm. These datasets were generated to determine how Social Opinion Mining can influence multiple application areas. Whilst all three datasets cater for one common social opinion dimension i.e., sentiment polarity, since it is the most explored dimension within the Opinion Mining research area, the Socio-Economic and Government dataset also focuses on additional opinion dimensions. This was done in order to cater for the lack of multidimensional datasets for Social Opinion Mining purposes. Given that all the three datasets cater for one common social opinion dimension in terms of sentiment polarity, the datasets can still be evaluated in terms of the influence of Social Opinion Mining in multiple application areas. However, the generation of the above-mentioned three social datasets enabled us to understand the importance of multiple social opinion dimensions that target multilingual data, which is currently largely unexplored in the Social Opinion Mining research area. For this reason, in this thesis we proceeded with the research direction of catering for multiple social opinion dimensions that target multilingual data (see research questions below). Moreover, the Socio-Economic and Government dataset in itself caters for multiple domains, such as Transport, Environment, and Taxation. In this regard, Social Opinion Mining can be applied in several domains and hence influence multiple application areas.

Research Question 3: *How can multidimensional and multilingual social opinion data be used for the development of classification models?*

To answer this question, we collected and annotated a social dataset in the Socio-Economic and Government domains, specifically Malta’s annual Government Budgets of 2018, 2019, and 2020. This is made up of multilingual data gathered from newswires and social networking services and is annotated for multiple social opinion dimensions, namely *subjectivity*, *sentiment polarity*, *emotion*, *irony*, and *sarcasm*, as well as for *negation*, *topic*, and *language*. This is the first annotated social dataset with multiple social opinion dimensions at both a European and national level, and in the context of the Maltese Socio-Economic and Government domains, making it highly beneficial and a first contribution of its kind for Malta. Furthermore, we explored how this multidimensional and multilingual social dataset can be used for classification purposes. This was validated through the development of different classification models for each social opinion dimension (mentioned above) and for different language levels, namely monolingual (English), code-switched (Maltese, Maltese-English), and multilingual (English, Maltese, Maltese-English). In this respect, several baseline classification models using traditional machine learning techniques and state-of-the-art deep neural network models were developed. In terms of contributions, these baseline models are a first at targeting different opinion dimensions, namely subjectivity, sentiment polarity, emotion, irony, and sarcasm, at a code-switched level using user-generated content in Maltese and Maltese-English, and at a multilingual level using user-generated content in English, Maltese, and Maltese-English.

Research Question 4: *How are fine-grained opinion search and summarisation affected through the use of a multidimensional Social Opinion Mining approach at an aspect-based level?*

The proposed multidimensional opinion search and summarisation approach presented in this thesis considerably improves upon existing state-of-the-art. While the latter focus on sentiment polarity only, the approach proposed in this thesis takes into consideration five social opinion dimensions, namely subjectivity, sentiment polarity, emotion, irony, and sarcasm. These all contributed towards the formulation of a more context-rich and fine-

grained representation of the real-life complexity of human opinions. The aspect-based multidimensional opinion results were summarised based on the use case being considered, i.e. the Malta Government Budget. A human evaluation comprising an extrinsic evaluation and usability evaluation was carried out to determine whether the fine-grained opinion search and summarisation at an aspect-based and multidimensional level provides added-value to the end-users, such as policy-makers and decision-takers. For this purpose, an Opinion Search and Summarisation system was developed. The extrinsic evaluation was based on four quality measures, improving upon existing state-of-the-art with the introduction of an additional measure and an adaptation of an existing one to be representative of the social data used within our study. The results obtained indicate that a more fine-grained aspect-based opinion summary based on the combined dimensions of subjectivity, sentiment polarity, emotion, and sarcasm or irony is 7.8% more informative and 15.6% more useful than an opinion summary based on sentiment polarity alone. This further highlights our contribution over current state-of-the-art opinion summarisation research which focuses only on sentiment polarity.

Research Question 5: *How can multidimensional Social Opinion Mining be applied to the Government domain?*

The proposed multidimensional Social Opinion Mining model was validated using a real-world use case within the Government domain, specifically the Malta Government Budget. This use case was chosen since Government Budgets focus on a different application area than the ones traditionally applied to, such as Finance, Marketing, Advertising, Sales, and Politics, and in itself it caters for multiple domains that are of national interest and importance, such as Transport, Environment, and Taxation. To answer this research question, we evaluated whether social opinion summarisation results about certain national domains and aspects such as the ones mentioned above can be used by the Government of Malta for policy formulation, policy-making, decision-making, and decision-taking. This human evaluation was carried out using an Opinion Search and Summarisation system that leveraged the novel multidimensional and multilingual social opinion dataset gener-

ated in the Socio-Economic and Government domains. The evaluation outcome indicates the suitability of the system towards the application in the Government domain. The same dataset was used to build baseline monolingual (English), code-switched (Maltese, Maltese-English), and multilingual (English, Maltese, Maltese-English) classification models for subjectivity detection, sentiment analysis, emotion analysis, sarcasm detection, and irony detection. Given that such models are capable of understanding English and Maltese data, both being Malta's official languages, they can be used by Governments, in this case the Government of Malta, to help in any decision-making and/or policy-making processes. Multidimensional Social Opinion Mining can be used by Governments to mine the opinions of citizens and residents of their respective countries to ensure that their voices on certain topics and issues of national importance are heard with the aim of positively impacting society at large. Moreover, this research area can be used by Governments in several other real-world scenarios, such as the public service in terms of online services offered and improvement of public service quality, Government measures such as budget and financial ones, and online public consultations focused on certain topics of national interest e.g., Climate Action Bill.

6.2 Future Directions

This section discusses a number of future directions of this thesis which shall build upon the presented research contributions.

- **Explore other dimensions that make up a social opinion:** The research within this thesis mainly focuses on five social opinion dimensions, namely *subjectivity*, *sentiment polarity*, *emotion*, *irony*, and *sarcasm*. However, as identified in the literature analysis (refer to Section 2.4), other dimensions, such as affect and mood, are also related to the other opinion dimensions. For example, research on multidimensional sentiment analysis shows that the human mood is very rich in social media, where a piece of text may contain multiple moods, such as calm and agreement [Huang et al., 2015]. Further research in the identified areas both separately e.g., mood analysis, and in conjunction with other opinion dimensions is needed to better understand

the factors that make up an opinion.

- **Correlation between social opinion dimensions:** From the analysis of this literature it is evident that there is a lack of research on any possible correlations between the social opinion dimensions e.g., between sentiment polarity and emotion. In this regard, [Troussas et al., 2016] specify that “emotions and polarities are mutually influenced by each other, conditioning opinion intensities and emotional strengths”. Similarly, [Koto and Adriani, 2015b] focused on finding the correlation between emotion and sentiment in social media. More in-depth research on such correlations shall help in better understanding the formulation of human opinions which are very complex.
- **Express social opinion dimensions in terms of more granularity:** Certain social opinion dimensions, such as sentiment polarity and emotion, can be expressed in a more fine-grained manner to better express the opinion dimension in question. For the dataset of Multidimensional and Multilingual Social Opinions for Malta’s Annual Government Budget, the sentiment polarity of each online post was annotated for a categorical value based on 3-levels, namely negative, neutral, positive, whereas the emotion of each online post was annotated for a categorical value based on 8-levels in accordance with Plutchik’s eight primary emotions namely joy, sadness, fear, anger, anticipation, surprise, disgust, trust [Plutchik, 1980]. In this case, the sentiment polarity can be expressed in more levels of granularity as a categorical value based on 5-levels, namely very negative, negative, neutral, positive, very positive. Moreover, for the emotion opinion dimension, new emotions can be generated if two primary emotions are combined based on Plutchik’s Wheel of Emotions. For example, if an online post is annotated for two emotions based on Plutchik’s eight primary emotions mentioned above, such as “joy” and “trust”, the combination of these primary emotions generates a new emotion “love”. These new emotions can be generated for the mentioned dataset in cases where raters who annotated this dataset did not all agree on the same emotion value.
- **Social Opinion Mining for low-resourced and/or code-switched languages:**

Further research focused on low-resourced languages, such as Maltese and Irish, and code-switched languages, such as Maltese-English and Irish-English, shall enable the development and advancement of certain language tools e.g., named entity recogniser, and resources e.g., social datasets, needed for the respective low-resourced and code-switched languages for Social Opinion Mining, in particular for subjectivity detection, sentiment analysis, emotion analysis, sarcasm detection, and irony detection (as separate research areas or otherwise). Whilst we initiated research in this regard, deep learning techniques and fine-tuned Transformer-based models are still under-explored. Fine-tuning Transformer-based models is extremely reliable for adapting to new NLP tasks e.g., from sentiment analysis to emotion analysis, and/or knowledge transfer from one language to another, such as cross-lingual adaptation where a labelled dataset is available in one language (e.g., English) which is then applied to another language such as a low-resourced language (e.g., Maltese) [Ruder, 2017].

- **Explainable aspect-based opinion summarisation:** Providing information on how the social opinion dimensions were extracted would enable an aspect-based opinion summarisation system to be more “explainable”, that is, by providing reference to the specific part of the online posts (spans) corresponding to the relevant opinion dimensions identified. Further evaluation of the impact of providing explainable results for each expressed social opinion dimension as part of each aspect-based opinion summary shall determine whether this facilitates certain processes such as decision-making and policy-making.
- **Dealing with imbalanced data:** The dataset of Multidimensional and Multilingual Social Opinions for Malta’s Annual Government Budget used for building baseline classification models for each Social Opinion Mining dimension is imbalanced, especially for the emotion, irony, and sarcasm classes. Most machine learning and deep learning classification algorithms are not capable of handling imbalanced classes and therefore end up being biased towards majority classes. This results in the trained classification models not being able to predict the correct outcome of unseen data, since there is not enough data for the model to learn from. Several

re-sampling techniques [Cateni et al., 2014, More, 2016], such as over-sampling (e.g., SMOTE [Chawla et al., 2002]), under-sampling (e.g., Tomek Link Removal [Tomek, 1976]), k-fold cross-validation, and a combination of all, can be used for handling an imbalanced dataset. Therefore, further experimentation should be carried out using re-sampling techniques in order to obtain better results for subjectivity detection, sentiment analysis, emotion analysis, sarcasm detection, and irony detection.

- **Aspect-based Opinion Mining for multiple social opinion dimensions:** Opinion Mining at an aspect-based level is usually applied for sentiment polarity only. Further aspect-based research is encouraged on other social opinion dimensions other than sentiment polarity, such as emotion, irony, and sarcasm, which are still unexplored. In terms of techniques, the use of Deep Learning approaches to carry out aspect-based Opinion Mining is still at an early stage and more research in this direction should be explored.
- **Real-world policy-making use case based on fine-grained opinion summaries:** Carry out a long-term evaluation in a real-world policy-making use case to determine the effectiveness of fine-grained opinion summaries at a national and/or international level. At a national level, this evaluation can be carried out by the Government through a Government entity e.g., Ministry, for a fixed, longer duration (e.g., months), to gather the general public opinion about matters that are of national importance. Such an evaluation shall determine whether the fine-grained opinion summaries generated are effective and hence should be taken in consideration for policy formulation and policy-making purposes, therefore, providing a nuanced-voice to the general public i.e., citizens and residents of the country in question.
- **Applicability of multidimensional Social Opinion Mining in multiple domains:** This thesis explored the applicability of multidimensional Social Opinion Mining in the Government domain, where the proposed multidimensional model catering for subjectivity, sentiment polarity, emotion, irony, and sarcasm, was validated using a real-world use case, specifically the Malta Government Budget. Further research on how multidimensional Social Opinion Mining affects other domains,

such as Politics e.g., Elections, and Finance e.g., Stock Trading, would provide interesting insights on which social opinion dimensions have a similar affect and impact across multiple domains (if any) and whether certain dimensions are more applicable to certain domains than others e.g., irony might be applicable to the Finance domain but less important to the Government domain. This research shall also help in better understanding people’s opinions especially when they are directed towards certain aspects within the respective domain and/or common aspects across multiple domains.

6.3 Final Remarks

This thesis focuses on multidimensional Social Opinion Mining which is a niche area within Opinion Mining research. This research can benefit multiple Opinion Mining tasks, such as opinion classification, aspect-based Opinion Mining, opinion search, and opinion summarisation. The multidimensional Social Opinion Mining model presented caters for **five** different social opinion dimensions, specifically *subjectivity*, *sentiment polarity*, *emotion*, *irony*, and *sarcasm*. This has been validated using a real-world use case within the Government domain, specifically the Malta Government Budget. In this regard, an Opinion Search and Summarisation system was developed to determine how a multidimensional Social Opinion Mining approach at an aspect-based level affects fine-grained opinion search and summarisation. Results obtained from the human evaluation carried out using the developed system indicated that a more fine-grained aspect-based opinion summary based on the combined dimensions of subjectivity, sentiment polarity, emotion, and sarcasm or irony is more informative and more useful than an opinion summary based on sentiment polarity only. This highlights our contribution over current state-of-the-art opinion summarisation research which tends to focus only on sentiment polarity .

Bibliography

- [Mer, a] “irony.” merriam-webster.com dictionary. <https://www.merriam-webster.com/dictionary/irony>. Date Accessed: 2022-02-27.
- [Mer, b] “mood.” merriam-webster.com dictionary. <https://www.merriam-webster.com/dictionary/mood>. Date Accessed: 2022-02-27.
- [Mer, c] “sarcasm.” merriam-webster.com dictionary. <https://www.merriam-webster.com/dictionary/sarcasm>. Date Accessed: 2022-02-27.
- [Abdul-Mageed et al., 2014] Abdul-Mageed, M., Diab, M., and Kübler, S. (2014). Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language*, 28(1):20–37.
- [Abdullah and Hadzikadic, 2017] Abdullah, M. and Hadzikadic, M. (2017). Sentiment analysis of twitter data: Emotions revealed regarding donald trump during the 2015-16 primary debates. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 760–764. IEEE.
- [Abdullah and Zolkepli, 2017] Abdullah, N. S. D. and Zolkepli, I. A. (2017). Sentiment analysis of online crowd input towards brand provocation in facebook, twitter, and instagram. In *Proceedings of the International Conference on Big Data and Internet of Thing*, pages 67–74. ACM.
- [Abulaish et al., 2009] Abulaish, M., Doja, M. N., Ahmad, T., et al. (2009). Feature and opinion mining for customer review summarization. In *International Conference on Pattern Recognition and Machine Intelligence*, pages 219–224. Springer.

- [Acheampong et al., 2021] Acheampong, F. A., Nunoo-Mensah, H., and Chen, W. (2021). Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, pages 1–41.
- [Agarap, 2018] Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.
- [Agarwal et al., 2018] Agarwal, A., Singh, R., and Toshniwal, D. (2018). Geospatial sentiment analysis using twitter data for uk-eu referendum. *Journal of Information and Optimization Sciences*, 39(1):303–317.
- [Agarwal et al., 2011] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. (2011). Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*, pages 30–38. Association for Computational Linguistics.
- [Agrawal et al., 2014] Agrawal, D., Budak, C., El Abbadi, A., Georgiou, T., and Yan, X. (2014). Big data in online social networks: User interaction analysis to model user behavior in social networks. In *DNIS*, pages 1–16. Springer.
- [Agrawal et al., 1994] Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499.
- [Ahuja and Dubey, 2017] Ahuja, S. and Dubey, G. (2017). Clustering and sentiment analysis on twitter data. In *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*, pages 1–5. IEEE.
- [AI4PublicPolicy, 2022] AI4PublicPolicy (2022). Reportbrain’s nlp for opinion mining: Going beyond sentiment analysis. <https://ai4publicpolicy.eu/blog-reportbrain-s-nlp-for-opinion-mining-going-beyond-sentiment-analysis/>. AI4PublicPolicy Project.
- [Aisopos et al., 2012] Aisopos, F., Papadakis, G., Tserpes, K., and Varvarigou, T. (2012). Content vs. context for sentiment analysis: a comparative analysis over microblogs. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, pages 187–196. ACM.

- [Akhtar et al., 2019] Akhtar, M. S., Chauhan, D. S., Ghosal, D., Poria, S., Ekbal, A., and Bhattacharyya, P. (2019). Multi-task learning for multi-modal emotion recognition and sentiment analysis. *arXiv preprint arXiv:1905.05812*.
- [Akhtar et al., 2020] Akhtar, M. S., Ekbal, A., and Cambria, E. (2020). How intense are you? predicting intensities of emotions and sentiments using stacked ensemble [application notes]. *IEEE Computational Intelligence Magazine*, 15(1):64–75.
- [Akhtar et al., 2017] Akhtar, M. S., Kumar, A., Ghosal, D., Ekbal, A., and Bhattacharyya, P. (2017). A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 540–546.
- [Alhumoud and Al Wazrah, 2021] Alhumoud, S. O. and Al Wazrah, A. A. (2021). Arabic sentiment analysis using recurrent neural networks: a review. *Artificial Intelligence Review*, pages 1–42.
- [Ali et al., 2020] Ali, S. M., Noorian, Z., Bagheri, E., Ding, C., and Al-Obeidat, F. (2020). Topic and sentiment aware microblog summarization for twitter. *Journal of Intelligent Information Systems*, 54(1):129–156.
- [Allen et al., 1955] Allen, K., Berry, M. M., Luehrs Jr, F. U., and Perry, J. W. (1955). Machine literature searching viii. operational criteria for designing information retrieval systems. *American Documentation (pre-1986)*, 6(2):93.
- [Altman, 1992] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.
- [Andrešić et al., 2017] Andrešić, D., Šaloun, P., and Anagnostopoulos, I. (2017). Efficient big data analysis on a single machine using apache spark and self-organizing map libraries. In *2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pages 1–5. IEEE.
- [Androutsopoulou et al., 2019] Androutsopoulou, A., Karacapilidis, N., Loukis, E., and Charalabidis, Y. (2019). Transforming the communication between citizens and government through ai-guided chatbots. *Government Information Quarterly*, 36(2):358–367.

- [Anjaria and Guddeti, 2014] Anjaria, M. and Guddeti, R. M. R. (2014). A novel sentiment analysis of social networks using supervised learning. *Social Network Analysis and Mining*, 4(1):181.
- [Aoudi and Malik, 2018] Aoudi, S. and Malik, A. (2018). Lexicon based sentiment comparison of iphone and android tweets during the iran-iraq earthquake. In *2018 Fifth HCT Information Technology Trends (ITT)*, pages 232–238. IEEE.
- [Appidi et al., 2020] Appidi, A. R., Srirangam, V. K., Suhas, D., and Shrivastava, M. (2020). Creation of corpus and analysis in code-mixed kannada-english twitter data for emotion prediction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6703–6709.
- [Aramanda et al., 2021] Aramanda, A., Abdul, S. M., and Vedala, R. (2021). Refining user ratings using user emotions for recommender systems. In *The 23rd International Conference on Information Integration and Web Intelligence*, pages 3–10.
- [Arslan et al., 2017] Arslan, Y., Birturk, A., Djumabaev, B., and Küçük, D. (2017). Real-time lexicon-based sentiment analysis experiments on twitter with a mild (more information, less data) approach. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 1892–1897. IEEE.
- [Attard et al., 2015] Attard, J., Orlandi, F., Scerri, S., and Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly*, 32(4):399–418.
- [Averchenkov et al., 2015] Averchenkov, V., Budylyskii, D., Podvesovskii, A., Averchenkov, A., Rytov, M., and Yakimov, A. (2015). Hierarchical deep learning: A promising technique for opinion monitoring and sentiment analysis in russian-language social networks. In *Creativity in Intelligent, Technologies and Data Science: First Conference, CIT&DS 2015, Volgograd, Russia, September 15–17, 2015, Proceedings*, pages 583–592. Springer.
- [Babanejad et al., 2020] Babanejad, N., Davoudi, H., An, A., and Papagelis, M. (2020). Affective and contextual embedding for sarcasm detection. In *Proceedings of the 28th In-*

- ternational Conference on Computational Linguistics*, pages 225–243, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- [Baccianella et al., 2010] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204.
- [Baccouche et al., 2018] Baccouche, A., Garcia-Zapirain, B., and Elmaghraby, A. (2018). Annotation technique for health-related tweets sentiment analysis. In *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 382–387. IEEE.
- [Bacchi et al., 2016] Bacchi, C., Uricchio, T., Bertini, M., and Del Bimbo, A. (2016). A multimodal feature learning approach for sentiment analysis of social network multimedia. *Multimedia Tools and Applications*, 75(5):2507–2525.
- [Baert et al., 2020] Baert, G., Gahbiche, S., Gadek, G., and Pauchet, A. (2020). Arabizi language models for sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 592–603.
- [Bahrainian and Dengel, 2013] Bahrainian, S.-A. and Dengel, A. (2013). Sentiment analysis and summarization of twitter data. In *2013 IEEE 16th International Conference on Computational Science and Engineering*, pages 227–234. IEEE.
- [Bakliwal et al., 2013] Bakliwal, A., Foster, J., van der Puil, J., O’Brien, R., Tounsi, L., and Hughes, M. (2013). Sentiment analysis of political tweets: Towards an accurate classifier. Association for Computational Linguistics.
- [Balage Filho et al., 2012] Balage Filho, P., Brun, C., and Rondeau, G. (2012). A graphical user interface for feature-based opinion mining. In *Proceedings of the Demonstration Session at the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5–8.
- [Balahur et al., 2009] Balahur, A., Kabadjov, M., Steinberger, J., Steinberger, R., and Montoyo, A. (2009). Summarizing opinions in blog threads. In *Proceedings of the 23rd*

Pacific Asia Conference on Language, Information and Computation, Volume 2, pages 606–613.

- [Balahur et al., 2012] Balahur, A., Kabadjov, M., Steinberger, J., Steinberger, R., and Montoyo, A. (2012). Challenges and solutions in the opinion summarization of user-generated content. *Journal of Intelligent Information Systems*, 39(2):375–398.
- [Balazs and Velásquez, 2016] Balazs, J. A. and Velásquez, J. D. (2016). Opinion mining and information fusion: A survey. *Information Fusion*, 27:95–110.
- [Bandhakavi et al., 2016] Bandhakavi, A., Wiratunga, N., Massie, S., and Deepak, P. (2016). Emotion-corpus guided lexicons for sentiment analysis on twitter. In *Research and Development in Intelligent Systems XXXIII: Incorporating Applications and Innovations in Intelligent Systems XXIV 33*, pages 71–85. Springer.
- [Bansal and Srivastava, 2018] Bansal, B. and Srivastava, S. (2018). On predicting elections with hybrid topic based sentiment analysis of tweets. *Procedia Computer Science*, 135:346–353.
- [Bansal et al., 2020] Bansal, S., Garimella, V., Suhane, A., Patro, J., and Mukherjee, A. (2020). Code-switching patterns can be an effective route to improve performance of downstream nlp applications: A case study of humour, sarcasm and hate speech detection. *arXiv preprint arXiv:2005.02295*.
- [Barapatre et al., 2016] Barapatre, D., Meena, M. J., and Ibrahim, S. S. (2016). Twitter data classification using side information. In *Proceedings of the 3rd International Symposium on Big Data and Cloud Computing Challenges (ISBCC-16’)*, pages 363–368. Springer.
- [Barberá, 2015] Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political analysis*, 23(1):76–91.
- [Baum and Petrie, 1966] Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563.

- [Becker et al., 2013] Becker, L., Erhart, G., Skiba, D., and Matula, V. (2013). Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 333–340.
- [Benamara et al., 2018] Benamara, F., Inkpen, D., and Taboada, M. (2018). Introduction to the special issue on language in social media: exploiting discourse and other contextual information. *Computational Linguistics*, 44(4):663–681.
- [Bird et al., 2009] Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* ” O’Reilly Media, Inc.”.
- [Blenn et al., 2012] Blenn, N., Charalampidou, K., and Doerr, C. (2012). Context-sensitive sentiment classification of short colloquial text. *NETWORKING 2012*, pages 97–108.
- [Bollen et al., 2011] Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8.
- [Borth et al., 2013] Borth, D., Ji, R., Chen, T., Breuel, T., and Chang, S.-F. (2013). Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232. ACM.
- [Bos and Frasincar, 2022] Bos, T. and Frasincar, F. (2022). Automatically building financial sentiment lexicons while accounting for negation. *Cognitive Computation*, 14(1):442–460.
- [Bosco et al., 2013] Bosco, C., Patti, V., and Bolioli, A. (2013). Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems*, 28(2):55–63.
- [Bottou, 2010] Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT’2010*, pages 177–186. Springer.

- [Bouazizi and Ohtsuki, 2017] Bouazizi, M. and Ohtsuki, T. (2017). A pattern-based approach for multi-class sentiment analysis in twitter. *IEEE Access*, 5:20617–20639.
- [Bouazizi and Ohtsuki, 2018] Bouazizi, M. and Ohtsuki, T. (2018). Multi-class sentiment analysis in twitter: What if classification is not the answer. *IEEE Access*, 6:64486–64502.
- [Bravo-Marquez et al., 2013] Bravo-Marquez, F., Mendoza, M., and Poblete, B. (2013). Combining strengths, emotions and polarities for boosting twitter sentiment analysis. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 2. ACM.
- [Bravo-Marquez et al., 2014] Bravo-Marquez, F., Mendoza, M., and Poblete, B. (2014). Meta-level sentiment models for big social data analysis. *Knowledge-Based Systems*, 69:86–99.
- [Breiman, 1996] Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.
- [Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- [Brereton et al., 2007] Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., and Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of systems and software*, 80(4):571–583.
- [Brodersen et al., 2010] Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE.
- [Brooke, 2013] Brooke, J. (2013). Sus: a retrospective. *Journal of usability studies*, 8(2):29–40.
- [Brooke et al., 1996] Brooke, J. et al. (1996). Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.
- [Buechel et al., 2020] Buechel, S., Rücker, S., and Hahn, U. (2020). Learning and evaluating emotion lexicons for 91 languages. *arXiv preprint arXiv:2005.05672*.

- [Bukhari et al., 2016] Bukhari, A., Qamar, U., and Ghazia, U. (2016). Urwf: user reputation based weightage framework for twitter micropost classification. *Information Systems and e-Business Management*, 15(3):623–659.
- [Buscaldi and Hernandez-Farias, 2015] Buscaldi, D. and Hernandez-Farias, I. (2015). Sentiment analysis on microblogs for natural disasters management: a study on the 2014 genoa floodings. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1185–1188. ACM.
- [Cai and Xia, 2015] Cai, G. and Xia, B. (2015). Convolutional neural networks for multimedia sentiment analysis. In *Natural Language Processing and Chinese Computing*, pages 159–167. Springer.
- [Cambria, 2016] Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107.
- [Cambria et al., 2020] Cambria, E., Li, Y., Xing, F. Z., Poria, S., and Kwok, K. (2020). Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 105–114.
- [Cambria et al., 2013] Cambria, E., Schuller, B., Xia, Y., and Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, (2):15–21.
- [Canales Zaragoza, 2018] Canales Zaragoza, L. (2018). Tackling the challenge of emotion annotation in text.
- [Carvalho and Plastino, 2021] Carvalho, J. and Plastino, A. (2021). On the evaluation and combination of state-of-the-art features in twitter sentiment analysis. *Artificial Intelligence Review*, 54(3):1887–1936.
- [Caschera et al., 2016] Caschera, M. C., Ferri, F., and Grifoni, P. (2016). Sentiment analysis from textual to multimodal features in digital environments. In *Proceedings of the 8th International Conference on Management of Digital EcoSystems*, pages 137–144. ACM.

- [Castellucci et al., 2015a] Castellucci, G., Croce, D., and Basili, R. (2015a). Acquiring a large scale polarity lexicon through unsupervised distributional methods. In *International Conference on Applications of Natural Language to Information Systems*, pages 73–86. Springer.
- [Castellucci et al., 2015b] Castellucci, G., Croce, D., and Basili, R. (2015b). Bootstrapping large scale polarity lexicons through advanced distributional methods. In *Congress of the Italian Association for Artificial Intelligence*, pages 329–342. Springer.
- [Castellucci et al., 2016] Castellucci, G., Croce, D., and Basili, R. (2016). A language independent method for generating large scale polarity lexicons. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 38–45.
- [Cateni et al., 2014] Cateni, S., Colla, V., and Vannucci, M. (2014). A method for re-sampling imbalanced datasets in binary classification tasks for real-world problems. *Neurocomputing*, 135:32–41.
- [Charalabidis et al., 2015] Charalabidis, Y., Maragoudakis, M., and Loukis, E. (2015). Opinion mining and sentiment analysis in policy formulation initiatives: The eu-community approach. In *International Conference on Electronic Participation*, pages 147–160. Springer.
- [Chawla et al., 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.
- [Chen et al., 2018a] Chen, C.-C., Huang, H.-H., and Chen, H.-H. (2018a). NtUSD-fin: a market sentiment dictionary for financial social media data applications. In *Proceedings of the 1st Financial Narrative Processing Workshop (FNP 2018)*.
- [Chen et al., 2020] Chen, C.-C., Huang, H.-H., and Chen, H.-H. (2020). Issues and perspectives from 10,000 annotated financial social media data. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6106–6110.

- [Chen et al., 2018b] Chen, C.-C., Huang, H.-H., Shiue, Y.-T., and Chen, H.-H. (2018b). Numeral understanding in financial tweets for fine-grained crowd-based forecasting. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 136–143. IEEE.
- [Chen et al., 2015] Chen, F., Gao, Y., Cao, D., and Ji, R. (2015). Multimodal hypergraph learning for microblog sentiment prediction. In *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, pages 1–6. IEEE.
- [Chen et al., 2017] Chen, M., Zhang, L.-L., Yu, X., and Liu, Y. (2017). Weighted co-training for cross-domain image sentiment classification. *Journal of Computer Science and Technology*, 32(4):714–725.
- [Chinchor, 1992] Chinchor, N. (1992). Muc-4 evaluation metrics. In *Proceedings of the 4th Conference on Message Understanding, MUC4 '92*, page 22–29, USA. Association for Computational Linguistics.
- [Chou et al., 2017] Chou, P.-H., Tsai, R. T.-H., and Hsu, J. Y.-j. (2017). Context-aware sentiment propagation using lda topic modeling on chinese conceptnet. *Soft Computing*, 21(11):2911–2921.
- [Cignarella et al., 2020] Cignarella, A. T., Basile, V., Sanguinetti, M., Bosco, C., Rosso, P., and Benamara, F. (2020). Multilingual irony detection with dependency syntax and neural models. *arXiv preprint arXiv:2011.05706*.
- [Clark and Niblett, 1989] Clark, P. and Niblett, T. (1989). The cn2 induction algorithm. *Machine learning*, 3(4):261–283.
- [Cohen, 1995] Cohen, W. W. (1995). Fast effective rule induction. In *Machine Learning Proceedings 1995*, pages 115–123. Elsevier.
- [Commission et al., 2022] Commission, E., Centre, J. R., Manzoni, M., Medaglia, R., Tangi, L., Van Noordt, C., Vaccari, L., and Gattwinkel, D. (2022). *AI Watch, road to the adoption of Artificial Intelligence by the public sector : a handbook for policy-makers, public administrations and relevant stakeholders*.

- [Condori and Pardo, 2017] Condori, R. E. L. and Pardo, T. A. S. (2017). Opinion summarization methods: Comparing and extending extractive and abstractive approaches. *Expert Systems with Applications*, 78:124–134.
- [Conroy and Dang, 2008] Conroy, J. and Dang, H. T. (2008). Mind the gap: Dangers of divorcing evaluations of summary content from linguistic quality. In *Proceedings of the 22nd international conference on computational linguistics (Coling 2008)*, pages 145–152.
- [Cook, 1977] Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18.
- [Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.
- [Cortis, 2013] Cortis, K. (2013). ACE: A concept extraction approach using linked open data. In *Making Sense of Microposts (#MSM2013) Concept Extraction Challenge*, pages 31–35.
- [Cortis and Davis, 2019] Cortis, K. and Davis, B. (2019). A social opinion gold standard for the malta government budget 2018. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 364–369.
- [Cortis and Davis, 2021a] Cortis, K. and Davis, B. (2021a). A dataset of multidimensional and multilingual social opinions for malta’s annual government budget. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 971–981.
- [Cortis and Davis, 2021b] Cortis, K. and Davis, B. (2021b). Over a decade of social opinion mining: a systematic review. *Artificial intelligence review*, 54(7):4873–4965.
- [Cortis and Davis, 2022] Cortis, K. and Davis, B. (2022). Baseline english-maltese classification models for subjectivity detection, sentiment analysis, emotion analysis, sarcasm detection and irony detection. In *Proceedings of the LREC 2022 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages (SIGUL 2022)*, pages 161–168. European Language Resources Association (ELRA).

- [Cortis and Davis, 2023] Cortis, K. and Davis, B. (2023). Social opinion search and summarisation: A multidimensional approach. (under review).
- [Cortis et al., 2017] Cortis, K., Freitas, A., Daudert, T., Huerlimann, M., Zarrouk, M., Handschuh, S., and Davis, B. (2017). Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535.
- [Cortis et al., 2021] Cortis, K., Verma, K., and Davis, B. (2021). Fine-tuning neural language models for multidimensional opinion mining of english-maltese social data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 309–314.
- [Cotfas et al., 2015] Cotfas, L.-A., Delcea, C., Roxin, I., and Paun, R. (2015). *Twitter Ontology-Driven Sentiment Analysis*, pages 131–139. Springer International Publishing, Cham.
- [Crammer et al., 2006] Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y. (2006). Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7(Mar):551–585.
- [Cresci et al., 2019] Cresci, S., Lillo, F., Regoli, D., Tardelli, S., and Tesconi, M. (2019). Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on twitter. *ACM Transactions on the Web (TWEB)*, 13(2):1–27.
- [Croce et al., 2020] Croce, D., Castellucci, G., and Basili, R. (2020). GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119, Online. Association for Computational Linguistics.
- [Cui et al., 2013] Cui, A., Zhang, H., Liu, Y., Zhang, M., and Ma, S. (2013). Lexicon-based sentiment analysis on topical chinese microblog messages. In *Semantic Web and Web Science*, pages 333–344. Springer.

- [Cui et al., 2011] Cui, A., Zhang, M., Liu, Y., and Ma, S. (2011). Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis. *Information retrieval technology*, pages 238–249.
- [Cvijikj and Michahelles, 2011] Cvijikj, I. P. and Michahelles, F. (2011). Understanding social media marketing: a case study on topics, categories and sentiment on a facebook brand page. In *Proceedings of the 15th international academic mindtrek conference: Envisioning future media environments*, pages 175–182. ACM.
- [Da Silva et al., 2014] Da Silva, N. F., Hruschka, E. R., and Hruschka, E. R. (2014). Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, 66:170–179.
- [Dambhare and Karale, 2017] Dambhare, S. S. and Karale, S. (2017). Smart map for smart city. In *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, pages 622–626. IEEE.
- [Dashtipour et al., 2016] Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y., Gelbukh, A., and Zhou, Q. (2016). Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8(4):757–771.
- [de Kauter et al., 2015] de Kauter, M. V., Breesch, D., and Hoste, V. (2015). Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with Applications*, 42:4999–5010.
- [Del Bosque and Garza, 2014] Del Bosque, L. P. and Garza, S. E. (2014). Aggressive text detection for cyberbullying. In *Mexican International Conference on Artificial Intelligence*, pages 221–232. Springer.
- [del Gobbo et al., 2021] del Gobbo, E., Fontanella, S., Sarra, A., and Fontanella, L. (2021). Emerging topics in brexit debate on twitter around the deadlines. *Social Indicators Research*, 156(2):669–688.
- [Delcea et al., 2014] Delcea, C., Cotfas, L.-A., and Paun, R. (2014). Understanding online social networks’ users—a twitter approach. In *International Conference on Computational Collective Intelligence*, pages 145–153. Springer.

- [Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.
- [Demszky et al., 2020] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.
- [Devillers et al., 2005] Devillers, L., Vidrascu, L., and Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422.
- [Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- [Diakopoulos and Shamma, 2010] Diakopoulos, N. A. and Shamma, D. A. (2010). Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1195–1198. ACM.
- [Do et al., 2019] Do, H. H., Prasad, P., Maag, A., and Alsadoon, A. (2019). Deep learning for aspect-based sentiment analysis: a comparative review. *Expert Systems with Applications*, 118:272–299.
- [dos Santos et al., 2018] dos Santos, A., Júnior, J. D. B., and de Arruda Camargo, H. (2018). Annotation of a corpus of tweets for sentiment analysis. In *International Conference on Computational Processing of the Portuguese Language*, pages 294–302. Springer.
- [Dragoni, 2018] Dragoni, M. (2018). Computational advertising in social networks: an opinion mining-based approach. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pages 1798–1804. ACM.
- [Dritsas et al., 2018] Dritsas, E., Livieris, I. E., Giotopoulos, K., and Theodorakopoulos, L. (2018). An apache spark implementation for graph-based hashtag sentiment classifi-

- cation on twitter. In *Proceedings of the 22nd Pan-Hellenic Conference on Informatics*, pages 255–260. ACM.
- [Drucker et al., 1997] Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., and Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161.
- [Duan et al., 2009] Duan, L., Tsang, I. W., Xu, D., and Chua, T.-S. (2009). Domain adaptation from multiple sources via auxiliary classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 289–296. ACM.
- [Dyba et al., 2007] Dyba, T., Dingsoyr, T., and Hanssen, G. K. (2007). Applying systematic reviews to diverse study types: An experience report. In *Empirical Software Engineering and Measurement, 2007. ESEM 2007. First International Symposium on*, pages 225–234. IEEE.
- [D’Asaro et al., 2017] D’Asaro, F. A., Di Gangi, M. A., Perticone, V., and Tabacchi, M. E. (2017). Computational intelligence and citizen communication in the smart city. *Informatik-Spektrum*, 40(1):25–34.
- [D’Avanzo and Pilato, 2015] D’Avanzo, E. and Pilato, G. (2015). Mining social network users opinions’ to aid buyers’ shopping decisions. *Computers in Human Behavior*, 51:1284–1294.
- [Eirinaki et al., 2012] Eirinaki, M., Pisal, S., and Singh, J. (2012). Feature-based opinion mining and ranking. *Journal of Computer and System Sciences*, 78(4):1175–1184.
- [Eke et al., 2020] Eke, C. I., Norman, A. A., Shuib, L., and Nweke, H. F. (2020). Sarcasm identification in textual data: systematic review, research challenges and open directions. *Artificial Intelligence Review*, 53(6):4215–4258.
- [Ekman, 1992] Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

- [Er et al., 2016] Er, M. J., Liu, F., Wang, N., Zhang, Y., and Pratama, M. (2016). User-level twitter sentiment analysis with a hybrid approach. In *International Symposium on Neural Networks*, pages 426–433. Springer.
- [Erdmann et al., 2014] Erdmann, M., Ikeda, K., Ishizaki, H., Hattori, G., and Takishima, Y. (2014). Feature based sentiment analysis of tweets in multiple languages. In *International Conference on Web Information Systems Engineering*, pages 109–124. Springer.
- [Farzindar and Inkpen, 2015] Farzindar, A. and Inkpen, D. (2015). Natural language processing for social media. *Synthesis Lectures on Human Language Technologies*, 8(2):1–166.
- [Fei and Li, 2020] Fei, H. and Li, P. (2020). Cross-lingual unsupervised sentiment classification with multi-view transfer learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5759–5771.
- [Feng et al., 2015] Feng, S., Song, K., Wang, D., and Yu, G. (2015). A word-emoticon mutual reinforcement ranking model for building sentiment lexicon from massive collection of microblogs. *World Wide Web*, 18(4):949.
- [Fersini et al., 2015] Fersini, E., Pozzi, F. A., and Messina, E. (2015). Detecting irony and sarcasm in microblogs: The role of expressive signals and ensemble classifiers. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pages 1–8. IEEE.
- [Fisher, 1925] Fisher, R. A. (1925). Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 22, pages 700–725. Cambridge University Press.
- [Flaes et al., 2016] Flaes, J. B., Rudinac, S., and Worring, M. (2016). What multimedia sentiment analysis says about city liveability. In *European Conference on Information Retrieval*, pages 824–829. Springer.
- [Fleiss and Cohen, 1973] Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

- [Freund et al., 1999] Freund, Y., Schapire, R., and Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612.
- [Frey and Dueck, 2007] Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *science*, 315(5814):972–976.
- [Furini and Montangero, 2016] Furini, M. and Montangero, M. (2016). Tsentiment: On gamifying twitter sentiment analysis. In *Computers and Communication (ISCC), 2016 IEEE Symposium on*, pages 91–96. IEEE.
- [Gallegos et al., 2016] Gallegos, L., Lerman, K., Huang, A., and Garcia, D. (2016). Geography of emotion: Where in a city are people happier? In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 569–574. International World Wide Web Conferences Steering Committee.
- [Gambino and Calvo, 2016] Gambino, O. J. and Calvo, H. (2016). A comparison between two spanish sentiment lexicons in the twitter sentiment analysis task. In *Ibero-American Conference on Artificial Intelligence*, pages 127–138. Springer.
- [Gandhe et al., 2018] Gandhe, K., Varde, A. S., and Du, X. (2018). Sentiment analysis of twitter data with hybrid learning for recommender applications. In *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 57–63. IEEE.
- [Ganesan and Zhai, 2012] Ganesan, K. and Zhai, C. (2012). Opinion-based entity ranking. *Information retrieval*, 15(2):116–150.
- [Gao et al., 2016] Gao, B., Berendt, B., and Vanschoren, J. (2016). Toward understanding online sentiment expression: an interdisciplinary approach with subgroup comparison and visualization. *Social Network Analysis and Mining*, 6(1):68.
- [Garg and Chatterjee, 2014] Garg, Y. and Chatterjee, N. (2014). Sentiment analysis of twitter feeds. In *International Conference on Big Data Analytics*, pages 33–52. Springer.

- [Gatt and Krahmer, 2018] Gatt, A. and Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.
- [Gauthier, 2001] Gauthier, T. D. (2001). Detecting trends using spearman’s rank correlation coefficient. *Environmental forensics*, 2(4):359–362.
- [Geetha et al., 2017] Geetha, M., Singha, P., and Sinha, S. (2017). Relationship between customer sentiment and online customer ratings for hotels-an empirical analysis. *Tourism Management*, 61:43–54.
- [Georgiadou et al., 2020] Georgiadou, E., Angelopoulos, S., and Drake, H. (2020). Big data analytics and international negotiations: Sentiment analysis of brexit negotiating outcomes. *International Journal of Information Management*, 51:102048.
- [Geurts et al., 2006] Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.
- [Ghiassi and Lee, 2018] Ghiassi, M. and Lee, S. (2018). A domain transferable lexicon set for twitter sentiment analysis using a supervised machine learning approach. *Expert Systems with Applications*, 106:197–216.
- [Ghiassi and Saidane, 2005] Ghiassi, M. and Saidane, H. (2005). A dynamic architecture for artificial neural networks. *Neurocomputing*, 63:397–413.
- [Ghosal et al., 2018] Ghosal, D., Akhtar, M. S., Ekbal, A., and Bhattacharyya, P. (2018). Deep ensemble model with the fusion of character, word and lexicon level information for emotion and sentiment prediction. In *International Conference on Neural Information Processing*, pages 162–174. Springer.
- [Giachanou and Crestani, 2016] Giachanou, A. and Crestani, F. (2016). Opinion retrieval in twitter: is proximity effective? In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 1146–1151. ACM.
- [Go et al., 2009] Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12).

- [Gonçalves et al., 2013] Gonçalves, P., Araújo, M., Benevenuto, F., and Cha, M. (2013). Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks*, pages 27–38. ACM.
- [Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.
- [Goonatilake and Herath, 2007] Goonatilake, R. and Herath, S. (2007). The volatility of the stock market and news. *International Research Journal of Finance and Economics*, 3(11):53–65.
- [Graham, 2015] Graham, Y. (2015). Re-evaluating automatic summarization with bleu and 192 shades of rouge. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 128–137.
- [Graves and Schmidhuber, 2005] Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610.
- [Greff et al., 2017] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. (2017). Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232.
- [Grießhaber et al., 2020] Grießhaber, D., Maucher, J., and Vu, N. T. (2020). Fine-tuning BERT for low-resource natural language understanding via active learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1158–1171, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- [Guerra et al., 2014] Guerra, P. C., Meira Jr, W., and Cardie, C. (2014). Sentiment analysis on evolving social streams: How self-report imbalances can help. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 443–452. ACM.

- [Guo and Li, 2019] Guo, X. and Li, J. (2019). A novel twitter sentiment analysis model with baseline correlation for financial market prediction with improved efficiency. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 472–477. IEEE.
- [Gupta et al., 2017] Gupta, U., Chatterjee, A., Srikanth, R., and Agrawal, P. (2017). A sentiment-and-semantics-based approach for emotion detection in textual conversations. *arXiv preprint arXiv:1707.06996*.
- [Hagge et al., 2017] Hagge, M., von Hoffen, M., Betzing, J. H., and Becker, J. (2017). Design and implementation of a toolkit for the aspect-based sentiment analysis of tweets. In *2017 IEEE 19th Conference on Business Informatics (CBI)*, volume 1, pages 379–387. IEEE.
- [Hanley and McNeil, 1982] Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.
- [Hao et al., 2017] Hao, Z., Cai, R., Yang, Y., Wen, W., and Liang, L. (2017). A dynamic conditional random field based framework for sentence-level sentiment analysis of chinese microblog. In *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, volume 1, pages 135–142. IEEE.
- [Hayashi et al., 2021] Hayashi, H., Budania, P., Wang, P., Ackerson, C., Neervannan, R., and Neubig, G. (2021). Wikiasp: A dataset for multi-domain aspect-based summarization. *Transactions of the Association for Computational Linguistics*, 9:211–225.
- [Heckerman et al., 1995] Heckerman, D., Geiger, D., and Chickering, D. M. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243.
- [Hemmatian and Sohrabi, 2019] Hemmatian, F. and Sohrabi, M. K. (2019). A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review*, 52(3):1495–1545.

- [Hernandez-Farias et al., 2014] Hernandez-Farias, I., Buscaldi, D., and Priego-Sánchez, B. (2014). Iradabe: Adapting english lexicons to the italian sentiment polarity classification task. In *First Italian Conference on Computational Linguistics (CLiC-it 2014) and the fourth International Workshop EVALITA2014*, pages 75–81.
- [Hilts and Yu, 2010] Hilts, A. and Yu, E. (2010). Modeling social media support for the elicitation of citizen opinion. In *Proceedings of the International Workshop on Modeling Social Media*, pages 1–4.
- [Hinton and Salakhutdinov, 2006] Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.
- [Hinton et al., 2012] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.
- [Hornik et al., 1989] Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feed-forward networks are universal approximators. *Neural networks*, 2(5):359–366.
- [Hossu and Parde, 2021] Hossu, P. and Parde, N. (2021). Using deep learning to correlate reddit posts with economic time series during the covid-19 pandemic. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, pages 8–14.
- [Hovy and Lin, 1999] Hovy, E. and Lin, C.-Y. (1999). Automatic text summarization and the summarist system. *Proceedings of the Tipster Text Program, phase III October 1996–October 1998*, pages 197–214.
- [Hridoy et al., 2015] Hridoy, S. A. A., Ekram, M. T., Islam, M. S., Ahmed, F., and Rahman, R. M. (2015). Localized twitter opinion mining using sentiment analysis. *Decision Analytics*, 2(1):8.
- [Hu and Liu, 2004] Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

- [Hu and Liu, 2006] Hu, M. and Liu, B. (2006). Opinion extraction and summarization on the web. In *Aaai*, volume 7, pages 1621–1624.
- [Hu et al., 2013] Hu, X., Tang, L., Tang, J., and Liu, H. (2013). Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 537–546. ACM.
- [Huang et al., 2006] Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K. (2006). Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501.
- [Huang et al., 2010] Huang, Y., Liu, Q., Zhang, S., and Metaxas, D. N. (2010). Image retrieval via probabilistic hypergraph ranking. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3376–3383. IEEE.
- [Huang et al., 2015] Huang, Y., Zhou, S., Huang, K., and Guan, J. (2015). Boosting financial trend prediction with twitter mood based on selective hidden markov models. In *International Conference on Database Systems for Advanced Applications*, pages 435–451. Springer.
- [Hubert et al., 2018] Hubert, R. B., Estevez, E., Maguitman, A., and Janowski, T. (2018). Examining government-citizen interactions on twitter using visual and sentiment analysis. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, page 55. ACM.
- [Hürlimann et al., 2016] Hürlimann, M., Davis, B., Cortis, K., Freitas, A., Handschuh, S., and Fernández, S. (2016). A twitter sentiment gold standard for the brexit referendum. In *SEMANTICS*, pages 193–196.
- [Hussain and Cambria, 2018] Hussain, A. and Cambria, E. (2018). Semi-supervised learning for big social data analysis. *Neurocomputing*, 275:1662–1673.
- [Hyun et al., 2020] Hyun, D., Cho, J., and Yu, H. (2020). Building large-scale english and korean datasets for aspect-level sentiment analysis in automotive domain. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 961–966.

- [Iskender et al., 2021] Iskender, N., Polzehl, T., and Möller, S. (2021). Reliability of human evaluation for text summarization: Lessons learned and challenges ahead. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 86–96.
- [Islam et al., 2020] Islam, K. I., Islam, M. S., and Amin, M. R. (2020). Sentiment analysis in bengali via transfer learning using multi-lingual bert. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pages 1–5. IEEE.
- [Javed et al., 2014] Javed, I., Afzal, H., Majeed, A., and Khan, B. (2014). Towards creation of linguistic resources for bilingual sentiment analysis of twitter data. In *International Conference on Applications of Natural Language to Data Bases/Information Systems*, pages 232–236. Springer.
- [Jaynes, 1957] Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review*, 106(4):620.
- [Jiang et al., 2020] Jiang, B., Hou, J., Zhou, W., Yang, C., Wang, S., and Pang, L. (2020). Metnet: A mutual enhanced transformation network for aspect-based sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 162–172.
- [Jiang et al., 2015] Jiang, F., Liu, Y.-Q., Luan, H.-B., Sun, J.-S., Zhu, X., Zhang, M., and Ma, S.-P. (2015). Microblog sentiment analysis with emoticon space model. *Journal of Computer Science and Technology*, 30(5):1120–1129.
- [Jiang et al., 2011] Jiang, L., Yu, M., Zhou, M., Liu, X., and Zhao, T. (2011). Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 151–160. Association for Computational Linguistics.
- [Jones et al., 2017] Jones, A., Ellman, J., and Jin, N. (2017). An application of sentiment analysis techniques to determine public opinion in social media. In *Proc. Int. Conf. Inf. Soc.*, pages 1–7.

- [Jones, 2007] Jones, K. S. (2007). Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449–1481.
- [Joulin et al., 2016] Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.
- [Kamal, 2015] Kamal, A. (2015). Review mining for feature based opinion summarization and visualization. *arXiv preprint arXiv:1504.03068*.
- [Kao and Huang, 2018] Kao, L.-J. and Huang, Y.-P. (2018). An effective social network sentiment mining model for healthcare product sales analysis. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2152–2157. IEEE.
- [Kaplan and Haenlein, 2010] Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68.
- [Kapočiūtė-Dzikienė et al., 2019] Kapočiūtė-Dzikienė, J., Damaševičius, R., and Woźniak, M. (2019). Sentiment analysis of lithuanian texts using traditional and deep learning approaches. *Computers*, 8(1):4.
- [Karyotis et al., 2017] Karyotis, C., Doctor, F., Iqbal, R., James, A., and Chang, V. (2017). A fuzzy computational model of emotion for cloud based sentiment analysis. *Information Sciences*.
- [Katsurai and Satoh, 2016] Katsurai, M. and Satoh, S. (2016). Image sentiment analysis using latent correlations among visual, textual, and sentiment views. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2837–2841. IEEE.
- [Katz et al., 2018] Katz, G., Heap, B., Wobcke, W., Bain, M., and Kannangara, S. (2018). Analysing tv audience engagement via twitter: Incremental segment-level opinion mining of second screen tweets. In *Pacific Rim International Conference on Artificial Intelligence*, pages 300–308. Springer.
- [Kemp, 2019] Kemp, S. (2019). Digital 2019 malta. DataReportal.

- [Khan et al., 2014] Khan, F. H., Bashir, S., and Qamar, U. (2014). Tom: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, 57:245–257.
- [Kim et al., 2011] Kim, H. D., Ganesan, K., Sondhi, P., and Zhai, C. (2011). Comprehensive review of opinion summarization.
- [Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- [Kitchenham, 2004] Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.
- [Kokab et al., 2022] Kokab, S. T., Asghar, S., and Naz, S. (2022). Transformer-based deep learning models for the sentiment analysis of social media data. *Array*, page 100157.
- [Kokkinogenis et al., 2015] Kokkinogenis, Z., Filguieras, J., Carvalho, S., Sarmento, L., and Rossetti, R. J. (2015). Mobility network evaluation in the user perspective: Real-time sensing of traffic information in twitter messages. *Advances in Artificial Transportation Systems and Simulation*, pages 219–234.
- [Kontopoulos et al., 2013] Kontopoulos, E., Berberidis, C., Dergiades, T., and Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert systems with applications*, 40(10):4065–4074.
- [Korakakis et al., 2017] Korakakis, M., Spyrou, E., and Mylonas, P. (2017). A survey on political event analysis in twitter. In *2017 12th international workshop on semantic and social media adaptation and personalization (SMAP)*, pages 14–19. IEEE.
- [Kothari et al., 2013] Kothari, A., Magdy, W., Darwish, K., Mourad, A., and Taei, A. (2013). Detecting comments on news articles in microblogs. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7.
- [Koto and Adriani, 2015a] Koto, F. and Adriani, M. (2015a). A comparative study on twitter sentiment analysis: Which features are good? In *International Conference on Applications of Natural Language to Information Systems*, pages 453–457. Springer.

- [Koto and Adriani, 2015b] Koto, F. and Adriani, M. (2015b). Hbe: Hashtag-based emotion lexicons for twitter sentiment analysis. In *Proceedings of the 7th Forum for Information Retrieval Evaluation*, pages 31–34. ACM.
- [Koto and Adriani, 2015c] Koto, F. and Adriani, M. (2015c). The use of pos sequence for analyzing sentence pattern in twitter sentiment analysis. In *Advanced Information Networking and Applications Workshops (WAINA), 2015 IEEE 29th International Conference on*, pages 547–551. IEEE.
- [Koto et al., 2020] Koto, F., Rahimi, A., Lau, J. H., and Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- [Kraus and Feuerriegel, 2017] Kraus, M. and Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*, 104:38–48.
- [Krippendorff, 2011] Krippendorff, K. (2011). Computing krippendorff’s alpha-reliability. https://repository.upenn.edu/asc_papers/43. Accessed: 2011-01-25.
- [Krippendorff, 2018] Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.
- [Kruspe et al., 2020] Kruspe, A., Häberle, M., Kuhn, I., and Zhu, X. X. (2020). Cross-language sentiment analysis of european twitter messages during the covid-19 pandemic. *arXiv preprint arXiv:2008.12172*.
- [Ku et al., 2006] Ku, L.-W., Liang, Y.-T., and Chen, H.-H. (2006). Opinion extraction, summarization and tracking in news and blog corpora. In *Proceedings of AAAI*, pages 100–107.
- [Kucher et al., 2018] Kucher, K., Paradis, C., and Kerren, A. (2018). The state of the art in sentiment visualization. In *Computer Graphics Forum*, volume 37, pages 71–96. Wiley Online Library.

- [Kumar and Garg, 2019] Kumar, A. and Garg, G. (2019). Sentiment analysis of multi-modal twitter data. *Multimedia Tools and Applications*, 78(17):24103–24119.
- [Lafferty et al., 2001] Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- [Lample and Conneau, 2019] Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.
- [Landis and Koch, 1977] Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- [Lau et al., 2014] Lau, R. Y., Li, C., and Liao, S. S. (2014). Social analytics: learning fuzzy product ontologies for aspect-oriented sentiment analysis. *Decision Support Systems*, 65:80–94.
- [LeCun et al., 1990] LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404.
- [Lek and Poo, 2013] Lek, H. H. and Poo, D. C. (2013). Aspect-based twitter sentiment classification. In *Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference on*, pages 366–373. IEEE.
- [Lewis, 1998] Lewis, D. D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages 4–15. Springer.
- [Li et al., 2018] Li, D., Rzepka, R., Ptaszynski, M., and Araki, K. (2018). Emoticon-aware recurrent neural network model for chinese sentiment analysis. In *2018 9th International Conference on Awareness Science and Technology (iCAST)*, pages 161–166. IEEE.

- [Li et al., 2016] Li, W., Li, Y., and Wang, Y. (2016). Chinese microblog sentiment analysis based on sentiment features. In *Asia-Pacific Web Conference*, pages 385–388. Springer.
- [Li and Xu, 2014] Li, W. and Xu, H. (2014). Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications*, 41(4):1742–1749.
- [Li and Li, 2013] Li, Y.-M. and Li, T.-Y. (2013). Deriving market intelligence from microblogs. *Decision Support Systems*, 55(1):206–217.
- [Ligthart et al., 2021] Ligthart, A., Catal, C., and Tekinerdogan, B. (2021). Systematic reviews in sentiment analysis: a tertiary study. *Artificial Intelligence Review*, pages 1–57.
- [Lima et al., 2015] Lima, A. C. E., de Castro, L. N., and Corchado, J. M. (2015). A polarity analysis framework for twitter messages. *Applied Mathematics and Computation*, 270:756–767.
- [Lin et al., 2018] Lin, B. Y., Xu, F. F., Zhu, K., and Hwang, S.-w. (2018). Mining cross-cultural differences and similarities in social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 709–719.
- [Lin and Hovy, 2003] Lin, C.-Y. and Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, pages 150–157.
- [Liu, 2007] Liu, B. (2007). Opinion mining. *Web data mining: Exploring hyperlinks, contents, and usage data*, pages 411–447.
- [Liu, 2010] Liu, B. (2010). Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition*. Taylor and Francis Group, Boca.
- [Liu, 2012] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

- [Liu, 2015] Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.
- [Liu et al., 2005] Liu, B., Hu, M., and Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351.
- [Liu and Zhang, 2012] Liu, B. and Zhang, L. (2012). *Mining Text Data*, chapter A Survey of Opinion Mining and Sentiment Analysis, pages 415–463. Springer US, Boston, MA.
- [Liu et al., 2018] Liu, Q., Hu, Y., Lei, Y., Wei, X., Liu, G., and Bi, W. (2018). Topic-based microblog polarity classification based on cascaded model. In *International Conference on Computational Science*, pages 206–220. Springer.
- [Liu et al., 2015] Liu, T., Jiang, F., Liu, Y., Zhang, M., and Ma, S. (2015). Do photos help express our feelings: Incorporating multimodal features into microblog sentiment analysis. In *Chinese National Conference on Social Media Processing*, pages 63–73. Springer.
- [Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- [Lloret et al., 2018] Lloret, E., Plaza, L., and Aker, A. (2018). The challenging task of summary evaluation: an overview. *Language Resources and Evaluation*, 52(1):101–148.
- [Lloyd, 1982] Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.
- [López et al., 2021] López, M., Martínez-Cámara, E., Luzón, M. V., and Herrera, F. (2021). Adops: Aspect discovery opinion summarisation methodology based on deep learning and subgroup discovery for generating explainable opinion summaries. *Knowledge-Based Systems*, 231:107455.

- [Loughran and McDonald, 2011] Loughran, T. and McDonald, B. (2011). "when is a liability not a liability? textual analysis, dictionaries, and 10-ks". *The Journal of Finance*, 66(1):35–65.
- [Makarenkov and Rokach, 2020] Makarenkov, V. and Rokach, L. (2020). Lessons learned from applying off-the-shelf bert: There is no silverbullet. *arXiv preprint arXiv:2009.07238*.
- [Mamta et al., 2020] Mamta, Ekbal, A., Bhattacharyya, P., Srivastava, S., Kumar, A., and Saha, T. (2020). Multi-domain tweet corpora for sentiment analysis: Resource creation and evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5046–5054, Marseille, France. European Language Resources Association.
- [Mancosu and Bobba, 2019] Mancosu, M. and Bobba, G. (2019). Using deep-learning algorithms to derive basic characteristics of social media users: The brexit campaign as a case study. *Plos one*, 14(1):e0211013.
- [Mani et al., 2002] Mani, I., Klein, G., House, D., Hirschman, L., Firmin, T., and Sundheim, B. (2002). Summac: a text summarization evaluation. *Natural Language Engineering*, 8(1):43–68.
- [Mansour et al., 2015] Mansour, R., Hady, M. F. A., Hosam, E., Amr, H., and Ashour, A. (2015). Feature selection for twitter sentiment analysis: An experimental study. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 92–103. Springer.
- [Martin, 2020] Martin, I. (2020). 84% say times of malta is their main source of online news. <https://timesofmalta.com/articles/view/times-of-malta-is-country-s-most-popular-news-site-and-paper-survey.800391>. Times of Malta.
- [Matthews, 1975] Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

- [Maynard et al., 2012] Maynard, D., Bontcheva, K., and Rout, D. (2012). Challenges in developing opinion mining tools for social media. *Proceedings of the@ NLP can u tag# usergeneratedcontent*, pages 15–22.
- [McCallum et al., 1998] McCallum, A., Nigam, K., et al. (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer.
- [McCullagh, 1984] McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, 16(3):285–292.
- [McNair and Droppleman, 1971] McNair, D.M., L. M. and Droppleman, L. (1971). *Profile of mood states*. Educational & Industrial testing service.
- [Medhat et al., 2014] Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.
- [Mehrabian, 1996] Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292.
- [Mejova and Srinivasan, 2012] Mejova, Y. and Srinivasan, P. (2012). Political speech in social media streams: Youtube comments and twitter posts. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 205–208. ACM.
- [Mejova et al., 2013] Mejova, Y., Srinivasan, P., and Boynton, B. (2013). Gop primary season on twitter: popular political sentiment in social media. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 517–526. ACM.
- [Mellon and Prosser, 2017] Mellon, J. and Prosser, C. (2017). Twitter and facebook are not representative of the general population: Political attitudes and demographics of british social media users. *Research & Politics*, 4(3):2053168017720008.
- [Meng et al., 2012] Meng, X., Wei, F., Liu, X., Zhou, M., Li, S., and Wang, H. (2012). Entity-centric topic-oriented opinion summarization in twitter. In *Proceedings of the*

- 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 379–387.
- [Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- [Miller, 1995] Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- [Min et al., 2013] Min, M., Lee, T., and Hsu, R. (2013). Role of emoticons in sentence-level sentiment classification. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 203–213. Springer.
- [Mirza and Osindero, 2014] Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.
- [Mishev et al., 2020] Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., and Trajanov, D. (2020). Evaluation of sentiment analysis in finance: from lexicons to transformers. *IEEE Access*, 8:131662–131682.
- [Moh et al., 2017] Moh, M., Moh, T.-S., Peng, Y., and Wu, L. (2017). On adverse drug event extractions using twitter sentiment analysis. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 6(1):18.
- [Mohammad et al., 2018] Mohammad, S., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018). Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.
- [Mohammad et al., 2013] Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA.
- [Mohammad and Turney, 2010] Mohammad, S. M. and Turney, P. D. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to*

- analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics.
- [Mohammad and Turney, 2013] Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.
- [Montejo-Raez et al., 2014] Montejo-Raez, A., Díaz-Galiano, M. C., Martínez-Santiago, F., and Ureña-López, L. (2014). Crowd explicit sentiment analysis. *Knowledge-Based Systems*, 69:134–139.
- [More, 2016] More, A. (2016). Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv preprint arXiv:1608.06048*.
- [Morency et al., 2011] Morency, L.-P., Mihalcea, R., and Doshi, P. (2011). Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176. ACM.
- [Moussa et al., 2018] Moussa, M. E., Mohamed, E. H., and Haggag, M. H. (2018). A survey on opinion summarization techniques for social media. *Future Computing and Informatics Journal*, 3(1):82–109.
- [Mozetič et al., 2016] Mozetič, I., Grčar, M., and Smailović, J. (2016). Twitter sentiment for 15 european languages. Slovenian language resource repository CLARIN.SI.
- [Mukherjee et al., 2020] Mukherjee, R., Peruri, H. C., Vishnu, U., Goyal, P., Bhattacharya, S., and Ganguly, N. (2020). Read what you need: Controllable aspect-based opinion summarization of tourist reviews. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1825–1828.
- [Mukhtar and Khan, 2019] Mukhtar, N. and Khan, M. A. (2019). Effective lexicon-based approach for urdu sentiment analysis. *Artificial Intelligence Review*, pages 1–28.
- [Müller et al., 2020] Müller, M., Salathé, M., and Kummervold, P. E. (2020). Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

- [Munezero et al., 2015] Munezero, M., Montero, C. S., Mozgovoy, M., and Sutinen, E. (2015). Emotwitter-a fine-grained visualization system for identifying enduring sentiments in tweets. *CICLing (2)*, 9042:78–91.
- [Nakov et al., 2013] Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., and Wilson, T. (2013). Semeval-2013 task 2: Sentiment analysis in twitter. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval-2013)*.
- [Nakov et al., 2016] Nakov, P., Ritter, A., Rosenthal, S., Stoyanov, V., and Sebastiani, F. (2016). SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval '16*, San Diego, California. Association for Computational Linguistics.
- [Narr et al., 2012] Narr, S., Hulphenhaus, M., and Albayrak, S. (2012). Language-independent twitter sentiment analysis. *Knowledge discovery and machine learning (KDML), LWA*, pages 12–14.
- [Naseem et al., 2020] Naseem, U., Razzak, I., Musial, K., and Imran, M. (2020). Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems*, 113:58–69.
- [Nausheen and Begum, 2018] Nausheen, F. and Begum, S. H. (2018). Sentiment analysis to predict election results using python. In *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, pages 1259–1262. IEEE.
- [Nguyen et al., 2020] Nguyen, D. Q., Vu, T., and Nguyen, A. T. (2020). BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.
- [Nguyen et al., 2018] Nguyen, M.-T., Lai, D. V., Nguyen, H. T., and Le Nguyen, M. (2018). Tsix: a human-involved-creation dataset for tweet summarization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

- [Nielsen, 2012] Nielsen, J. (2012). Usability 101: Introduction to usability (2012). URL: <http://www.nngroup.com/articles/usability-101-introduction-to-usability/>[Accessed November 2016], 9:35.
- [Nozza et al., 2020] Nozza, D., Bianchi, F., and Hovy, D. (2020). What the [mask]? making sense of language-specific bert models. *arXiv preprint arXiv:2003.02912*.
- [Öhman et al., 2020] Öhman, E., Pàmies, M., Kajava, K., and Tiedemann, J. (2020). XED: A multilingual dataset for sentiment analysis and emotion detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6542–6552, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- [Ontotext, 2016] Ontotext (2016). #brexit twitter analysis.
- [Orăsan, 2019] Orăsan, C. (2019). Automatic summarisation: 25 years on. *Natural Language Engineering*, 25(6):735–751.
- [Orellana-Rodriguez et al., 2013] Orellana-Rodriguez, C., Diaz-Aviles, E., and Nejdl, W. (2013). Mining emotions in short films: user comments or crowdsourcing? In *Proceedings of the 22nd International Conference on World Wide Web*, pages 69–70. ACM.
- [Orellana-Rodriguez et al., 2015] Orellana-Rodriguez, C., Diaz-Aviles, E., and Nejdl, W. (2015). Mining affective context in short films for emotion-aware recommendation. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 185–194. ACM.
- [Ortis et al., 2018] Ortis, A., Farinella, G. M., Torrisi, G., and Battiato, S. (2018). Visual sentiment analysis based on on objective text description of images. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE.
- [Ortony et al., 1988] Ortony, A., Clore, G., and Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press.
- [Ou et al., 2014] Ou, G., Chen, W., Li, B., Wang, T., Yang, D., and Wong, K.-F. (2014). Clusm: an unsupervised model for microblog sentiment analysis incorporating link in-

- formation. In *International Conference on Database Systems for Advanced Applications*, pages 481–494. Springer.
- [Padurariu and Breaban, 2019] Padurariu, C. and Breaban, M. E. (2019). Dealing with data imbalance in text classification. *Procedia Computer Science*, 159:736–745.
- [Pai and Alathur, 2018] Pai, R. R. and Alathur, S. (2018). Assessing mobile health applications with twitter analytics. *International journal of medical informatics*, 113:72–84.
- [Palinkas et al., 2015] Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., and Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and policy in mental health and mental health services research*, 42(5):533–544.
- [Pan and Yang, 2009] Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.
- [Pandarachalil et al., 2015] Pandarachalil, R., Sendhilkumar, S., and Mahalakshmi, G. (2015). Twitter sentiment analysis for large-scale data: an unsupervised approach. *Cognitive computation*, 7(2):254–262.
- [Pandey et al., 2017] Pandey, A. C., Rajpoot, D. S., and Saraswat, M. (2017). Twitter sentiment analysis using hybrid cuckoo search method. *Information Processing & Management*, 53(4):764–779.
- [Pang and Lee, 2008] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1-2):1–135.
- [Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- [Patwa et al., 2020] Patwa, P., Aguilar, G., Kar, S., Pandey, S., PYKL, S., Gambäck, B., Chakraborty, T., Solorio, T., and Das, A. (2020). Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the 14th International*

- Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain. Association for Computational Linguistics.
- [Pecar, 2018] Pecar, S. (2018). Towards opinion summarization of customer reviews. In *Proceedings of ACL 2018, Student Research Workshop*, pages 1–8.
- [Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- [Pereira, 2021] Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, 54(2):1087–1115.
- [Peters et al., 2018] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- [Plutchik, 1980] Plutchik, R. (1980). Chapter 1 - a general psychoevolutionary theory of emotion. In Plutchik, R. and Kellerman, H., editors, *Theories of Emotion*, pages 3 – 33. Academic Press.
- [Polignano et al., 2019] Polignano, M., Basile, P., De Gemmis, M., Semeraro, G., and Basile, V. (2019). Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR.
- [Pontiki et al., 2016] Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, pages 19–30.
- [Poria et al., 2016] Poria, S., Cambria, E., Howard, N., Huang, G.-B., and Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59.

- [Porshnev and Redkin, 2014] Porshnev, A. and Redkin, I. (2014). Analysis of twitter users' mood for prediction of gold and silver prices in the stock market. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 190–197. Springer.
- [Porshnev et al., 2014] Porshnev, A., Redkin, I., and Karpov, N. (2014). Modelling movement of stock market indexes with data from emoticons of twitter users. In *Russian Summer School in Information Retrieval*, pages 297–306. Springer.
- [Potthast and Becker, 2010] Potthast, M. and Becker, S. (2010). Opinion summarization of web comments. In *European Conference on Information Retrieval*, pages 668–669. Springer.
- [Prabowo and Thelwall, 2009] Prabowo, R. and Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2):143–157.
- [Pustejovsky and Stubbs, 2012] Pustejovsky, J. and Stubbs, A. (2012). *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications.* ” O'Reilly Media, Inc.”.
- [Qaisi and Aljarah, 2016] Qaisi, L. M. and Aljarah, I. (2016). A twitter sentiment analysis for cloud providers: a case study of azure vs. aws. In *Computer Science and Information Technology (CSIT), 2016 7th International Conference on*, pages 1–6. IEEE.
- [Quinlan, 1986] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.
- [Rabelo et al., 2012] Rabelo, J. C., Prudêncio, R. B., and Barros, F. A. (2012). Using link structure to infer opinions in social networks. In *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*, pages 681–685. IEEE.
- [Radford et al., 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

- [Radhika and Sankar, 2017] Radhika, N. and Sankar, S. (2017). Personalized language-independent music recommendation system. In *2017 International Conference on Intelligent Computing and Control (I2C2)*, pages 1–6. IEEE.
- [Rai et al., 2018] Rai, A., Minsker, B., Diesner, J., Karahalios, K., and Sun, Y. (2018). Identification of landscape preferences by using social media analysis. In *2018 International Workshop on Social Sensing (SocialSens)*, pages 44–49. IEEE.
- [Raja and Swamynathan, 2016] Raja, M. and Swamynathan, S. (2016). Tweet sentiment analyzer: Sentiment score estimation method for assessing the value of opinions in tweets. In *Proceedings of the International Conference on Advances in Information Communication Technology & Computing*, page 83. ACM.
- [Ramirez et al., 2010] Ramirez, I., Sprechmann, P., and Sapiro, G. (2010). Classification and clustering via dictionary learning with structured incoherence and shared features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3501–3508. IEEE.
- [Rani and Kumar, 2019] Rani, S. and Kumar, P. (2019). A journey of indian languages over sentiment analysis: a systematic review. *Artificial Intelligence Review*, 52(2):1415–1462.
- [Årup Nielsen, 2011] Årup Nielsen, F. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Making Sense of Microposts (#MSM2011)*, pages 93–98.
- [Rathan et al., 2018] Rathan, M., Hulipalled, V. R., Venugopal, K., and Patnaik, L. (2018). Consumer insight mining: Aspect based twitter opinion mining of mobile phone reviews. *Applied Soft Computing*, 68:765–773.
- [Ravi and Ravi, 2015] Ravi, K. and Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89:14 – 46.

- [Ray and Chakrabarti, 2017] Ray, P. and Chakrabarti, A. (2017). Twitter sentiment analysis for product review using lexicon method. In *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, pages 211–216. IEEE.
- [Reiter and Belz, 2009] Reiter, E. and Belz, A. (2009). An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.
- [Reyes et al., 2013] Reyes, A., Rosso, P., and Veale, T. (2013). A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 47(1):239–268.
- [Rosenthal et al., 2017] Rosenthal, S., Farra, N., and Nakov, P. (2017). SemEval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.
- [Rosenthal et al., 2015] Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., and Stoyanov, V. (2015). Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 451–463.
- [Rosenthal et al., 2014] Rosenthal, S., Ritter, A., Nakov, P., and Stoyanov, V. (2014). Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*.
- [Rosner and Borg, 2022] Rosner, M. and Borg, C. (2022). Project european language equality (ele) d1.25 report on the maltese language.
- [Rosner et al., 2012] Rosner, M., Joachimsen, J., Rehm, G., and Uszkoreit, H. (2012). *The Maltese language in the digital age*. Springer.
- [Rousseeuw, 1984] Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880.

- [Rout et al., 2018] Rout, J. K., Choo, K.-K. R., Dash, A. K., Bakshi, S., Jena, S. K., and Williams, K. L. (2018). A model for sentiment and emotion analysis of unstructured social media text. *Electronic Commerce Research*, 18(1):181–199.
- [Rout et al., 2017] Rout, J. K., Singh, S., Jena, S. K., and Bakshi, S. (2017). Deceptive review detection using labeled and unlabeled data. *Multimedia Tools and Applications*, 76(3):3187–3211.
- [Ruder, 2017] Ruder, S. (2017). Transfer Learning - Machine Learning’s Next Frontier. <http://ruder.io/transfer-learning/>.
- [Rui et al., 2013] Rui, H., Liu, Y., and Whinston, A. (2013). Whose and what chatter matters? the effect of tweets on movie sales. *Decision Support Systems*, 55(4):863–870.
- [Rumelhart et al., 1985] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.
- [Sahni et al., 2017] Sahni, T., Chandak, C., Chedeti, N. R., and Singh, M. (2017). Efficient twitter sentiment classification using subjective distant supervision. In *2017 9th International Conference on Communication Systems and Networks (COMSNETS)*, pages 548–553. IEEE.
- [Saías et al., 2018] Saías, J., Mourão, M., and Oliveira, E. (2018). Detailing sentiment analysis to consider entity aspects: an approach for portuguese short texts.
- [Saif et al., 2013] Saif, H., Fernandez, M., He, Y., and Alani, H. (2013). Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold.
- [Saini et al., 2018] Saini, S., Rao, R., Vaichole, V., Rane, A., and Abin, D. (2018). Emotion recognition using multimodal approach. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCCUBEA)*, pages 1–4. IEEE.
- [Salton and McGill, 1986] Salton, G. and McGill, M. J. (1986). Introduction to modern information retrieval.

- [Salvatore et al., 2021] Salvatore, C., Biffignandi, S., and Bianchi, A. (2021). Social media and twitter data quality for new social indicators. *Social Indicators Research*, 156(2):601–630.
- [Samoylov, 2014] Samoylov, A. B. (2014). Evaluation of the delta tf-idf features for sentiment analysis. In *International Conference on Analysis of Images, Social Networks and Texts_x000D_*, pages 207–212. Springer.
- [Sandoval-Almazan and Valle-Cruz, 2018] Sandoval-Almazan, R. and Valle-Cruz, D. (2018). Facebook impact and sentiment analysis on political campaigns. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, page 56. ACM.
- [Sanh et al., 2019] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- [Santarcangelo et al., 2015] Santarcangelo, V., Oddo, G., Pilato, M., Valenti, F., and Fornaro, C. (2015). Social opinion mining: an approach for italian language. In *Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference on*, pages 693–697. IEEE.
- [Sarsam et al., 2020] Sarsam, S. M., Al-Samarraie, H., Alzahrani, A. I., and Wright, B. (2020). Sarcasm detection using machine learning algorithms in twitter: A systematic review. *International Journal of Market Research*, 62(5):578–598.
- [Schembri, 2019] Schembri, S. (2019). Malta: The ultimate ai launchpad - a strategy and vision for artificial intelligence in malta 2030. https://malta.ai/wp-content/uploads/2019/10/Malta_The_Ultimate_AI_Launchpad_vFinal.pdf. Parliamentary Secretariat for Financial Services, Digital Economy and Innovation, Office of the Prime Minister.
- [Schuster, 2003] Schuster, T. (2003). Meta-communication and market dynamics. reflexive interactions of financial markets and the mass media.

- [scikit learn, a] scikit learn. Bernoulli naïve bayes. https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html.
- [scikit learn, b] scikit learn. Complement naïve bayes. https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.ComplementNB.html.
- [scikit learn, c] scikit learn. Decision tree. <https://scikit-learn.org/stable/modules/tree.html>.
- [scikit learn, d] scikit learn. Linear support vector classification. <https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html>.
- [scikit learn, e] scikit learn. Logistic regression. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.
- [scikit learn, f] scikit learn. Nu-support support vector classification. <https://scikit-learn.org/stable/modules/generated/sklearn.svm.NuSVC.html>.
- [scikit learn, g] scikit learn. Random forest. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>.
- [scikit learn, h] scikit learn. Support vector classification. <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>.
- [Serrano-Guerrero et al., 2015] Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., and Herrera-Viedma, E. (2015). Sentiment analysis: a review and comparative analysis of web services. *Information Sciences*, 311:18–38.
- [Shamim et al., 2015] Shamim, A., Balakrishnan, V., and Tahir, M. (2015). Evaluation of opinion visualization techniques. *Information visualization*, 14(4):339–358.
- [Shamim et al., 2021] Shamim, A., Qureshi, M. A., Jabeen, F., Liaqat, M., Bilal, M., Jembre, Y. Z., and Attique, M. (2021). Multi-attribute online decision-making driven by opinion mining. *Mathematics*, 9(8):833.
- [Shamma et al., 2009] Shamma, D. A., Kennedy, L., and Churchill, E. F. (2009). Tweet the debates: understanding community annotation of uncollected sources. In *Proceedings of the first SIGMM workshop on Social media*, pages 3–10. ACM.

- [Sheth et al., 2014] Sheth, A., Jadhav, A., Kapanipathi, P., Lu, C., Purohit, H., Smith, G. A., and Wang, W. (2014). Twitris: A system for collective social intelligence. In *Encyclopedia of Social Network Analysis and Mining*, pages 2240–2253. Springer.
- [Shukri et al., 2015] Shukri, S. E., Yaghi, R. I., Aljarah, I., and Alsawalqah, H. (2015). Twitter sentiment analysis: A case study in the automotive industry. In *Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on*, pages 1–5. IEEE.
- [Singh et al., 2018] Singh, N., Roy, N., and Gangopadhyay, A. (2018). Analyzing the sentiment of crowd for improving the emergency response services. In *2018 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 1–8. IEEE.
- [Sinha, 2014] Sinha, N. (2014). Using big data in finance: Example of sentiment-extraction from news articles. <http://www.federalreserve.gov/econresdata/notes/feds-notes/2014/using-big-data-in-finance-example-of-sentiment-extraction-from-news-articles-20140326.html>. Accessed 29-March-2016.
- [Sobhani et al., 2015] Sobhani, P., Inkpen, D., and Matwin, S. (2015). From argumentation mining to stance classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 67–77.
- [Socher et al., 2013] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.
- [Somasundaran and Wiebe, 2010] Somasundaran, S. and Wiebe, J. (2010). Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 116–124.

- [Song and Gruzd, 2017] Song, M. Y.-J. and Gruzd, A. (2017). Examining sentiments and popularity of pro-and anti-vaccination videos on youtube. In *Proceedings of the 8th International Conference on Social Media & Society*, page 17. ACM.
- [Sotiropoulos et al., 2019] Sotiropoulos, K., Byers, J. W., Pratikakis, P., and Tsourakakis, C. E. (2019). Twittermancer: Predicting interactions on twitter accurately. *arXiv preprint arXiv:1904.11119*.
- [Soulis et al., 2013] Soulis, K., Varlamis, I., Giannakoulopoulos, A., and Charatsev, F. (2013). A tool for the visualisation of public opinion. *International Journal of Electronic Governance*, 6(3):218–231.
- [Speriosu et al., 2011] Speriosu, M., Sudan, N., Upadhyay, S., and Baldrige, J. (2011). Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 53–63. Association for Computational Linguistics.
- [Stojanovski et al., 2018] Stojanovski, D., Strezoski, G., Madjarov, G., Dimitrovski, I., and Chorbev, I. (2018). Deep neural network architecture for sentiment analysis and emotion identification of twitter messages. *Multimedia Tools and Applications*, 77(24):32213–32242.
- [Streijl et al., 2016] Streijl, R. C., Winkler, S., and Hands, D. S. (2016). Mean opinion score (mos) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, 22(2):213–227.
- [Strubell et al., 2019] Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*.
- [Sumanth and Inkpen, 2015] Sumanth, C. and Inkpen, D. (2015). How much does word sense disambiguation help in sentiment analysis of micropost data? In *Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 115–121.

- [Sun et al., 2019] Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics*, pages 194–206. Springer.
- [Supovitz et al., 2018] Supovitz, J., Daly, A. J., and Del Fresno, M. (2018). The common core debate on twitter and the rise of the activist public. *Journal of Educational Change*, 19(4):419–440.
- [Susanto et al., 2020] Susanto, Y., Livingstone, A. G., Ng, B. C., and Cambria, E. (2020). The hourglass model revisited. *IEEE Intelligent Systems*, 35(5):96–102.
- [Sykora et al., 2020] Sykora, M., Elayan, S., and Jackson, T. W. (2020). A qualitative analysis of sarcasm, irony and related# hashtags on twitter. *Big Data & Society*, 7(2):2053951720972735.
- [Taddy, 2013] Taddy, M. (2013). Measuring political sentiment on twitter: factor optimal design for multinomial inverse regression. *Technometrics*, 55(4):415–425.
- [Takala et al., 2014] Takala, P., Malo, P., Sinha, A., and Ahlgren, O. (2014). Gold-standard for topic-specific sentiment analysis of economic texts. In *LREC*, volume 2014, pages 2152–2157. Citeseer.
- [Tan et al., 2014] Tan, S. S., Soon, L.-K., Lim, T. Y., Tang, E. K., and Loo, C. K. (2014). Learning the mapping rules for sentiment analysis. In *Proceedings of the 5th International Workshop on Web-scale Knowledge Representation Retrieval & Reasoning*, pages 19–22. ACM.
- [Teixeira and Laureano, 2017] Teixeira, A. and Laureano, R. M. (2017). Data extraction and preparation to perform a sentiment analysis using open source tools: The example of a facebook fashion brand page. In *2017 12th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6. IEEE.
- [Tetlock et al., 2008] Tetlock, P. C., Saar-Tsechansky, M., and Macskassy, S. (2008). More than words: Quantifying language to measure firms’ fundamentals. *The Journal of Finance*, 63(3):1437–1467.

- [Thelwall et al., 2011] Thelwall, M., Buckley, K., and Paltoglou, G. (2011). Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418.
- [Thelwall et al., 2012] Thelwall, M., Buckley, K., and Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173.
- [Tomek, 1976] Tomek, I. (1976). Two modifications of cnn. *IEEE Trans. Systems, Man and Cybernetics*, 6:769–772.
- [Tong et al., 2017] Tong, Y., Zhou, B., and Huang, J. (2017). Topic-adaptive sentiment analysis on tweets via learning from multi-sources data. In *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, volume 1, pages 241–246. IEEE.
- [Troussas et al., 2016] Troussas, C., Krouska, A., and Virvou, M. (2016). Evaluation of ensemble-based sentiment classifiers for twitter data. In *Information, Intelligence, Systems & Applications (IISA), 2016 7th International Conference on*, pages 1–6. IEEE.
- [Trupthi et al., 2017] Trupthi, M., Pabboju, S., and Narasimha, G. (2017). Sentiment analysis on twitter using streaming api. In *2017 IEEE 7th International Advance Computing Conference (IACC)*, pages 915–919. IEEE.
- [Tsytsarau and Palpanas, 2012] Tsytsarau, M. and Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3):478–514.
- [Tullis and Stetson, 2004] Tullis, T. S. and Stetson, J. N. (2004). A comparison of questionnaires for assessing website usability. In *Usability professional association conference*, volume 1, pages 1–12. Minneapolis, USA.
- [Tumasjan et al., 2010] Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10(1):178–185.

- [Tunkelang, 2009] Tunkelang, D. (2009). Faceted search. *Synthesis lectures on information concepts, retrieval, and services*, 1(1):1–80.
- [Unankard et al., 2014] Unankard, S., Li, X., Sharaf, M., Zhong, J., and Li, X. (2014). Predicting elections from social networks based on sub-event detection and sentiment analysis. In *International Conference on Web Information Systems Engineering*, pages 1–16. Springer.
- [van der Lee et al., 2021] van der Lee, C., Gatt, A., van Miltenburg, E., and Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.
- [Van Hee, 2017] Van Hee, C. (2017). *Can machines sense irony?: exploring automatic irony detection on social media*. PhD thesis, Ghent University.
- [Van Hee et al., 2018a] Van Hee, C., Lefever, E., and Hoste, V. (2018a). Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.
- [Van Hee et al., 2018b] Van Hee, C., Lefever, E., and Hoste, V. (2018b). Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.
- [Varshney and Gupta, 2014] Varshney, N. and Gupta, S. (2014). Mining churning factors in indian telecommunication sector using social media analytics. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 405–413. Springer.
- [Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.
- [Vilarinho and Ruiz, 2018] Vilarinho, G. and Ruiz, E. (2018). Global centrality measures in word graphs for twitter sentiment analysis. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 55–60. IEEE.

- [Vincent et al., 2008] Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM.
- [Vo et al., 2017] Vo, T. H., Nguyen, T. T., Pham, H. A., and Van Le, T. (2017). An efficient hybrid model for vietnamese sentiment analysis. In *Asian Conference on Intelligent Information and Database Systems*, pages 227–237. Springer.
- [Wadawadagi and Pagi, 2020] Wadawadagi, R. and Pagi, V. (2020). Sentiment analysis with deep neural networks: comparative study and performance assessment. *Artificial Intelligence Review*, 53:6155–6195.
- [Walha et al., 2016] Walha, A., Ghozzi, F., and Gargouri, F. (2016). Etl design toward social network opinion analysis. In *Computer and Information Science 2015*, pages 235–249. Springer.
- [Wang and Wu, 2015] Wang, F. and Wu, Y. (2015). Sentiment-bearing new words mining: Exploiting emoticons and latent polarities. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 166–179. Springer.
- [Wang et al., 2012] Wang, H., Can, D., Kazemzadeh, A., Bar, F., and Narayanan, S. (2012). A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120. Association for Computational Linguistics.
- [Wang et al., 2018a] Wang, H., Zhang, F., Hou, M., Xie, X., Guo, M., and Liu, Q. (2018a). Shine: Signed heterogeneous information network embedding for sentiment link prediction. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 592–600. ACM.
- [Wang and Ye, 2013] Wang, J.-H. and Ye, T.-W. (2013). Unsupervised opinion targets expansion and modification relation identification for microblog sentiment analysis. In *International Conference on Social Informatics*, pages 255–267. Springer.

- [Wang et al., 2014a] Wang, M., Cao, D., Li, L., Li, S., and Ji, R. (2014a). Microblog sentiment analysis based on cross-media bag-of-words model. In *Proceedings of international conference on internet multimedia computing and service*, page 76. ACM.
- [Wang and Manning, 2012] Wang, S. and Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2*, pages 90–94. Association for Computational Linguistics.
- [Wang et al., 2016] Wang, Y., Feng, S., Wang, D., Zhang, Y., and Yu, G. (2016). Context-aware chinese microblog sentiment classification with bidirectional lstm. In *Asia-Pacific Web Conference*, pages 594–606. Springer.
- [Wang et al., 2018b] Wang, Y., Kim, K., Lee, B., and Youn, H. Y. (2018b). Word clustering based on pos feature for efficient twitter sentiment analysis. *Human-centric Computing and Information Sciences*, 8(1):17.
- [Wang et al., 2014b] Wang, Z., Yu, Z., Chen, L., and Guo, B. (2014b). Sentiment detection and visualization of chinese micro-blog. In *Data Science and Advanced Analytics (DSAA), 2014 International Conference on*, pages 251–257. IEEE.
- [Wehrmann et al., 2017] Wehrmann, J., Becker, W., Cagnini, H. E., and Barros, R. C. (2017). A character-based convolutional neural network for language-agnostic twitter sentiment analysis. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2384–2391. IEEE.
- [Weichselbraun et al., 2017] Weichselbraun, A., Gindl, S., Fischer, F., Vakulenko, S., and Scharl, A. (2017). Aspect-based extraction and analysis of affective knowledge from social media streams. *IEEE Intelligent Systems*, 32(3):80–88.
- [Weiss et al., 2015] Weiss, S. M., Indurkha, N., and Zhang, T. (2015). Data sources for prediction: Databases, hybrid data and the web. In *Fundamentals of Predictive Text Mining*, pages 147–164. Springer.
- [Wilson et al., 2005] Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on*

- human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.
- [Witten et al., 2016] Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
- [Wolf et al., 2019] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
- [Wu et al., 2016] Wu, F., Huang, Y., Song, Y., and Liu, S. (2016). Towards building a high-quality microblog-specific chinese sentiment lexicon. *Decision Support Systems*, 87:39–49.
- [Xia et al., 2020] Xia, P., Wu, S., and Van Durme, B. (2020). Which *BERT? A survey organizing contextualized encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7516–7533, Online. Association for Computational Linguistics.
- [Xing et al., 2020] Xing, F., Malandri, L., Zhang, Y., and Cambria, E. (2020). Financial sentiment analysis: An investigation into common mistakes and silver bullets. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 978–987.
- [Xu et al., 2020] Xu, J., Masuda, K., Nishizaki, H., Fukumoto, F., and Suzuki, Y. (2020). Semi-automatic construction and refinement of an annotated corpus for a deep learning framework for emotion classification. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1611–1617.
- [Xu et al., 2012] Xu, J.-M., Zhu, X., and Bellmore, A. (2012). Fast learning for sentiment analysis on bullying. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 10. ACM.
- [Xuereb, 2020] Xuereb, M. (2020). ‘we consulted online media’ - justice minister on rule-of-law reform bills. <https://timesofmalta.com/articles/view/we-consulted-soc>

- ial-media-justice-minister-on-rule-of-law-reform-bills.825589. Times of Malta.
- [Yadav and Vishwakarma, 2020] Yadav, A. and Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6):4335–4385.
- [Yadollahi et al., 2017] Yadollahi, A., Shahraki, A. G., and Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):1–33.
- [Yan et al., 2014] Yan, G., He, W., Shen, J., and Tang, C. (2014). A bilingual approach for conducting chinese and english social media sentiment analysis. *Computer Networks*, 75:491–503.
- [Yang et al., 2014] Yang, C.-H., Chen, J.-D., and Kao, H.-Y. (2014). Competition component identification on twitter. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 584–595. Springer.
- [Yang et al., 2019] Yang, W., Zhang, H., and Lin, J. (2019). Simple applications of bert for ad hoc document retrieval. *arXiv preprint arXiv:1903.10972*.
- [Yang et al., 2016] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.
- [Yimam et al., 2020] Yimam, S. M., Alemayehu, H. M., Ayele, A., and Biemann, C. (2020). Exploring amharic sentiment analysis from social media texts: Building annotation tools and classification models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1048–1060.
- [You et al., 2015] You, Q., Luo, J., Jin, H., and Yang, J. (2015). Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*, pages 381–388.

- [Yu et al., 2011] Yu, H.-F., Huang, F.-L., and Lin, C.-J. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1-2):41–75.
- [Yuan et al., 2015] Yuan, J., You, Q., and Luo, J. (2015). Sentiment analysis using social multimedia. In *Multimedia Data Mining and Analytics*, pages 31–59. Springer.
- [Zainuddin et al., 2016a] Zainuddin, N., Selamat, A., and Ibrahim, R. (2016a). Improving twitter aspect-based sentiment analysis using hybrid approach. In *Asian Conference on Intelligent Information and Database Systems*, pages 151–160. Springer.
- [Zainuddin et al., 2016b] Zainuddin, N., Selamat, A., and Ibrahim, R. (2016b). Twitter feature selection and classification using support vector machine for aspect-based sentiment analysis. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 269–279. Springer.
- [Zainuddin et al., 2018] Zainuddin, N., Selamat, A., and Ibrahim, R. (2018). Hybrid sentiment classification on twitter aspect-based sentiment analysis. *Applied Intelligence*, 48(5):1218–1232.
- [Zhang et al., 2012] Zhang, L., Jia, Y., Zhou, B., and Han, Y. (2012). Microblogging sentiment analysis using emotional vector. In *Cloud and Green Computing (CGC), 2012 Second International Conference on*, pages 430–433. IEEE.
- [Zhang et al., 2013] Zhang, L., Pei, S., Deng, L., Han, Y., Zhao, J., and Hong, F. (2013). Microblog sentiment analysis based on emoticon networks model. In *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*, pages 134–138. ACM.
- [Zhang et al., 2017] Zhang, S., Zhang, X., and Chan, J. (2017). A word-character convolutional neural network for language-agnostic twitter sentiment analysis. In *Proceedings of the 22nd Australasian Document Computing Symposium*, page 12. ACM.
- [Zhang et al., 2015] Zhang, Y., Shang, L., and Jia, X. (2015). Sentiment analysis on microblogging by integrating text and image features. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 52–63. Springer.

- [Zhao et al., 2019] Zhao, Y., Xu, X., and Wang, M. (2019). Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. *International Journal of Hospitality Management*, 76:111–121.
- [Zhou et al., 2017] Zhou, X., Tao, X., Rahman, M. M., and Zhang, J. (2017). Coupling topic modelling in opinion mining for social media analysis. In *Proceedings of the International Conference on Web Intelligence*, pages 533–540. ACM.
- [Zhou et al., 2014] Zhou, Z., Zhang, X., and Sanderson, M. (2014). Sentiment analysis on twitter through topic-based lexicon expansion. In *Australasian Database Conference*, pages 98–109. Springer.
- [Zhu et al., 2013] Zhu, S., Xu, B., Zheng, D., and Zhao, T. (2013). Chinese microblog sentiment analysis based on semi-supervised learning. In *Semantic Web and Web Science*, pages 325–331. Springer.
- [Zhu and Ghahramani, 2002] Zhu, X. and Ghahramani, Z. (2002). Learning from labeled and unlabeled data with label propagation.
- [Zimbra et al., 2016] Zimbra, D., Ghiassi, M., and Lee, S. (2016). Brand-related twitter sentiment analysis using feature engineering and the dynamic architecture for artificial neural networks. In *System Sciences (HICSS), 2016 49th Hawaii International Conference on*, pages 1930–1938. IEEE.

Appendix A

Malta Government Budget

Dataset Annotation Guidelines

Malta Government Budget Annotation Guidelines

What is the goal of the project?

To annotate online posts (i.e. comments from newspaper articles and tweets from Twitter) written in Maltese and/or English about Malta's annual Government Budget for multiple social opinion dimensions, namely subjectivity/objectivity, sentiment polarity, emotion, sarcasm, and irony.

What is each tag called and how is it used?

Annotation Types

Each online post shall be annotated with the following 8 annotation types i.e., 5 different social opinion dimensions (*subjectivity, sentiment polarity, emotion, sarcasm, and irony*), and 3 others (*negation, off-topic and language*):

1. **Subjectivity:** binary value, with 1 referring to subjective posts and 0 to objective posts.

Subjectivity Analysis

Subjectivity

Determines whether a piece of text expresses an opinion, in terms of personal feelings or beliefs



Objectivity

Refers to sentences that express some factual information about the world



2. **Sentiment Polarity:** categorical value (3-levels) for the sentiment polarity of the online post (negative, neutral, positive).

Sentiment Analysis

Discovery of subjective expressions that describe people's sentiments, views or attitudes

toward entities, events, topics and their properties



POSITIVE



NEUTRAL

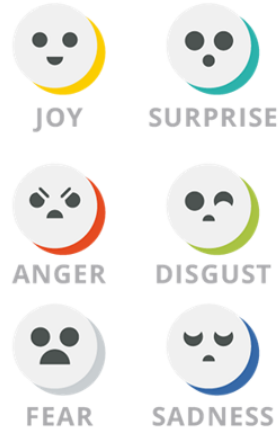


NEGATIVE

3. **Emotion:** categorical value for the emotion of the online post based on Robert Plutchik's 8 basic emotions (joy, trust, fear, surprise, sadness, anticipation, anger, and disgust)¹.

Emotion Analysis

Identification of human emotion from facial expressions or expressed subjective feelings and thoughts, such as anger, fear, sadness, disgust and surprise



4. **Sarcasm:** binary value, with 1 referring to sarcasm in online posts.
5. **Irony:** binary value, with 1 referring to irony in online posts.

Sarcasm and Irony Detection

Irony

Usually used to convey, the opposite meaning of the actual things you say, but its purpose is not intended to hurt the other person

Sarcasm

Holds the “characteristic” of meaning the opposite of what you say, but unlike irony, it is used to hurt the other person

6. **Negation:** binary value, with 1 referring to negated online posts. A negated post refers to the opposite of what is conveyed due to certain grammatical operations/terms such as ‘not’.
7. **Off-topic:** binary value, with 1 referring to off-topic online posts that are not related to the Government Budget, even if they are of a political nature.
8. **Language:** numerical value, with 0 referring to online posts in English, 1 referring to posts in Maltese, and 2 referring to Maltese-English (Maltenglish) code-switched posts, and 3 referring to any other language/code-switched posts. Note that '2' refers to an online post that contains switching (at least one normal/slang word) from one language to another (even multiple times) at a sentence/clause/phrase level.

Annotation Examples

Refer to the Malta Budget 2019 dataset template, which contains 5 online posts annotated from each data source i.e., Times of Malta, MaltaToday, The Malta Independent and Twitter.

¹ <https://www.6seconds.org/2020/08/11/plutchik-wheel-emotions>

Further Information

The following is further information in relation to the grey areas that the annotator may encounter when annotating the online posts:

- Retweets of **objective** posts are classified as still being **objective** in our case, even though they may reflect support to a particular budget measure.
- Retweets of **subjective** posts are classified as being **subjective**, that is, reflecting the opinion of the original post.
- If additional text is added to an objective retweet, this signifies an opinion on top of the original online post, hence is classified as being **subjective**.
- In the context of **objective** tweets, emotions are somewhat different than those for **subjective** tweets (which reflect the person's emotions) where *joy* is annotated to posts which shall bring a better quality of life and/or concrete support for certain budget measures.
- An online post having a *positive sentiment polarity* does not necessarily mean that the **emotion** reflected is a "*positive*" one i.e., *joy* or *trust*, since certain emotions reflect the user's (i.e. the user making the post) reaction to certain online posts (comments/tweets) made by other users. For example, online post with ID: 20190211 ("Then WHAT, exactly do you want?! Spending power and disposable income has been on the increase these last few years, so whatever business you're in, you MUST be better off; unless you are dealing with Stamperija Falluta, because then you have cause to be worried about your business/cash flow!") conveys a *positive sentiment polarity* and a *surprise emotion*.
- Certain online posts that convey **irony/sarcasm** may influence the respective **sentiment polarity** and/or **emotion**. However, this is not always the case. For example, online post with ID: 20190363 ("Min jaf x'jingibed mill inieher bhallissa - tant xoghol, tant srum barranin li ntlaghu buex isdir hsara lil Malta, u xorta johorgu figuri b'sahhithom hekk."), conveys a *positive sentiment polarity*, *joy emotion*, and **sarcasm**.
- It is important to note for certain special characters, such as an exclamation mark (!), and/or emoticons, such as 😊, since these might affect the overall online post value of a particular annotation type e.g., **sentiment polarity** or **emotion**.
- Any reference to other political events/occurrences (e.g., references to a particular opposition Minister or measure introduced in a previous legislature) should be annotated as being **off-topic**, unless a direct comparison with the existing government budget is made, in which case it would be classified as being **on-topic**.
- Any online post that refers directly and/or is related to a particular budget measure announced by the Government for the respective year, should be annotated as being **on-topic**. For example, online post with ID: 20190014 ("Stqsi lil dawk li int ukoll kont tircievi kontijiet tad-dawl u l-ilma gholjien, staqsi lil dawk il-pensonanti li tahtkom ghal 25 sema ma hadux zieda, staqsi lil dawk l-eluf li kienu jirregistraw ghax-xoghol tahtkom u lllum qed jaqighu x'jieklu, staqsi lil istudenti li lllum mhux qed ihallsu ghal ezamijiet u ghat-trasport... Nahseb li m'ghandikx mera id-dar.....") is **on-topic** due to the sentence before last, which refers to two budget measures i.e., "Free public transport extended to 14- and 15-year-olds and all full-time students aged 20 and over." and "All Matsec exam fees to be scrapped", even though the first part of the post is totally off-topic.
- If an online post is referring/mentioning a general subject/domain, such as, rent, electricity and/or water, and is not related to any budget measure, this should be annotated as being **off-topic**. For example, online post with ID: 20190206 ("u ejja nahseb qieghad tmurt bil guh self employed") is **off-topic** since this has nothing to do with a work related budget measure for self-employed i.e., "Self-employed forced to shut up shop will be eligible for unemployment benefit until they find work".
- In the case of **off-topic** online posts, the overall general context of the said post (composed of one or more sentences) should be taken in consideration when carrying out all the other

annotations i.e., subjectivity, sentiment polarity, emotion, sarcasm, irony, negation, and language.

- Online posts that are written in Maltese with the exception of one word e.g., budget, are still classified as being written in Maltese-English i.e., for the **language** annotation.
- If a link is provided in an online post, this should be ignored and not followed further to carry out the annotations. Therefore, all annotations should be based on the existing text provided in the online post, including any link text that might be relevant.
- If the name of an entity e.g., Person/Organisation within an online post is in a language that is different to the post's main language, this shall not be taken as being code-switched (2) or in another language (3), since it reflects the entity's original name. However, if the entity has both a Maltese and English name, the entity's written language is considered for the **language** annotation, e.g., Teatru Rjal (Maltese) vs. National Opera House (English).
- Should an online post contain multiple word/phrases that indicate **negation**, focus should only be given to those sentences/clauses/phrases that are **on-topic**, i.e., refers directly and/or is related to a particular budget measure announced by the Government for the respective year. Otherwise, the negations should be ignored.

What parts of the text do you want annotated, and what should be left alone?

Online posts from four different data sources have been collected, namely three Maltese newswires in the Times of Malta, MaltaToday, and The Malta Independent and one social networking service in Twitter. Each online post will be given a single value (label) for each of the 8 annotation types above. Note that one online post can be composed of more than one sentence. In such a case, the annotations should reflect the overall understanding of the entire online post by the annotator. For example:

- If an online post is composed of three sentences that signify a *negative sentiment polarity* in one sentence, and *positive sentiment polarity* in two sentences, the overall **sentiment polarity** should be *positive*.
- If an online post contains multiple sentences that convey different annotation representations e.g., experiencing both the *joy* and *surprise* emotions for the **emotion** annotation type, the primary **emotion** that is closest to the Government Budget topic (especially in cases where multiple topics are referenced in the online post), should be chosen.

How will the annotation be created?

A spreadsheet that is prepopulated with all the online posts per annual Government Budget shall be used by each annotator. The values (labels) for each annotation type shall be in the form of a drop-down list.

Other information

- The annotation types above are based on the first dataset that was annotated for the Malta Government Budget 2018. The annotation guidelines above were updated based on the feedback received from the annotators that carried out the task for this dataset and other observations. More information can be found here: <https://aclanthology.org/D19-5547.pdf>
- A summary of the Government Budget measures for each year should be primarily read by the annotator (i.e., before starting with the annotation process) to be familiar with the upcoming year's national measures. This shall help the annotator in understanding the opinions written by the general public and determine which ones are relevant, valid and/or otherwise. The following are the suggested articles for each year's Government Budget:
 - Budget 2018: <https://timesofmalta.com/articles/view/budget-2018-at-a-glance.660017>

- Budget 2019: <https://timesofmalta.com/articles/view/budget-2019-at-a-glance.692338>
- Budget 2020: <https://timesofmalta.com/articles/view/budget-2020-at-a-glance-whats-in-it-for-you.742272>
- The annotation guidelines shall be evaluated by each annotator and task owner after the first 25 online posts, and then after the first 100 online posts from each different data source are annotated, to check whether any clarifications are required to the guidelines. In such a case, the guidelines will be updated accordingly after a common understanding is reached.
- During the annotation process any interesting observations and feedback should be written down by the annotator, so that these will be discussed further with the task owner.

Appendix B

Full Evaluation Results for each Social Opinion Dimension using Machine Learning-based Models

Malta Budget 2018, 2019, 2020 Annual Government Budget Dataset [English - 4,732 online posts]

Subjectivity								
Classifier (word level TF-IDF)	Logistic Regression (LR)	Support Vector Machine (SVM) - LinearSVC	Support Vector Machine (SVM) - NuSVC	Support Vector Machine (SVM) - SVC	Bernoulli Naïve Bayes (BNB)	Complement Naïve Bayes (CNB)	Decision Trees (DT)	Random Forest (RF)
<i>Precision (weighted)</i>								
Mean	0.898985	0.892643	0.897768	0.407041	0.872429	0.898181	0.850321	0.886455
Standard deviation	0.072063	0.063519	0.072184	0.00097	0.060323	0.062577	0.071806	0.073218
Execution time (seconds)	0.19448066	0.123159	104.593976	130.738797	0.079820	0.039925	3.113672	24.685547
<i>Recall (weighted)</i>								
Mean	0.890517	0.884181	0.885863	0.637997	0.854592	0.888197	0.84403	0.879733
Standard deviation	0.080987	0.069411	0.086278	0.000761	0.079871	0.070466	0.066909	0.08201
Execution time (seconds)	0.19647622	0.118683	123.809086	135.010668	0.052841	0.032908	3.084726	25.071534
<i>F1 Score (weighted)</i>								
Mean	0.883841	0.879541	0.876273	0.496998	0.840135	0.883805	0.836732	0.8721
Standard deviation	0.090688	0.076603	0.099753	0.000954	0.100051	0.077748	0.083394	0.09332
Execution time (seconds)	0.21741724	0.113696	127.066246	131.721909	0.050866	0.031843	3.142780	30.266122
<i>Balanced Accuracy</i>								
Mean	0.866635	0.86671	0.855156	0.5	0.811524	0.873531	0.825418	0.85322
Standard deviation	0.109981	0.095431	0.118991	0	0.11106	0.09623	0.103128	0.11421
Execution time (seconds)	0.19448042	0.121662	130.349560	133.181255	0.057853	0.028923	3.182049	29.180988
<i>ROC AUC Score One-vs-Rest (weighted)</i>								
Mean	0.955468	NA	0.954053	0.940225	0.934616	0.954115	0.836103	0.951584
Standard deviation	0.04679	NA	0.048718	0.059692	0.052903	0.050101	0.097632	0.045572
Execution time (seconds)	0.23138189	NA	118.210097	135.354865	0.054814	0.032912	3.130797	29.914939
<i>ROC AUC Score One-vs-One (weighted)</i>								
Mean	0.955468	NA	0.954059	0.940235	0.934616	0.954115	0.834177	0.951524
Standard deviation	0.04679	NA	0.048714	0.059685	0.052903	0.050101	0.105521	0.043646
Execution time (seconds)	0.22839069	NA	122.366131	131.516467	0.054852	0.031915	3.147292	31.231760

Malta Budget 2018, 2019, 2020 Annual Government Budget Dataset [English - 4,732 online posts]

Sentiment Polarity

Classifier (word level TF-IDF)	Logistic Regression (LR)	Support Vector Machine (SVM) - LinearSVC	Support Vector Machine (SVM) - NuSVC	Support Vector Machine (SVM) - SVC	Bernoulli Naïve Bayes (BNB)	Complement Naive Bayes (CNB)	Decision Trees (DT)	Random Forest (RF)
<i>Precision (weighted)</i>								
Mean	0.795001	0.778484	0.798937	0.30236	0.795785	0.80954	0.73527	0.781681
Standard deviation	0.062405	0.052479	0.046434	0.000418	0.051838	0.060617	0.060647	0.065411
Execution time (seconds)	5.4145174	0.501660	170.696718	148.801829	0.058840	0.049865	5.265391	36.224598
<i>Recall (weighted)</i>								
Mean	0.782981	0.772412	0.78361	0.549873	0.777069	0.786152	0.729729	0.771151
Standard deviation	0.063678	0.052843	0.051936	0.00038	0.058514	0.072718	0.056989	0.060436
Execution time (seconds)	5.77855206	0.409940	163.333380	137.293424	0.053856	0.046876	5.272669	33.758583
<i>F1 Score (weighted)</i>								
Mean	0.773488	0.766855	0.777319	0.390174	0.776828	0.783019	0.730343	0.762214
Standard deviation	0.070612	0.054157	0.053882	0.000444	0.053359	0.073829	0.056692	0.067856
Execution time (seconds)	5.46837783	0.420839	160.116101	135.064040	0.057847	0.043883	5.935525	33.789001
<i>Balanced Accuracy</i>								
Mean	0.722771	0.717044	0.739063	0.333333	0.727495	0.766624	0.683661	0.712355
Standard deviation	0.077534	0.058836	0.059551	0	0.056173	0.075009	0.059166	0.075032
Execution time (seconds)	5.37862086	0.408905	164.670662	141.175434	0.073802	0.054852	5.606422	35.492793
<i>ROC AUC Score One-vs-Rest (weighted)</i>								
Mean	0.912313	NA	0.90798	0.844436	0.883816	0.902395	0.773283	0.914512
Standard deviation	0.038343	NA	0.03353	0.075816	0.044054	0.042505	0.038441	0.032395
Execution time (seconds)	5.66984057	NA	164.010988	134.826475	0.081781	0.060838	5.571950	36.478816
<i>ROC AUC Score One-vs-One (weighted)</i>								
Mean	0.910059	NA	0.906579	0.855395	0.880396	0.9027	0.760329	0.905257
Standard deviation	0.041414	NA	0.035748	0.075942	0.040654	0.041147	0.044179	0.039628
Execution time (seconds)	5.47237062	NA	167.190399	145.875279	0.090758	0.073803	5.700552	35.290663

Malta Budget 2018, 2019, 2020 Annual Government Budget Dataset [English - 4,732 online posts]

<i>Emotion</i>								
Classifier (word level TF-IDF)	Logistic Regression (LR)	Support Vector Machine (SVM) - LinearSVC	Support Vector Machine (SVM) - NuSVC	Support Vector Machine (SVM) - SVC	Bernoulli Naïve Bayes (BNB)	Complement Naive Bayes (CNB)	Decision Trees (DT)	Random Forest (RF)
<i>Precision (weighted)</i>								
Mean	0.548196	0.566701	0.560293	0.174376	0.550509	0.621616	0.532135	0.535862
Standard deviation	0.039689	0.039983	0.033416	0.000792	0.048384	0.05453	0.040589	0.047694
Execution time (seconds)	10.3918054	0.845738	254.507192	143.336993	0.088849	0.084555	6.113413	46.290185
<i>Recall (weighted)</i>								
Mean	0.604407	0.595956	0.600175	0.417582	0.59849	0.604845	0.542484	0.587076
Standard deviation	0.032677	0.05162	0.040466	0.000948	0.051026	0.066484	0.048064	0.044735
Execution time (seconds)	10.1568472	0.767979	255.186209	154.663881	0.069727	0.068815	7.023755	46.964180
<i>F1 Score (weighted)</i>								
Mean	0.558523	0.573032	0.565908	0.246018	0.559174	0.597985	0.529232	0.538796
Standard deviation	0.028066	0.04086	0.032799	0.000952	0.050814	0.059299	0.0468	0.044317
Execution time (seconds)	11.186094	0.751996	258.918981	154.321250	0.080784	0.071808	6.897779	44.708179
<i>Balanced Accuracy</i>								
Mean	0.247898	0.282854	0.268255	0.125	0.248973	0.319283	0.273307	0.227477
Standard deviation	0.023119	0.025369	0.025894	0	0.034061	0.035876	0.028211	0.026811
Execution time (seconds)	11.4114909	0.756940	256.932391	137.393610	0.071808	0.055851	5.900373	45.253153
<i>ROC AUC Score One-vs-Rest (weighted)</i>								
Mean	0.853009	NA	0.843479	0.808009	0.809012	0.842161	0.672194	0.833134
Standard deviation	0.040707	NA	0.043098	0.058414	0.037702	0.046722	0.038364	0.044798
Execution time (seconds)	11.16116	NA	255.994479	176.364910	0.118683	0.113696	7.099911	46.124341
<i>ROC AUC Score One-vs-One (weighted)</i>								
Mean	0.846007	NA	0.84232	0.826195	0.778992	0.832531	0.631816	0.809016
Standard deviation	0.032956	NA	0.033994	0.03558	0.031803	0.036357	0.022187	0.039032
Execution time (seconds)	11.4214673	NA	250.715391	179.370892	0.404145	0.350278	6.648542	46.428714

Malta Budget 2018, 2019, 2020 Annual Government Budget Dataset [English - 4,732 online posts]

Irony

Classifier (word level TF-IDF)	Logistic Regression (LR)	Support Vector Machine (SVM) - LinearSVC	Support Vector Machine (SVM) - NuSVC	Support Vector Machine (SVM) - SVC	Bernoulli Naïve Bayes (BNB)	Complement Naive Bayes (CNB)	Decision Trees (DT)	Random Forest (RF)
<i>Precision (weighted)</i>								
Mean	0.921714	0.927716	0.921682	0.921714	0.942605	0.93525	0.940631	0.925926
Standard deviation	0.00121	0.012979	0.001195	0.00121	0.021765	0.007726	0.016209	0.012548
Execution time (seconds)	0.29027605	0.210649	109.151075	19.827127	0.058806	0.045687	2.646639	27.191777
<i>Recall (weighted)</i>								
Mean	0.960059	0.956679	0.959214	0.960059	0.906602	0.923291	0.936398	0.960271
Standard deviation	0.000629	0.006348	0.001348	0.000629	0.0782	0.040272	0.031942	0.000838
Execution time (seconds)	0.27635837	0.191393	119.175861	19.559973	0.049901	0.040920	2.430729	28.119870
<i>F1 Score (weighted)</i>								
Mean	0.940496	0.940348	0.940073	0.940496	0.917422	0.92766	0.933126	0.940999
Standard deviation	0.000932	0.003619	0.000979	0.000932	0.046925	0.023662	0.019436	0.001675
Execution time (seconds)	0.27877164	0.173042	119.959047	17.620518	0.049693	0.046877	2.391442	26.070671
<i>Balanced Accuracy</i>								
Mean	0.5	0.510847	0.49956	0.5	0.535921	0.561683	0.560174	0.5
Standard deviation	0	0.019195	0.00073	0	0.073153	0.038838	0.040204	0
Execution time (seconds)	0.28799415	0.183587	116.872080	17.797195	0.052840	0.038994	2.428777	25.013881
<i>ROC AUC Score One-vs-Rest (weighted)</i>								
Mean	0.884479	NA	0.85741	0.727756	0.799145	0.774789	0.561589	0.841656
Standard deviation	0.107637	NA	0.075926	0.057515	0.162267	0.091664	0.065727	0.086882
Execution time (seconds)	0.26599312	NA	117.358007	14.447102	0.054545	0.051324	2.261174	22.918039
<i>ROC AUC Score One-vs-One (weighted)</i>								
Mean	0.884479	NA	0.85741	0.727698	0.799145	0.774789	0.557476	0.847413
Standard deviation	0.107637	NA	0.075911	0.05741	0.162267	0.091664	0.059009	0.08669
Execution time (seconds)	0.26377821	NA	109.193628	13.356499	0.049637	0.047904	2.570560	26.821255

Malta Budget 2018, 2019, 2020 Annual Government Budget Dataset [English - 4,732 online posts]

Sarcasm								
Classifier (word level TF-IDF)	Logistic Regression (LR)	Support Vector Machine (SVM) - LinearSVC	Support Vector Machine (SVM) - NuSVC	Support Vector Machine (SVM) - SVC	Bernoulli Naïve Bayes (BNB)	Complement Naive Bayes (CNB)	Decision Trees (DT)	Random Forest (RF)
<i>Precision (weighted)</i>								
Mean	0.957769	0.960873	0.961544	0.957769	0.971384	0.964114	0.966282	0.960081
Standard deviation	0.001215	0.006431	0.007027	0.001215	0.011126	0.007928	0.006366	0.007199
Execution time (seconds)	0.21846509	0.130650	54.466874	9.446687	0.067820	0.051862	2.954386	21.712917
<i>Recall (weighted)</i>								
Mean	0.978656	0.976966	0.978867	0.978656	0.95034	0.945691	0.966189	0.978656
Standard deviation	0.000621	0.003051	0.000018	0.000621	0.041363	0.036439	0.016034	0.000621
Execution time (seconds)	0.15210295	0.104263	59.422618	9.365024	0.053854	0.051411	2.892107	20.147659
<i>F1 Score (weighted)</i>								
Mean	0.9681	0.967914	0.968939	0.9681	0.956555	0.954114	0.962334	0.968481
Standard deviation	0.000925	0.001536	0.001145	0.000925	0.023294	0.021283	0.00797	0.001867
Execution time (seconds)	0.20296693	0.100734	61.085205	9.225725	0.058864	0.047869	2.910279	21.632849
<i>Balanced Accuracy</i>								
Mean	0.5	0.508466	0.509438	0.5	0.558462	0.566324	0.539809	0.5
Standard deviation	0	0.018602	0.018891	0	0.055718	0.072913	0.043947	0
Execution time (seconds)	0.16256809	0.086801	60.195354	9.083780	0.073783	0.047872	3.097657	21.620252
<i>ROC AUC Score One-vs-Rest (weighted)</i>								
Mean	0.912198	NA	0.880103	0.732216	0.871998	0.789496	0.529917	0.797657
Standard deviation	0.119032	NA	0.10574	0.103001	0.128511	0.116001	0.03557	0.122636
Execution time (seconds)	0.16458941	NA	61.262658	9.305136	0.065821	0.051895	3.004213	21.408348
<i>ROC AUC Score One-vs-One (weighted)</i>								
Mean	0.912198	NA	0.880168	0.732173	0.871998	0.789496	0.525456	0.773359
Standard deviation	0.119032	NA	0.105727	0.103051	0.128511	0.116001	0.038351	0.134297
Execution time (seconds)	0.1964426	NA	61.603465	9.020720	0.059840	0.052858	3.260226	21.209943

Malta Budget 2018, 2019, 2020 Annual Government Budget Dataset [Maltese-English and Maltese - 1,596 online posts]

Subjectivity								
Classifier (word level TF-IDF)	Logistic Regression (LR)	Support Vector Machine (SVM) - LinearSVC	Support Vector Machine (SVM) - NuSVC	Support Vector Machine (SVM) - SVC	Bernoulli Naïve Bayes (BNB)	Complement Naïve Bayes (CNB)	Decision Trees (DT)	Random Forest (RF)
<i>Precision (weighted)</i>								
Mean	0.862451	0.861485	0.866554	0.284987	0.857786	0.886867	0.85643	0.891391
Standard deviation	0.074381	0.079464	0.07402	0.002525	0.054658	0.062724	0.091902	0.075283
Execution time (seconds)	0.129895	0.084772	20.668743	25.094199	0.091816	0.059177	1.20102763	10.547154
<i>Recall (weighted)</i>								
Mean	0.843911	0.843848	0.848903	0.533836	0.7907	0.859611	0.828805	0.850142
Standard deviation	0.095566	0.089469	0.09074	0.002364	0.122041	0.09738	0.118021	0.12241
Execution time (seconds)	0.089760	0.063829	21.443317	25.180806	0.057847	0.051864	1.25270867	10.687201
<i>F1 Score (weighted)</i>								
Mean	0.839627	0.841091	0.845513	0.371596	0.772777	0.854936	0.840003	0.847637
Standard deviation	0.103584	0.092145	0.096013	0.002719	0.15322	0.105658	0.101674	0.144409
Execution time (seconds)	0.104720	0.070811	22.468027	25.430298	0.062830	0.054854	1.33644176	9.726062
<i>Balanced Accuracy</i>								
Mean	0.843608	0.842783	0.847955	0.5	0.802879	0.864388	0.837593	0.85449
Standard deviation	0.088045	0.08498	0.084147	0	0.114555	0.09073	0.101996	0.117348
Execution time (seconds)	0.104722	0.067819	23.194986	26.418688	0.064828	0.045877	1.30309176	10.348607
<i>ROC AUC Score One-vs-Rest (weighted)</i>								
Mean	0.92519	NA	0.923772	0.904334	0.92877	0.925488	0.827352	0.931932
Standard deviation	0.076354	NA	0.074861	0.101011	0.063919	0.065249	0.110385	0.066796
Execution time (seconds)	0.109705	NA	18.551733	25.642880	0.064828	0.057846	1.27232814	10.702531
<i>ROC AUC Score One-vs-One (weighted)</i>								
Mean	0.92519	NA	0.923772	0.904286	0.92877	0.925488	0.829902	0.930892
Standard deviation	0.076354	NA	0.074861	0.101045	0.063919	0.065249	0.111208	0.068315
Execution time (seconds)	0.096741	NA	20.245687	24.869214	0.060837	0.052859	1.32327628	9.559262

Malta Budget 2018, 2019, 2020 Annual Government Budget Dataset [Maltese-English and Maltese - 1,596 online posts]

Sentiment Polarity

Classifier (word level TF-IDF)	Logistic Regression (LR)	Support Vector Machine (SVM) - LinearSVC	Support Vector Machine (SVM) - NuSVC	Support Vector Machine (SVM) - SVC	Bernoulli Naïve Bayes (BNB)	Complement Naïve Bayes (CNB)	Decision Trees (DT)	Random Forest (RF)
<i>Precision (weighted)</i>								
Mean	0.712071	0.754208	0.750617	0.34394	0.597098	0.733453	0.737661	0.744739
Standard deviation	0.084666	0.08806	0.098858	0.00149	0.069804	0.092582	0.065601	0.080513
Execution time (seconds)	2.516273	0.213429	22.208073	16.003161	0.048542	0.037902	1.42250943	14.954082
<i>Recall (weighted)</i>								
Mean	0.733215	0.759524	0.752044	0.586462	0.65862	0.745079	0.72818	0.760177
Standard deviation	0.083641	0.091074	0.102141	0.001271	0.062376	0.102146	0.088846	0.098198
Execution time (seconds)	2.375292	0.151594	22.071631	16.312656	0.041888	0.034874	1.39884591	14.118991
<i>F1 Score (weighted)</i>								
Mean	0.689206	0.739622	0.725397	0.433592	0.593306	0.724719	0.713753	0.727601
Standard deviation	0.081683	0.096397	0.106766	0.001532	0.060966	0.102363	0.088923	0.100209
Execution time (seconds)	2.689799	0.158576	21.958898	15.809725	0.040890	0.032912	1.38199282	13.871472
<i>Balanced Accuracy</i>								
Mean	0.562462	0.638019	0.618941	0.333333	0.449516	0.612975	0.600848	0.621904
Standard deviation	0.063686	0.091922	0.101573	0	0.049415	0.08929	0.08756	0.075925
Execution time (seconds)	3.114425	0.141652	21.857579	15.570223	0.038935	0.033907	1.35929084	13.851295
<i>ROC AUC Score One-vs-Rest (weighted)</i>								
Mean	0.861349	NA	0.860645	0.849768	0.816897	0.847137	0.761013	0.872381
Standard deviation	0.073484	NA	0.070865	0.066748	0.094059	0.073179	0.064646	0.080658
Execution time (seconds)	2.975994	NA	22.806798	15.565908	0.072767	0.050864	1.40495062	13.920745
<i>ROC AUC Score One-vs-One (weighted)</i>								
Mean	0.82469	NA	0.82771	0.815131	0.76629	0.805232	0.725014	0.843829
Standard deviation	0.078794	NA	0.081344	0.07605	0.093455	0.071976	0.05989	0.081538
Execution time (seconds)	2.549638	NA	22.049991	15.904867	0.067820	0.067419	1.44623232	14.052663

Malta Budget 2018, 2019, 2020 Annual Government Budget Dataset [Maltese-English and Maltese - 1,596 online posts]

Emotion								
Classifier (word level TF-IDF)	Logistic Regression (LR)	Support Vector Machine (SVM) - LinearSVC	Support Vector Machine (SVM) - NuSVC	Support Vector Machine (SVM) - SVC	Bernoulli Naïve Bayes (BNB)	Complement Naive Bayes (CNB)	Decision Trees (DT)	Random Forest (RF)
<i>Precision (weighted)</i>								
Mean	0.387068	0.431473	0.377562	0.164858	0.305901	0.459716	0.39745	0.427607
Standard deviation	0.066252	0.061812	0.074974	0.002172	0.054344	0.069449	0.058036	0.074333
Execution time (seconds)	7.551333	0.449753	32.588877	20.594411	0.064750	0.047839	1.819197	16.964402
<i>Recall (weighted)</i>								
Mean	0.465527	0.462366	0.436057	0.406018	0.437999	0.478094	0.414151	0.491219
Standard deviation	0.045378	0.066458	0.046885	0.002684	0.024645	0.065296	0.066066	0.069835
Execution time (seconds)	10.347057	0.401165	31.269639	20.561232	0.056847	0.044879	1.825574	16.861066
<i>F1 Score (weighted)</i>								
Mean	0.376882	0.427224	0.375519	0.234498	0.314026	0.432851	0.402914	0.422871
Standard deviation	0.034389	0.054136	0.049605	0.002648	0.035831	0.047303	0.058992	0.062922
Execution time (seconds)	9.159186	0.386930	32.618278	21.102636	0.054854	0.042886	1.776890	16.799053
<i>Balanced Accuracy</i>								
Mean	0.205188	0.275239	0.241458	0.125	0.162991	0.254581	0.234674	0.264129
Standard deviation	0.022074	0.062415	0.05161	0	0.02325	0.031079	0.051085	0.052787
Execution time (seconds)	10.099896	0.438826	31.778672	21.219770	0.063835	0.040889	1.792011	16.790550
<i>ROC AUC Score One-vs-Rest (weighted)</i>								
Mean	0.767849	NA	0.762188	0.76848	0.708964	0.76157	0.612182	0.760453
Standard deviation	0.073249	NA	0.056033	0.072421	0.05639	0.063002	0.043863	0.071702
Execution time (seconds)	8.023407	NA	31.812180	20.873440	0.099729	0.096740	1.838538	16.673572
<i>ROC AUC Score One-vs-One (weighted)</i>								
Mean	0.773025	NA	0.786097	0.773825	0.69129	0.754403	0.592553	0.765885
Standard deviation	0.056762	NA	0.051787	0.059956	0.040458	0.04216	0.026722	0.057295
Execution time (seconds)	8.785131	NA	31.889935	20.980258	0.338132	0.322177	2.040579	16.954129

Malta Budget 2018, 2019, 2020 Annual Government Budget Dataset [Maltese-English and Maltese - 1,596 online posts]

Irony

Classifier (word level TF-IDF)	Logistic Regression (LR)	Support Vector Machine (SVM) - LinearSVC	Support Vector Machine (SVM) - NuSVC	Support Vector Machine (SVM) - SVC	Bernoulli Naïve Bayes (BNB)	Complement Naive Bayes (CNB)	Decision Trees (DT)	Random Forest (RF)
<i>Precision (weighted)</i>								
Mean	0.836859	0.86909	0.836656	0.836859	0.852891	0.89866	0.873688	0.836809
Standard deviation	0.005147	0.040563	0.005331	0.005147	0.029189	0.031776	0.028173	0.00519
Execution time (seconds)	0.11469173	0.0967376	16.61064672	3.92442131	0.05186105	0.075795412	1.66899276	13.56156516
<i>Recall (weighted)</i>								
Mean	0.914796	0.913546	0.912296	0.914796	0.86783	0.886018	0.882237	0.914171
Standard deviation	0.002812	0.008177	0.007225	0.002812	0.056065	0.061891	0.061374	0.00378
Execution time (seconds)	0.11070323	0.0837767	16.64851737	3.819279432	0.04687452	0.07679534	1.44721699	13.34571409
<i>F1 Score (weighted)</i>								
Mean	0.874091	0.87929	0.872834	0.874091	0.855613	0.884021	0.873438	0.874916
Standard deviation	0.00409	0.009558	0.005784	0.00409	0.031979	0.042444	0.0393	0.005964
Execution time (seconds)	0.1171968	0.0907581	14.78843832	4.01030612	0.04587793	0.081782103	1.40874577	13.67780948
<i>Balanced Accuracy</i>								
Mean	0.5	0.518964	0.49863	0.5	0.507704	0.611068	0.563341	0.503229
Standard deviation	0	0.030468	0.003139	0	0.034832	0.057851	0.073932	0.010876
Execution time (seconds)	0.1059804	0.0817819	15.39215565	3.994776249	0.04388213	0.056846857	1.3636682	14.00300074
<i>ROC AUC Score One-vs-Rest (weighted)</i>								
Mean	0.859442	NA	0.847283	0.742453	0.721696	0.813684	0.567714	0.796062
Standard deviation	0.13529	NA	0.109094	0.096753	0.170878	0.121596	0.065654	0.144008
Execution time (seconds)	0.08976007	NA	17.20572114	3.929448128	0.09474683	0.072317362	1.36001778	13.38377023
<i>ROC AUC Score One-vs-One (weighted)</i>								
Mean	0.859442	NA	0.847388	0.74231	0.721696	0.813684	0.57439	0.777627
Standard deviation	0.13529	NA	0.109125	0.096519	0.170878	0.121596	0.053828	0.158774
Execution time (seconds)	0.07978654	NA	17.28489923	4.128341436	0.08976078	0.063830137	1.35951686	13.25396919

Malta Budget 2018, 2019, 2020 Annual Government Budget Dataset [Maltese-English and Maltese - 1,596 online posts]

Sarcasm								
Classifier (word level TF-IDF)	Logistic Regression (LR)	Support Vector Machine (SVM) - LinearSVC	Support Vector Machine (SVM) - NuSVC	Support Vector Machine (SVM) - SVC	Bernoulli Naïve Bayes (BNB)	Complement Naïve Bayes (CNB)	Decision Trees (DT)	Random Forest (RF)
<i>Precision (weighted)</i>								
Mean	0.909438	0.909316	0.909408	0.909438	0.919784	0.922637	0.911011	0.909408
Standard deviation	0.005668	0.005827	0.005705	0.005668	0.021538	0.023581	0.007774	0.005705
Execution time (seconds)	0.151207	0.088764	13.917876	2.559507	0.057128	0.040259	1.037858	9.694102
<i>Recall (weighted)</i>								
Mean	0.95364	0.95114	0.953015	0.95364	0.929823	0.901034	0.927362	0.953015
Standard deviation	0.002974	0.009126	0.004112	0.002974	0.027158	0.056516	0.036635	0.004112
Execution time (seconds)	0.115200	0.104719	14.255334	3.196795	0.043569	0.038864	1.070471	9.530663
<i>F1 Score (weighted)</i>								
Mean	0.931012	0.929747	0.930699	0.931012	0.921376	0.9097	0.916654	0.930699
Standard deviation	0.004388	0.006875	0.004848	0.004388	0.016343	0.036802	0.023696	0.004848
Execution time (seconds)	0.092263	0.061836	16.349689	3.304753	0.043839	0.039893	0.995368	9.581498
<i>Balanced Accuracy</i>								
Mean	0.5	0.498684	0.499671	0.5	0.507068	0.530167	0.50345	0.499671
Standard deviation	0	0.003947	0.000987	0	0.035031	0.087631	0.032568	0.000987
Execution time (seconds)	0.099243	0.055852	15.832864	3.323404	0.048871	0.045880	0.974012	10.244848
<i>ROC AUC Score One-vs-Rest (weighted)</i>								
Mean	0.797687	NA	0.75976	0.656775	0.754832	0.727621	0.494897	0.668778
Standard deviation	0.230739	NA	0.187019	0.113749	0.15093	0.223064	0.027353	0.120771
Execution time (seconds)	0.122673	NA	14.185053	2.311158	0.056881	0.038895	0.898229	9.697888
<i>ROC AUC Score One-vs-One (weighted)</i>								
Mean	0.797687	NA	0.75976	0.657057	0.754832	0.727621	0.497296	0.698804
Standard deviation	0.230739	NA	0.187019	0.114051	0.15093	0.223064	0.028828	0.140094
Execution time (seconds)	0.094748	NA	15.517448	2.199332	0.063794	0.047872	0.922534	9.395439

Appendix C

Full Evaluation Results for each Social Opinion Dimension using Neural-based Models

Subjectivity

Model	Max Length	Learning rate	Epsilon value	Train set amount (70%)	Validation set amount (20%)	Test set amount (10%)
BERT	128	0.00002	0.00000001	4470	1277	640
DistilBERT	128	0.00002	0.00000001	4470	1277	640
BERTweet (base)	128	0.00002	0.00000001	4470	1277	640
BERTweet (covid19)	128	0.00002	0.00000001	4470	1277	640

Subjectivity

Model	Epoch	Training Loss	Validation Loss	Validation Accuracy	Training Time	Validation Time
BERT	1	0.34	0.22	92.27	12.53	100.19
	2	0.18	0.17	94.14	13.14	106.32
	3	0.12	0.17	94.6	13.08	107.39
	4	0.08	0.2	95	13.07	107.14
DistilBERT	1	0.36	0.22	91.87	3.65	29.83
	2	0.19	0.18	93.35	3.66	29.66
	3	0.14	0.19	93.51	3.64	29.65
	4	0.11	0.18	94.52	3.65	29.63
BERTweet (base)	1	0.41	0.27	91	12.55	99.38
	2	0.25	0.21	93.12	13.2	105.77
	3	0.2	0.2	94.13	13.21	107.37
	4	0.16	0.2	94.61	13.16	107.49
BERTweet (covid19)	1	0.41	0.24	92.58	22.99	188.48
	2	0.25	0.22	93.2	22.93	188.57
	3	0.2	0.19	94.14	22.92	188.47
	4	0.15	0.2	94.14	22.91	188.16

Subjectivity

Model	Threshold		precision	recall	f1-score	support	AUC	True Negative Rate (Specificity)	True Positive Rate (Sensitivity)	Total MCC
BERT	0.49	0	0.94	0.95	0.95	380	0.983	0.95	0.911538462	0.864
		1	0.93	0.91	0.91	260				
		accuracy			0.93	640				
		macro avg	0.93	0.93	0.93	640				
		weighted avg	0.93	0.93	0.93	640				
DistilBERT	0.49	0	0.92	0.96	0.94	380	0.9799	0.957894737	0.884615385	0.851
		1	0.93	0.88	0.91	260				
		accuracy			0.93	640				
		macro avg	0.93	0.92	0.92	640				
		weighted avg	0.93	0.93	0.93	640				
BERTweet (base)	0.49	0	0.92	0.97	0.94	380	0.9697	0.965789474	0.880769231	0.857
		1	0.95	0.88	0.91	260				
		accuracy			0.93	640				
		macro avg	0.93	0.92	0.93	640				
		weighted avg	0.93	0.93	0.93	640				
BERTweet (covid19)	0.49	0	0.93	0.98	0.96	380	0.9748	0.978947368	0.896153846	0.887
		1	0.97	0.9	0.93	260				
		accuracy			0.95	640				
		macro avg	0.95	0.94	0.94	640				
		weighted avg	0.95	0.95	0.94	640				

Sentiment Polarity

Model	Max Length	Learning rate	Epsilon value	Train set amount (70%)	Validation set amount (20%)	Test set amount (10%)
BERT	128	0.00002	0.00000001	4470	1277	640
DistilBERT	128	0.00002	0.00000001	4470	1277	640
BERTweet (base)	128	0.00002	0.00000001	4470	1277	640
BERTweet (covid19)	128	0.00002	0.00000001	4470	1277	640

Sentiment Polarity

Model	Epoch	Training Loss	Validation Loss	Validation Accuracy	Training Time	Validation Time
BERT	1	0.7	0.5	82.15	14.08	114.11
	2	0.49	0.45	82.54	14.21	115.77
	3	0.39	0.43	84.57	14.2	115.79
	4	0.33	0.43	84.25	14.24	115.91
DistilBERT	1	0.72	0.5	82.23	6.15	49.63
	2	0.5	0.45	83.26	6.46	52.07
	3	0.42	0.42	84.18	6.6	53.64
	4	0.37	0.41	84.73	6.53	53.6
BERTweet (base)	1	0.78	0.57	81.36	13.36	103.62
	2	0.57	0.48	85.58	13.12	106.2
	3	0.47	0.45	86.3	13.11	106.64
	4	0.43	0.44	86.14	13.09	106.6
BERTweet (covid19)	1	0.78	0.55	82.39	13.01	105.12
	2	0.56	0.47	85.9	13.1	106.49
	3	0.46	0.44	86.21	13.1	106.52
	4	0.41		86.99	13.11	106.75

Sentiment Polarity

Model		precision	recall	f1-score	support	AUC	Total MCC
BERT	0	0.71	0.8	0.76	123	0.9578	0.748
	1	0.83	0.77	0.8	161	0.9451	
	2	0.91	0.9	0.91	356	0.9611	
	accuracy			0.85	640		
	macro avg	0.82	0.83	0.82	640		
	weighted avg	0.85	0.85	0.85	640		
DistilBERT	0	0.7	0.79	0.74	123	0.9514	0.71
	1	0.78	0.75	0.76	161	0.9309	
	2	0.9	0.88	0.89	356	0.9491	
	accuracy			0.83	640		
	macro avg	0.79	0.8	0.8	640		
	weighted avg	0.83	0.83	0.83	640		
BERTweet (base)	0	0.75	0.85	0.79	123	0.9614	0.772
	1	0.83	0.77	0.8	161	0.9308	
	2	0.92	0.91	0.92	356	0.9578	
	accuracy			0.86	640		
	macro avg	0.83	0.84	0.84	640		
	weighted avg	0.87	0.86	0.86	640		
BERTweet (covid19)	0	0.77	0.85	0.81	123	0.9534	0.781
	1	0.86	0.77	0.81	161	0.934	
	2	0.91	0.92	0.92	356	0.9638	
	accuracy			0.87	640		
	macro avg	0.85	0.85	0.85	640		
	weighted avg	0.87	0.87	0.87	640		

Emotion

Model	Max Length	Learning rate	Epsilon value	Train set amount (70%)	Validation set amount (20%)	Test set amount (10%)
BERT	128	0.00002	0.00000001	4470	1277	640
DistilBERT	128	0.00002	0.00000001	4470	1277	640

Emotion

Model	Epoch	Training Loss	Validation Loss	Validation Accuracy	Training Time	Validation Time
BERT	1	1.38	1.17	57.86	13.83	113
	2	1.13	1.01	64.52	13.61	111.77
	3	1.01	0.97	65.53	13.57	111.71
	4	0.95	0.96	66.72	13.55	111.47
DistilBERT	1	1.41	1.14	60.36	6.86	56.68
	2	1.15	1.06	61.93	6.85	55.34
	3	1.05	1	65.62	6.72	55.77
	4	0.97	0.98	66.09	6.81	56.06

Emotion

Model		precision	recall	f1-score	support	AUC	Total MCC
BERT	0	0.74	0.87	0.8	264	0.9026	0.495
	1	0.44	0.11	0.18	36	0.8171	
	2	1	0	0	7	0.8	
	3	1	0	0	18	0.9353	
	4	1	0	0	25	0.8474	
	5	0.31	0.74	0.44	50	0.9135	
	6	0.14	0.03	0.05	37	0.8938	
	7	0.73	0.7	0.72	203	0.8663	
	accuracy			0.65	640		
	macro avg	0.67	0.31	0.27	640		
weighted avg	0.67	0.65	0.6	640			
DistilBERT	0	0.74	0.87	0.8	264	0.9131	0.484
	1	0.25	0.03	0.05	36	0.821	
	2	1	0	0	7	0.8188	
	3	1	0	0	18	0.9346	
	4	1	0	0	25	0.8317	
	5	0.32	60	0.41	50	0.9038	
	6	0.2	0.16	0.18	37	0.8925	
	7	0.72	0.71	0.72	203	0.8817	
	accuracy			0.64	640		
	macro avg	0.65	0.3	0.27	640		
weighted avg	0.66	0.64	0.6	640			

Emotion

Model	Max Length	Learning rate	Epsilon value	Train set amount (70%)	Validation set amount (20%)	Test set amount (10%)
BERTweet (base)	128	0.00002	0.00000001	4470	1277	640
BERTweet (covid19)	128	0.00002	0.00000001	4470	1277	640

Emotion

Model	Epoch	Training Loss	Validation Loss	Validation Accuracy	Training Time	Validation Time
BERTweet (base)	1	1.62	1.38	58.65	12.97	107.11
	2	1.36	1.22	62.02	13.1	106.92
	3	1.24	1.15	64.69	13.17	106.8
	4	1.18	1.13	65.4	13.06	106.69
BERTweet (covid19)	1	1.62	1.36	58.65	12.92	105.52
	2	1.35	1.2	63.44	13.06	106.66
	3	1.22	1.14	65.64	13.15	106.65
	4	1.16	1.12	66.19	13.12	106.54

Emotion

Model		precision	recall	f1-score	support	AUC	Total MCC
BERTweet (base)	0	0.75	0.83	0.79	264	0.8935	0.478
	1	1	0	0	36	0.7068	
	2	1	0	0	7	0.8617	
	3	1	0	0	18	0.9184	
	4	1	0	0	25	0.8201	
	5	0.31	0	0.45	50	0.898	
	6	0	0	0	37	0.8771	
	7	0.67	0.72	0.70	203	0.8597	
	accuracy			0.63	640		
	macro avg	0.72	0.29	0.24	640		
weighted avg	0.68	0.63	0.58	640			
BERTweet (covid19)	0	0.77	0.84	0.8	264	0.8983	0.501
	1	1	0	0	36	0.7216	
	2	1	0	0	7	0.8438	
	3	1	0	0	18	0.921	
	4	1	0	0	25	0.8289	
	5	0.31	0.86	0.46	50	0.91	
	6	1	0	0	37	0.8932	
	7	0.7	0.73	0.72	203	0.8567	
	accuracy			0.65	640		
	macro avg	0.85	0.3	0.25	640		
weighted avg	0.76	0.65	0.59	640			

Irony

Model	Max Length	Learning rate	Epsilon value	Train set amount (70%)	Validation set amount (20%)	Test set amount (10%)
BERT	128	0.00002	0.00000001	4470	1277	640
DistilBERT	128	0.00002	0.00000001	4470	1277	640
BERTweet (base)	128	0.00002	0.00000001	4470	1277	640
BERTweet (covid19)	128	0.00002	0.00000001	4470	1277	640

Irony

Model	Epoch	Training Loss	Validation Loss	Validation Accuracy	Training Time	Validation Time
BERT	1	0.2	0.15	94.84	22.62	185.92
	2	0.15	0.14	94.84	22.63	186.2
	3	0.12	0.14	94.6	22.67	186.42
	4	0.09	0.18	94.52	22.66	186.5
DistilBERT	1	0.2	0.15	94.84	11.4	93.31
	2	0.15	0.14	94.84	11.44	93.89
	3	0.13	0.14	94.84	11.47	94.07
	4	0.11	0.15	94.91	11.42	94.1
BERTweet (base)	1	0.24	0.2	94.84	23.27	190.95
	2	0.19	0.18	94.84	23.26	191.39
	3	0.18	0.16	94.84	23.41	191.83
	4	0.16	0.17	94.84	23.41	192
BERTweet (covid19)	1	0.24	0.17	94.84	23.38	191.18
	2	0.18	0.17	94.84	23.31	191.88
	3	0.17	0.16	94.84	23.35	191.85
	4	0.16	0.16	94.84	23.42	192.12

Irony

Model	Threshold		precision	recall	f1-score	support	AUC	True Negative Rate (Specificity)	True Positive Rate (Sensitivity)	Total MCC
BERT	0.49	0	0.95	0.99	0.97	607	0.8831	0.993410214	0.090909091	0.179
		1	0.43	0.09	0.15	33				
		accuracy			0.95	640				
		macro avg	0.69	0.54	0.56	640				
		weighted avg	0.93	0.95	0.93	640				
DistilBERT	0.49	0	0.95	1	0.98	607	0.8957	1	0.060606061	0.24
		1	1	0.06	0.11	33				
		accuracy			0.95	640				
		macro avg	0.98	0.53	0.54	640				
		weighted avg	0.95	0.95	0.93	640				
BERTweet (base)	0.49	0	0.95	1	0.97	607	0.8623	1	0	0
		1	1	0	0	33				
		accuracy			0.95	640				
		macro avg	0.97	0.5	0.49	640				
		weighted avg	0.95	0.95	0.92	640				
BERTweet (covid19)	0.49	0	0.95	1	0.97	607	0.8869	1	0	0
		1	1	0	0	33				
		accuracy			0.95	640				
		macro avg	0.97	0.5	0.49	640				
		weighted avg	0.95	0.95	0.92	640				

Sarcasm

Model	Max Length	Learning rate	Epsilon value	Train set amount (70%)	Validation set amount (20%)	Test set amount (10%)
BERT	128	0.00002	0.00000001	4470	1277	640
DistilBERT	128	0.00002	0.00000001	4470	1277	640
BERTweet (base)	128	0.00002	0.00000001	4470	1277	640
BERTweet (covid19)	128	0.00002	0.00000001	4470	1277	640

Sarcasm

Model	Epoch	Training Loss	Validation Loss	Validation Accuracy	Training Time	Validation Time
BERT	1	0.14	0.09	97.26	22.61	185.62
	2	0.09	0.11	97.26	22.64	186.3
	3	0.06	0.13	96.87	22.65	186.32
	4	0.05	0.13	96.4	22.63	186.29
DistilBERT	1	0.15	0.09	97.26	11.59	94.87
	2	0.1	0.09	97.26	11.61	95.58
	3	0.08	0.1	97.18	11.61	95.56
	4	0.06	0.1	97.02	11.6	95.6
BERTweet (base)	1	0.19	0.13	97.26	23.29	191.01
	2	0.14	0.12	97.26	23.24	191.04
	3	0.13	0.11	97.26	23.31	191.57
	4	0.11	0.11	97.26	23.34	191.72
BERTweet (covid19)	1	0.19	0.13	97.26	23.24	190.87
	2	0.14	0.11	97.26	23.26	190.93
	3	0.13	0.1	97.26	23.23	191.01
	4	0.12	0.11	97.26	23.25	191.14

Sarcasm

Model	Threshold		precision	recall	f1-score	support	AUC	True Negative Rate (Specificity)	True Positive Rate (Sensitivity)	Total MCC
BERT	0.49	0	0.97	0.99	0.98	622	0.8583	0.990353698	0.055555556	0.073
		1	0.14	0.6	0.08	18				
		accuracy			0.96	640				
		macro avg	0.56	0.52	0.53	640				
		weighted avg	0.95	0.96	0.96	640				
DistilBERT	0.49	0	0.97	1	0.99	622	0.8787	0.998392283	0.111111111	0.265
		1	0.67	0.11	0.19	18				
		accuracy			0.97	640				
		macro avg	0.82	0.55	0.59	640				
		weighted avg	0.97	0.97	0.96	640				
BERTweet (base)	0.49	0	0.97	1	0.99	622	0.8727	1	0	0
		1	1	0	0	18				
		accuracy			0.97	640				
		macro avg	0.99	0.5	0.49	640				
		weighted avg	0.97	0.97	0.96	640				
BERTweet (covid19)	0.49	0	0.97	1	0.99	622	0.7922	1	0	0
		1	1	0	0	18				
		accuracy			0.97	640				
		macro avg	0.99	0.5	0.49	640				
		weighted avg	0.97	0.97	0.96	640				

Appendix D

DCU Research Ethics Committee Application



**Dublin City University
RESEARCH ETHICS COMMITTEE**

**APPLICATION FOR APPROVAL OF A PROJECT
INVOLVING HUMAN PARTICIPANTS**

Application No. (office use only) DCUREC/2021/_____

Please read the following information carefully before completing your application. Failure to adhere to these guidelines will make your submission ineligible for review.

- **Applications must be submitted via the Research Ethics Application Portal [here](#) – no hardcopy required. All queries relating to submission should be e-mailed to the DCU Research Ethics Committee (REC) at rec@dcu.ie**
- **Section 4 of this form addresses the possible data protection issues of the proposed research and it must be completed prior to making a formal REC application.**
- **Student applicants must include their supervisor as an investigator on the Research Ethics Application Portal – this applies to all masters by research and PhD students. The form should be checked, approved and signed by the supervisor in advance of submission to REC. ***NB – Taught Masters and Undergraduate students apply for ethical review via their local ethics review panel, not via REC.*****
- **The application should consist of one electronic file only, with an electronic signature from the PI (and supervisor if applicable). The completed application must incorporate all supplementary documentation, especially those being given to the proposed participants. The application will go through an initial triage process and will be returned to the applicant(s) if the form is incomplete or documentation is missing. If extensive changes are required, it will be reviewed at the next REC committee meeting. The application must be proofread and spellchecked before submission to the REC.**
- **All sections of the application form must be answered as instructed and within the word limits given.**

Applications which do not adhere to all of these requirements will not be accepted for review and will be returned directly to the applicant.

Applications must be completed on the form; answers in the form of attachments will not be accepted, except where indicated. No hardcopy applications will be accepted. **Research must not commence until written approval has been received from the Research Ethics Committee.**

Note: If your research requires approval from the [Biological Safety Committee \(BSC\)](#) this must be in place prior to REC submission. Contact bio.safety@dcu.ie. Please attach the responses from these committees to this submission as directed below.

PROJECT TITLE	Malta Government Budget Social Opinion Search and Summarisation
PRINCIPAL INVESTIGATOR(S) <i>The named Principal Investigator is the person with primary responsibility for the research project. In the case of PhD/D.Ed/MSc Research projects the supervisor must be listed as Principal Investigator, in addition to the student.</i>	Keith Cortis, Dr. Brian Davis
START AND END DATE	23/10/2021-30/11/2021
LEVEL OF RISK <i>Please indicate whether this project requires (a) notification (b) expedited or (c) full committee review. Justification for your choice is required under section 3.1</i>	Notification

1. ADMINISTRATIVE DETAILS

PROJECT TYPE: (mark Y to as many as apply)	Research Project	...	Funded Consultancy	...
			Clinical Trial	...
	Student Research Project (please indicate level below, e.g. PhD/D.Ed./MSc Research)	Y	Other - Please Describe:	...
	PhD / Other Doctorate	Y		
	D.Ed.			
	MSc Research	...		

1.1 INVESTIGATOR CONTACT DETAILS

PRINCIPAL INVESTIGATOR(S): In the case of PhD/D.Ed./MSc Research projects the supervisor must be listed as Principal Investigator. *Doctoral researchers and Research Masters may be listed as Principal Investigators, depending on the conventions of the discipline and on the individual case. It should be made clear, in subsequent sections of this application, who is carrying out the research procedures.*

NAME	SCHOOL/UNIT	EMAIL
Keith Cortis	ADAPT School of Computing	keith.cortis@adaptcentre.ie
Dr Brian Davis	ADAPT School of Computing	brian.davis@adaptcentre.ie

OTHER INVESTIGATORS:

NAME	SCHOOL/UNIT	EMAIL

1.2 WILL THE RESEARCH BE UNDERTAKEN ON-SITE AT DUBLIN CITY UNIVERSITY?

YES or NO
NO

If NO, state details of the off-campus location – provide details of the approval to gain access to that location in section 2.7.

Research will be carried out online using a meeting platform such as Zoom, via a web application.

1.3 WILL THIS RESEARCH INVOLVE ANIMALS?

YES or NO
NO

If YES, please provide details on the outcome from BRAG and attach copies of approval(s) received etc.

--

1.4 HAS THIS RESEARCH PROPOSAL BEEN SUBMITTED TO ANOTHER ETHICS COMMITTEE?

YES or NO
NO

If YES, please provide details on the outcome and attach copies of approval(s) received etc.

--

1.4.1 HAS THIS RESEARCH PROPOSAL BEEN REFUSED ETHICAL APPROVAL FROM THIS OR ANOTHER RESEARCH ETHICS COMMITTEE PREVIOUSLY?

If YES, please provide details.

NA

DECLARATION BY PRINCIPAL INVESTIGATOR(S)

The information contained herein is, to the best of my knowledge and belief, accurate. I have read the University's current research ethics guidelines, and accept responsibility for the conduct of the procedures set out in the attached application in accordance with the form guidelines, the [REC guidelines](#), the University's [Conflict of Interest Policy](#), its [Code of Good Research Practice](#) and any other condition laid down by the Dublin City University Research Ethics Committee. I have attempted to identify all risks related to the research that may arise in conducting this research and acknowledge my obligations and the rights of the participants.

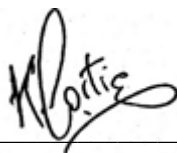
If there exists any affiliation or financial interest for researcher(s) in this research or its outcomes or any other circumstances which might represent a perceived, potential or actual conflict of interest this should be declared in accordance with the University's [Conflict of Interest Policy](#).

I and my co-investigators and/or supporting staff have the appropriate qualifications, experience and facilities to conduct the research set out in the attached application and to deal with any emergencies and contingencies related to the research that may arise. Supervisor(s) signature(s) is / are required as evidence that they have read and approve this submission.

Please note:

1. Any amendments to the original approved proposal must receive prior REC approval.
2. As a condition of approval investigators are required to document and report immediately to the Secretary of the Research Ethics Committee any adverse events, any issues which might negatively impact on the conduct of the research and/or any complaint from a participant relating to their participation in the study.

Electronic Signature(s):



Principal investigator(s): _____

Print Name(s) here: Keith Cortis _____

Date: 01/10/2021

I, the main supervisor of this research proposal, have read and approve this submission.

Supervisor(s) signature (where relevant): Brian Davis _____

Print Name(s) here: Brian Davis _____

Date: 01/10/2021

2. PROJECT OUTLINE

2.1 LAY DESCRIPTION, AIMS & JUSTIFICATION, METHODOLOGY (Approx.900 words)

Please outline, in terms that any non-expert would understand, what your research project is about, including what participants will be required to do. Please explain any technical terms or discipline-specific phrases. State the aims and significance of the project. Where relevant, state the specific hypothesis to be tested. Please provide a brief description of background research, a justification as to why this research project should proceed in that context and an explanation of any expected benefits to the community. **NB** – all references cited should be listed in an attached bibliography. Provide an outline of the proposed method and state who is doing which task – include details of data collection techniques, the tasks participants will be asked to do, the estimated time commitment involved, and how data will be analysed. If the project includes any procedure which is beyond already established and accepted techniques, please include a description of it. There should be enough detail provided to facilitate ethical review, but applicants are encouraged to keep it as succinct as possible.

Outline: This research project focuses on fine-grained opinion search and opinion summarisation at a multidimensional level for a real-world use case, namely the Malta Government Budget. Five social opinion dimensions are taken into consideration, namely subjectivity, sentiment polarity, emotion, irony, and sarcasm.

What participants are required to do: Participants shall be asked to carry a number of tasks using an opinion search and summarisation tool (web-based), that offers:

1. A *search functionality* for rich fine-grained social opinions of different aspects within social data, either via natural language or faceted navigation.
2. Visualisation of returned results in terms of an aspect-based feature-based *opinion summary*.

Technical terms/discipline-specific phrases:

- *Social Opinion Mining* on data obtained from social sources is an evolving research domain tasked with the identification of several opinion dimensions, such as subjectivity, sentiment polarity, emotion, irony, and sarcasm, from noisy user-generated content (all content is considered as social data) spread across heterogeneous sources, such as social media platforms and online newswires commenting sections (Cortis and Davis, 2021a).
- *Subjectivity* determines whether a sentence expresses an opinion –in terms of personal feelings or beliefs– or not, in which case a sentence expresses objectivity. Objectivity refers to sentences that express some factual information about the world (Liu, 2020).
- *Sentiment* determines the polarity (positive/negative/neutral) and strength/intensity (through a numeric rating score e.g., 1–5 stars, or level of depth e.g., low/high/medium) of an expressed opinion (Liu, 2010).
- *Emotion* refers to a person’s subjective feelings and thoughts, such as love, joy, surprise, anger, sadness, and fear (Liu, 2010).
- *Irony* is usually used to convey, the opposite meaning of the actual things you say, but its purpose is not intended to hurt the other person.
- *Sarcasm* holds the “characteristic” of meaning the opposite of what you say, but unlike irony, it is used to hurt the other person.
- An *aspect-level* or *feature-level* (Hu and Liu, 2004) (Pontiki et al., 2016) approach classifies a particular opinion dimension, such as sentiment polarity and/or emotion, for a given entity/aspect, such as a product or service, therefore producing fine-grained Opinion Mining.

Aims and significance of project: This main objective of this research project is to evaluate whether multidimensional opinion mining at an aspect-based level provides benefits for fine-grained opinion search and opinion summarisation.

Hypothesis to be tested: Does a multidimensional social opinion mining approach at an aspect-based level provide any additional benefits for fine-grained opinion search and opinion summarisation, when compared to a one-dimensional approach?

Background research:

The research area of Opinion Mining (Liu and Zhang, 2012), also referred to as Sentiment Analysis (Pang and Lee), is a popular and extremely valuable research area, especially for the exploitation of user-generated content extracted from social sources, such as social media platforms e.g., Twitter and newswires commenting sections (all content considered as social data). This Natural

Language Processing (Manning et al., 2014) task of Opinion Mining is considered as very challenging, especially when applied on social data, due to its semi-structured and/or unstructured format. This Natural Language Processing task is carried out to understand opinions as expressed by humans, which can help in the advancement of several real-world application areas, such as Politics (e.g., election prediction), Marketing, Advertising, Sales (e.g., brand/product awareness), Technology (e.g., company perception), and Finance (e.g., stock market prediction); industries, such as Hospitality (e.g., restaurant recommendations), and domains, such as Sports (e.g., fan sentiment), E-commerce (e.g., product recommendations), and Government (e.g., government-citizen interaction).

The opinion search and summarisation tool that will be used for this evaluation leverages a novel multidimensional and multilingual social opinion dataset in the socioeconomic domain, specifically Malta's annual Government Budget, which comprises social data from the 2018, 2019, and 2020 budgets (Cortis and Davis, 2021b). This social opinion dataset is valuable for multiple research applications, namely tools and resources for low-resources languages such as Maltese, Natural Language processing approaches for the analysis and processing of mixed-language (Maltese and English) user-generated content and fine-grained opinion search and summarisation at a multidimensional level (main focus of this evaluation). In fact, the research recently published by (Cortis et al., 2021) leveraged this social opinion dataset to fine-tune pre-trained neural language models that cater for the i) English, Maltese and Maltese-English languages as well as ii) five different social opinion dimensions, namely subjectivity, sentiment polarity, emotion, irony, and sarcasm.

Expected benefits to community: The Government of Malta can use social opinion summarisation results about certain national aspects (e.g., Environment domain, Social Security benefits) for policy formulation, policy-making, decision-making, and decision-taking. These results provide a nuanced voice to citizens who use social media platforms to make their opinions known and/or provide feedback about anything, in this case Malta's annual Government Budget. User-generated content has always been invaluable for certain needs, such as improving an entity's service/perception and tracking citizen opinion.

Outline of proposed method (data collection techniques, tasks for participants, estimated time commitment, how data will be analysed):

- Data collection techniques: Results from the tasks carried out by each participant will be answered as part of an evaluation questionnaire that will be provided. A short pre-evaluation questionnaire consisting of a few generic questions about this research area will be also presented to evaluate the proficiency of each participant in this regard.
- Tasks for participants: Each participant will be asked to carry out 4-5 tasks using the opinion search and summarisation tool (web-based).
- Estimated time commitment: 60 minutes.
- How data will be analysed: A quantitative and qualitative analysis of the results of the pre-evaluation and evaluation questionnaires will be carried out.

2.2 INVESTIGATORS' QUALIFICATIONS, EXPERIENCE AND SKILLS (Approx. 200 words)

List the academic qualifications and outline the experience and skills *relevant to this project* that the PI, other researchers and any supporting staff have in carrying out the research and in dealing with any emergencies, unexpected outcomes, or contingencies that may arise. **State specifically who will be carrying out the research procedures.**

Keith Cortis is a part-time PhD student with the main research focusing on multidimensional Social Opinion Mining. He is in possession of:

- 2013, Master of Applied Science (DERI), NUI Galway
- 2010, Bachelor of Science in Information Technology (Honours) (Melit.), Computer Science and Artificial Intelligence, University of Malta

Keith has experience in the ICT domain as a full-time professional and has worked in both academia and industry on numerous fields and domains ranging from social semantic web, language technologies, robotics, and government.

Dr. Brian Davis is an Assistant Professor at the DCU School of Computing, a member of the ADAPT research centre, and an SFI Funded Investigator. He is in possession of:

- 2018, Postgraduate Diploma in Academic Practice, NUI Galway

- 2017, Postgraduate Certificate in Teaching and Learning, NUI Galway
- 2013, PhD in Engineering by Research (DERI), NUI Galway
- 2004, MSc in Computer Science by Research, Trinity College Dublin
- 2001, BSc in Applied Computational Linguistics & German, Dublin City University

Brian's core expertise intersects with Natural Language Processing (NLP), Ontology Engineering. Other research interests include: Computer and Data Ethics, NLP for social media, cross lingual opinion mining from social media for the finance and political domains and combining visualisation and textualisation using Natural Language Generation.

2.3 PARTICIPANT PROFILE

List and very briefly describe each participant group where applicable. For instance, participant group 1 will consist of..., participant group 2 will consist of... etc. Provide the number, age range and source of participants. Please provide a justification of your proposed sample size.

All participants shall be over 16 years of age, are required to have basic digital skills (e.g., familiar with using a PC, browser, etc.), able to understand the English and Maltese languages (since data used for evaluation contains data in both languages) and have basic knowledge of the annual Malta Government Budget. There will be 6 participant groups in total, each consisting of 3 people. Each participant group shall be evaluating three opinion search and summarisation systems, one focusing on sentiment polarity, one on sentiment polarity and emotion, and one on sentiment polarity, emotion, and irony or sarcasm. Each system shall require all participants to complete a set of tasks. An additional participant group will consist of 3-5 subject matter experts within the Government of Malta, where they shall be asked to carry out the same opinion search and summarisation systems and a follow-up interview.

The sample size (18-23) satisfies the requirements for reliable System Usability Scale (SUS) evaluations (Tullis and Stetson, 2004).

2.4 PARTICIPANT RECRUITMENT

Please provide specific details as to how you will be recruiting participants. How will people be informed that you are doing this research? How will they be approached and asked if they are willing to participate? If you are mailing or phoning people, please explain how you have obtained their names and contact details. If a recruitment advertisement is to be used, please ensure you attach a copy to this application (Approx. 100 words).

A list of potential participants satisfying the participant profile discussed above will be drawn and subsequently invited via a communication channel (email, phone, etc.). The list will consist of people who are acquaintances (e.g., work) of the PIs.

2.5 IS IT LIKELY THAT ANY PARTICIPANTS COULD BE CONSIDERED POTENTIALLY VULNERABLE?

Are some or all participants vulnerable in any way? (e.g. by virtue of the group they belong to, people who have undergone traumatic or adverse emotional events, people with diminished cognitive ability, power relations between researchers and participants etc.)?

YES or NO
NO

If Yes, please state and describe what this vulnerability (or vulnerabilities) is and justify why this research is being done with such participants

2.6 WILL THE IDENTITY OF THE PARTICIPANTS BE PROTECTED?

YES or NO
YES

If NO, please explain why

IF YOU ANSWERED YES TO 2.6, PLEASE ANSWER THE FOLLOWING QUESTION:

2.7 HOW WILL THE ANONYMITY OF THE PARTICIPANTS BE RESPECTED?

Please bear in mind that where the sample size is very small, it may be impossible to guarantee anonymity/confidentiality of participant identity. Participants involved in such projects need to be advised of this limitation in the Plain Language Statement/Information Sheet. If you intend to fully anonymize the data, please provide details.

The identity of the participants will not be collected since it is not within scope of the human evaluation to be carried out.

2.8 LEGAL LIMITATIONS TO DATA CONFIDENTIALITY

Participants need to be made aware that confidentiality of information provided cannot always be guaranteed by researchers and can only be protected within the limitations of the law - i.e., it is possible for data to be subject to subpoena, freedom of information claim or mandated reporting by some professions. This information should be included in your Plain Language Statement and Informed Consent Form. Depending on the research proposal and academic discipline, you may need to state additional specific limitations.

State how and where participants will be informed of these limitations.

NA – refer to previous answer.

2.9 CHILD PARTICIPANTS (anyone under 18 years old)

If your participants include children, you **must** confirm that you are in compliance with the research specific guidelines as detailed in [Keeping Children Safe - Policies and Procedures supporting Child Protection at DCU](#).

Please indicate your compliance with the following guidelines:	Mark here
We confirm that we have read and agree to act in accordance with the DCU Child Protection policy and procedures	
We confirm that we have put in place safeguards for the children participating in the research	
We confirm that we have supports in place for children who may disclose current or historical abuse (whether or not this is the focus of the research)	

2.10 PLEASE EXPLAIN WHEN, HOW, WHERE, AND TO WHOM RESULTS WILL BE DISSEMINATED, INCLUDING WHETHER PARTICIPANTS WILL BE PROVIDED WITH ANY INFORMATION AS TO THE FINDINGS OR OUTCOMES OF THE PROJECT?

After the overall results are analysed, these will be published in an academic conference. The participants will be provided with a copy of the study for their reference. However, it's important to note that no personal identities shall be included within this study, since the main aim is to evaluate the proposed technique (in the form of three systems) only.

2.11 ARE OTHER APPROVALS REQUIRED TO GAIN ACCESS TO ANOTHER LOCATION, ORGANISATION, SCHOOL ETC.?

YES or NO
NO

If YES, please specify from whom and attach a copy of the approval documentation. If this is not yet available, please explain when this will be obtained.

3. RISK AND RISK MANAGEMENT

3.1 EXPLAIN AND JUSTIFY THE STATED LEVEL OF RISK TO PARTICIPANTS

You must provide a justification for the stated level of risk and its corresponding level of review (Full Committee, Expedited, Notification), as indicated on the cover page of your application. Note that the level of risk may be influenced by the vulnerability of the research group, the methods employed and the nature of the research itself. For further information on risk levels, please refer to the Levels of Review information on the [Research Support Services website](#).

No risks to participants since evaluation is focused on the proposed technique and not the people evaluating it.

3.2 POTENTIAL RISKS TO PARTICIPANTS AND RISK MANAGEMENT PROCEDURES

Identify, as far as possible, all potential risks to participants (physical, psychological, social, legal, economic, etc.), associated with the proposed research. Will your research involve deception, investigation of participants involved in illegal activities, performance of any acts which might diminish the self-esteem of participants or cause them to experience embarrassment, regret or depression, administration of any substance or agent, collection of body tissues or fluid samples, use of non-treatment of placebo control conditions, collection and/or testing of DNA samples, administration of ionising radiation? Please explain what risk management procedures will be put in place to minimise these risks.

No potential risks since none of the mentioned research shall be carried out during this evaluation.

3.3 ARE THERE LIKELY TO BE ANY BENEFITS (DIRECT OR INDIRECT) TO PARTICIPANTS FROM THIS RESEARCH?

YES or NO
NO

If YES, provide details

--

3.4 ARE THERE ANY SPECIFIC RISKS TO RESEARCHERS?

Examples include use of dangerous materials, asking certain types of questions, research being undertaken in certain locations, researchers working alone in isolated areas, etc.

YES or NO
NO

If YES, please describe and explain what risk management procedures will be put in place to minimise these risks

--

3.5 DEALING WITH ADVERSE / UNEXPECTED OUTCOMES

Please describe what measures/protocols you have put in place in the event that there are any unexpected outcomes or adverse effects to participants arising from involvement in the project.

There are no risks for unexpected outcomes or adverse effects to participants.

3.6 SUPPORT FOR PARTICIPANTS

Depending on risks to participants you may need to consider having additional support for participants during/after the study. Consider whether your project would require additional support, e.g., external counselling available to participants. Please advise what support will be available.

Not applicable for proposed evaluation.

3.7 HOW WILL THE CONDUCT OF THE PROJECT BE MONITORED?

Please explain how the principal investigator will monitor the conduct of the project (especially where several people are involved in recruiting or interviewing, administering procedures, etc.) to ensure that it conforms to the procedures set out in this application. In the case of student projects please give details of how the supervisor(s) will monitor the conduct of the project.

The PI will initially pilot the human evaluation with 3 participants to test the robustness of the system (Note that these same participants shall not be used for the human evaluation). Once this is complete, the PI shall carry out an initial session with each participant group (i.e., each evaluator) to explain the main objectives of the evaluation and what is expected from their end. Recruitment shall be carried out solely by the PI, therefore, no other people are involved in this process.

3.8 DO YOU PROPOSE TO OFFER PAYMENTS OR INCENTIVES TO PARTICIPANTS?

YES or NO
NO

If YES, please provide further details

--

3.9 DO ANY OF THE RESEARCHERS ON THIS PROJECT HAVE A PERSONAL, PHILOSOPHICAL, FINANCIAL, POLITICAL, IDEOLOGICAL, OR COMMERCIAL INTEREST IN ITS OUTCOME THAT MIGHT INFLUENCE THE INTEGRITY OF THE RESEARCH, OR BIAS THE CONDUCT OR REPORTING OF THE RESEARCH, OR UNDULY DELAY OR OTHERWISE AFFECT THEIR PUBLICATION?

YES or NO
NO

If YES, please specify how this conflict of interest will be addressed

--

4. PERSONAL DATA

Definition of Personal Data

Personal data is any information about a living person, where that person is either identified or could be identified, from the data itself or when it is combined with other data. Typical examples of personal data in a research context are:

- a) paper based records e.g. consent forms, research participant files, patient records, interview notes etc.
- b) electronic records e.g. database of participant details, online survey returns, photos, audio & visual recordings, IP addresses, diagnostic / clinical imaging etc.
- c) other e.g. genetic data, biometric data, clinical or medical samples etc.

Note: If personal data is to be obtained and / or processed in the course of the proposed research then there are certain legal obligations and principles to be followed. These are set out in the 2016 General Data Protection Regulation (GDPR) and associated Irish Law.

Any data that is fully and completely anonymous is not considered to be 'personal data'. However, any data that is merely pseudo-anonymised is deemed to be 'personal data'.

Further information on data protection issues is available from the University's [Data Protection Unit \(DPU\)](#). You should also consider consulting with your Unit's [GDPR Advocate](#) for help and advice on filling out this section of the form.

(A) Your knowledge of Data Protection

Have you taken and completed the online data protection training course ('Data Protection Course') that is available to all staff and students through the DCU Loop System ?	YES or NO	YES
--	-----------	-----

If you answered 'No' to the previous question then the DPU strongly recommends that all applicants complete the course on Loop before completing section # 4 of the REC Application Form.

If you experience difficulties in accessing the Loop course at the link above, please contact the [Teaching Enhancement Unit](#) for assistance.

(B) Initial Assessment of whether any of the data to be used in the proposed research is 'Personal Data' (see definition above)

1	Will the proposed research include living human subjects? <i>Rationale – personal data applies only to living individuals.</i>	YES or NO	YES
2	Will the proposed research use any data that can be linked to an identified, or an identifiable, person? <i>Rationale – to be personal data it must be possible to associate it with an identified, or an identifiable, living person.</i>	YES or NO	NO
3	Will the proposed research use any data identifiers that can be linked to a living person? Examples are a participant's name, code or ID number, their address, their IP address etc. <i>Rationale: fully anonymised data is not deemed to be 'personal data' but data that has been deemed to be merely pseudo-anonymised is deemed to be 'personal data'.</i>	YES or NO	NO

If you answered 'Yes' to any of the questions 1 to 3 in sub-section (B), then continue to sub-section (C) and answer questions 1-8. If you answered 'No' to all of the questions 1 to 3 in sub-section (B), then proceed directly to section # 5 of this Application Form.

(C) Assessing the degree of risk inherent in the personal data			
1	Will the proposed research involve the use of <u>personal data</u> on individuals that reveals any of the following attributes or characteristics about them? (State 'Yes' or 'No' as appropriate to all of the following)		
	<i>Racial or Ethnic Origin</i>	YES or NO	NO
	<i>Political Opinions</i>	YES or NO	NO
	<i>Religious or Philosophical Beliefs</i>	YES or NO	NO
	<i>Trade Union Membership</i>	YES or NO	NO
	<i>Genetic Data</i>	YES or NO	NO
	<i>Biometric Data</i>	YES or NO	NO
	<i>Data Concerning Health</i>	YES or NO	NO
	<i>Data concerning a Person's Sex Life or Sexual Orientation</i>	YES or NO	NO
2	Will the proposed research involve the use of <u>personal data</u> relating to children or vulnerable individuals? <i>A child, for data protection purposes, is defined as an individual below 18 years of age. Where the processing relates to 'electronic marketing' the age limit is reduced to 16 years. A vulnerable individual may be anyone who is unable to consent to, or to oppose, the processing of his or her data for any reason, including disability.</i>	YES or NO	NO
3	Will the proposed research involve the use of data relating to an individual's criminal convictions and / or offences?	YES or NO	NO
4	Will the proposed research involve the large-scale processing of <u>personal data</u> ? <i>This may include: a wide range or large volume of personal data; processing which takes place over a large geographical area; processing where a large number of people are affected (e.g. over 100 individuals); or where the processing is extensive or it has potential long-lasting effects on individuals.</i>	YES or NO	NO
5	Will the proposed research involve any form of <u>automated processing</u> of personal data? <i>In particular, to analyse or predict aspects concerning that person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.</i>	YES or NO	NO
6	Will the proposed research involve the sharing or transferring of any personal data to a 3 rd party outside of DCU? <i>For example, other research partners, providers of translation or transcription services, etc.</i> <i>For clarity, this question is not intended to refer to any standard software services already provided by DCU, for example the university's email system or its cloud-based storage provider (Google Drive).</i>	YES or NO	NO

7	<p>Will the proposed research require the sharing or processing of personal data outside the EU or the EEA? (e.g. the US, the UK, Canada, Australia, China etc.)</p> <p><i>The EEA refers to the 'European Economic Area' (i.e. the EU plus Norway, Liechtenstein and Iceland).</i></p>	YES or NO	NO
8	<p>Will the proposed research involve the matching or combining of separate datasets of information on individuals in a way that would exceed their reasonable expectations of privacy?</p> <p><i>This is especially important where two or more previously anonymous datasets are combined in such a way so as to allow for the identification of individuals. An example would be combining mobile phone location data along with any other dataset to identify individuals.</i></p>	YES or NO	NO

Important Point: Next Step

If you answered 'Yes' to one or more of the questions 1 to 8 in sub-section (C) **you must contact the Data Protection Unit (DPU) prior to submitting this application form to the REC.** The DPU will assess whether there are any further data protection issues to be addressed or additional procedures to be followed.

5. DATA / SAMPLE STORAGE, SECURITY AND DISPOSAL

For the purpose of this section the term 'Data' includes personal data that is in a raw or a processed state (e.g. interview audiotape, transcript or analysis, etc.). The term 'Samples' include body fluids and/or tissue samples.

5.1 HOW AND WHERE WILL THE DATA / SAMPLES BE STORED?

DCU recommends that any data stored electronically offsite should utilise the DCU Google Drive. Alternative offsite storage will need to be justified and must meet data protection and GDPR compliance requirements.

No recordings or transcripts shall be required from the 6 participant groups (3 people each). As for the additional participant group (3-5 subject matter experts), the replies to the interview questions shall be stored electronically on the DCU Google Drive - secured. The same applies for the answers to the questionnaires that each participant shall be required to fill in before and during the human evaluation.

5.2 WHO WILL HAVE ACCESS TO DATA / SAMPLES?

If people other than the main researchers have access, please name who they are and explain for what purpose.

Only main researchers.

5.3 HOW LONG IS THE DATA TO BE HELD OR RETAINED?

*Note that, with very few exceptions, **Personal Data** may not be retained indefinitely. It is up to the research team to establish an upper retention limit for each category of Personal Data used within the project and to ensure it is applied at the expiry of that limit.*

No personal data shall be collected during the human evaluation.

5.4 WILL THE PERSONAL DATA BE USED AT A LATER DATE FOR THE PURPOSE OF PUBLICATION OF THE RESULTS OF THE RESEARCH?

YES or NO

NO

Where it is intended that the personal data used in the project will be used at a later date for the purposes of publication please explain how consent to do so will be obtained.

--

5.5 IF THE DATA/SAMPLES ARE TO BE DISPOSED OF AT THE END OF THE PROJECT PLEASE EXPLAIN HOW, WHEN AND BY WHOM THIS WILL BE DONE?

*Note that simply deleting files is not sufficiently secure. The additional steps to be taken to maintain data security should be given. **Personal data** must be disposed of in a safe and secure manner at the end of its retention period. If the data is stored in (a) a paper-based format, then shredding or disposal via a secure bin is recommended; or (b) in an electronic-based format, then deletion of the record or the full anonymization of the data is recommended. If data/samples are **not** being disposed of, please justify that intention.*

<p>How will the data/samples be disposed?</p> <p>Please describe the means by which the personal data will be deleted or destroyed. This includes personal data held in hard copy and digital formats.</p>	NA
<p>When will the data/samples be disposed?</p> <p>Please indicate the intended retention period of the personal data, and reasons for this retention period. Please note that retention periods must be GDPR compliant and must be consistent with the DCU Retention Policy.</p>	NA
<p>By whom will the data/samples be disposed?</p> <p>Please indicate the designated team member(s) with responsibility for deletion and/or destruction of the research project's personal data.</p>	NA

6. FUNDING OF THE RESEARCH

6.1 HOW IS THIS WORK BEING FUNDED?

Work is part of the PhD, therefore no funding is being used.
--

6.2 PROJECT GRANT NUMBER (If relevant and/or known – otherwise mark as N/A)

NA

6.3 DOES THE PROJECT REQUIRE APPROVAL BEFORE CONSIDERATION FOR FUNDING BY A GRANTING BODY?

YES or NO
NO

6.4 HOW WILL PARTICIPANTS BE INFORMED OF THE SOURCE OF THE FUNDING? (E.g. included in the Plain Language Statement)

NA

6.5 DO THE FUNDERS OF THIS PROJECT HAVE A PERSONAL, FINANCIAL, POLITICAL, IDEOLOGICAL, OR COMMERCIAL INTEREST IN ITS OUTCOME THAT MIGHT COMPROMISE THE INDEPENDENCE AND INTEGRITY OF THE RESEARCH, OR BIAS THE CONDUCT OR REPORTING OF THE RESEARCH, OR UNDULY DELAY OR OTHERWISE AFFECT THEIR PUBLICATION?

YES or NO
NO

If YES, please specify how this conflict of interest will be addressed

7. PLAIN LANGUAGE STATEMENT (Attach to this document. Approx. 400 words)

A Plain Language Statement (PLS) should be used in all cases. This is written information in plain language that you will be providing to participants, outlining the nature of their involvement in the project and inviting their participation. The PLS should specifically describe what will be expected of participants, the risks and inconveniences for them, and other information relevant to their involvement. Please note that the language used must reflect the participant age group and corresponding comprehension level– if your participants have different comprehension levels (e.g. both adults and children) then separate forms should be prepared for each group. The PLS can be embedded in an email to which an online survey is attached, or handed/sent to individuals in advance of their consent being sought. See the link to sample templates on the [Ethics Approval section](#) of the Research Support Services website.

PLEASE CONFIRM WHETHER THE FOLLOWING ISSUES HAVE BEEN ADDRESSED IN YOUR PLAIN LANGUAGE STATEMENT/ INFORMATION SHEET FOR PARTICIPANTS:

	YES or NO
Introductory Statement (PI and researcher names, school, title of the research)	YES
What is this research about?	YES
Why is this research being conducted?	YES
What will the participant be expected to do/have to do if they decide to participate in the research study?	YES
How will their privacy be protected?	YES
How will the data be used and subsequently disposed of?	NO
What are the legal limitations to data confidentiality?	NO
Are there any benefits of taking part in the research study?	YES
Are there any risks of taking part in the research study?	YES
Confirmation that participants can change their mind at any stage and withdraw from the study	YES
How will participants find out what happens with the project?	NO
Contact details for further information (including REC contact details)	YES
Details relating to GDPR Compliance where Personal Data is being sought	NO

If any of these issues are marked NO, please justify their exclusion:

Several statements listed as NO since they are not applicable due to the reasons mentioned in this ethics form (refer to previous sections for more information).

8. INFORMED CONSENT FORM (Attach to this document. Approx. 300 words)

In most cases where interviews or focus groups are taking place, an Informed Consent Form is required. This is an important document requiring participants to indicate their consent to participate in the study and give their signature. In cases where an anonymous questionnaire is being used, it is not enough to include a tick box in the questionnaire. Participants should indicate their consent to each aspect of the research in a staged manner by checking mandatory checkboxes.

See link to sample templates on the [Ethics Approval section](#) of the Research Support Services website.

NB – IF AN INFORMED CONSENT FORM IS NOT BEING USED, THE REASON FOR THIS MUST BE JUSTIFIED HERE.

9. ASSENT FORM & PLAIN LANGUAGE STATEMENT FOR CHILDREN (Attach to this document.)

A child specific Plain Language Statement (PLS) should be used in research where children will be involved. The PLS must be written in a way that is understandable for children within your targeted age group. It also must state, in plain language, the nature of their involvement in the project and inviting their participation. The PLS should specifically describe what will be expected of participants, the risks and inconveniences for them, and other information relevant to their involvement. In addition, child participants should also be provided with an Assent Form. Parents/guardians will be provided with the Informed Consent Form, but each child should provide assent before taking part in the research. The Assent Form needs to be understandable to the age-group you are targeting. See link to sample templates on the [Ethics Approval Section](#) of the Research Support Services website.

NB – IF AN ASSENT FORM IS NOT BEING USED, THE REASON FOR THIS MUST BE JUSTIFIED HERE.

10. SUBMISSION CHECKLIST (Attach to this document)

Please confirm that all supplementary information is included in your application (in electronic copy). If questionnaire or interview questions are submitted in draft form, please indicate this by putting (draft) after YES. A copy of the final documentation must be submitted for final approval when available.

My application has been collated as one electronic file which includes the following documentation:	INCLUDED (mark as YES)	NOT APPLICABLE (mark as N/A)
Bibliography	YES	
Recruitment advertisement		N/A
Plain language statement/Information Statement	YES	
Informed Consent form	YES	
Informed Assent form		N/A
Evidence of external approvals related to the research		N/A
Questionnaire / Survey		N/A
Interview / Focus Group Questions		Usability evaluation answers to the questions will be based on System Usability Scale (Brooke, 1986), Mean Opinion Score (Streijl et al., 2016), and a free-text format field (for any further comments by evaluators).
Debriefing material		N/A
Other (e.g. BSC approval review letter, Data Protection Impact Assessment)		N/A

DUBLIN CITY UNIVERSITY

Bibliography

- Keith Cortis and Brian Davis. 2021a. Over a decade of social opinion mining: a systematic review. *Artificial intelligence review*, pages 1–93. <https://link.springer.com/article/10.1007/s10462-021-10030-2>
- Keith Cortis and Brian Davis. 2021b. A dataset of multidimensional and multilingual social opinions for malta's annual government budget. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 971–981. <https://ojs.aaai.org/index.php/ICWSM/article/view/18120>
- Keith Cortis, Kanishk Verma and Brian Davis. 2021. Fine-tuning Neural Language Models for Multidimensional Opinion Mining of English-Maltese Social Data. In *Proceedings of Recent Advances in Natural Language Processing*. <https://ranlp.org/ranlp2021/proceedings.pdf>
- Liu, B., 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010), pp.627-666.
- Hu, M. and Liu, B., 2004, August. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177).
- Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O. and Hoste, V., 2016, January. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation* (pp. 19-30).
- Liu, B. and Zhang, L., 2012. A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415-463). Springer, Boston, MA.
- Pang, B. and Lee, L. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.* 2, 1–2 (January 2008), 1–135. DOI:<https://doi.org/10.1561/1500000011>
- Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S. and McClosky, D., 2014, June. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55-60).
- Tullis, T.S. and Stetson, J.N., 2004, June. A comparison of questionnaires for assessing website usability. In *Usability professional association conference* (Vol. 1, pp. 1-12).
- Brooke, J., 1986. System usability scale (SUS): a quick-and-dirty method of system evaluation user information. *Reading, UK: Digital Equipment Co Ltd*, 43, pp.1-7.
- Streijl, R.C., Winkler, S. and Hands, D.S., 2016. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, 22(2), pp.213-227.

DUBLIN CITY UNIVERSITY

Malta Government Budget Social Opinion Search and Summarisation – Plain Language Statement

Introduction to the Research Study

- **Research Study Title:** Malta Government Budget Social Opinion Search and Summarisation
- **Research Overview:** This research project focuses on fine-grained opinion search and opinion summarisation at a multidimensional level for a real-world use case, namely the Malta Government Budget. Five social opinion dimensions are taken into consideration, namely subjectivity, sentiment polarity, emotion, irony, and sarcasm.
- **University Department:** School of Computing
- **Principal Investigators:**
 - Keith Cortis, email: keith.cortis@adaptcentre.ie
 - Dr. Brian Davis, email: brian.davis@adaptcentre.ie

Privacy Notice

No personal data shall be collected for this study since it is not within scope of the human evaluation to be carried out.

Details of what participant involvement in the Research Study will require

Human evaluation will consist of the following:

- Short online session to explain the main objectives of the evaluation and what is expected from each participant
- Fill in a short pre-evaluation questionnaire
- Use an Opinion Search and Summarisation online tool via a browser to carry out a set of pre-determined tasks. This tool offers participants the possibility of:
 - *Searching* for rich fine-grained social opinions of different aspects within the social data provided about the Malta Government Budget, either via natural language or faceted navigation;
 - Visualisation of returned results in terms of an aspect-based feature-based *opinion summary*.
- Fill in an evaluation questionnaire about the tasks carried out
- Follow-up interview about the tasks carried out (only to Subject Matter Experts)

Estimated time commitment for all activities: 60 minutes

Potential risks to participants from involvement in the Research Study (if greater than that encountered in everyday life)

No risks to participants since evaluation is focused on the proposed technique and not the people evaluating it.

Benefits to participants from involvement in the Research Study

The benefits of taking part in this research study include increased knowledge of opinion mining research and how it can be used e.g., for policy-making and decision-making, to provide a nuanced voice to citizens who use social media platforms to make their opinions known and/or provide feedback about anything, in this case Malta's annual Government Budget.

Involvement in the Research Study

Involvement in this research study is voluntary and the participants can withdraw their consent at any point of the human evaluation session.

If participants have concerns about this study and wish to contact an independent person, please contact:

The Secretary, Dublin City University Research Ethics Committee, c/o Research and Innovation Support, Dublin City University, Dublin 9. Tel 01-7008000, e-mail rec@dcu.ie

DUBLIN CITY UNIVERSITY

Malta Government Budget Social Opinion Search and Summarisation – Informed Consent Form

Research Study Title

- **Research Study Title:** Malta Government Budget Social Opinion Search and Summarisation
- **Research Overview:** This research project focuses on fine-grained opinion search and opinion summarisation at a multidimensional level for a real-world use case, namely the Malta Government Budget. Five social opinion dimensions are taken into consideration, namely subjectivity, sentiment polarity, emotion, irony, and sarcasm.
- **University Department:** ADAPT, School of Computing
- **Principal Investigators:**
 - Keith Cortis, email: keith.cortis@adaptcentre.ie
 - Dr. Brian Davis, email: brian.davis@adaptcentre.ie

Clarification of the purpose of the research

No personal data shall be collected for this study since it is not within scope of the human evaluation to be carried out.

Confirmation of particular requirements as highlighted in the Plain Language Statement

Participant – please complete the following (Circle Yes or No for each question)

I have read the Plain Language Statement (or had it read to me)	Yes/No
I understand the information provided	Yes/No
I understand the information provided in relation to data protection	Yes/No
I have had an opportunity to ask questions and discuss this study	Yes/No
I have received satisfactory answers to all my questions	Yes/No
I am aware that my interview will be audiotaped	Yes/No

Confirmation that involvement in the Research Study is voluntary

I am involving myself in this research study on a voluntary basis.	Yes/No
I may withdraw from the Research Study at any point of the evaluation.	Yes/No

Note: No personal data shall be gathered and stored for the purposes of this Research Study.

Signature:

I have read and understood the information in this form. My questions and concerns have been answered by the researchers, and I have a copy of this consent form. Therefore, I consent to take part in this research project

Participants Signature: _____

Name in Block Capitals: _____

Witness: _____

Date: _____

Appendix E

Human Evaluation - Pre-Evaluation Questionnaire

A short pre-evaluation questionnaire consisting of a few generic questions about the participant and this research area.

***Required**

1. What is your professional background? * (Please mark only one)

Academia

Private Sector

Public Sector

Public Service

Student

Other: _____

2. Which sector do you work in? * (Please mark only one)

ICT

Other: _____

3. Select the highest level of education you have * (Please mark only one)

- MQF Level 8 (Doctorate Degree)
 - MQF Level 7 (Master's Degree/Postgraduate Diploma/Postgraduate Certificate)
 - MQF Level 6 (Bachelor's Degree)
 - MQF Level 5 (Undergraduate Diploma/Undergraduate Certificate)
 - MQF Level 4 (Matriculation Certificate/Advanced Level/Intermediate Level)
 - MQF Level 3 (General Education SEC Grade 1-5)
 - MQF Level 2 (General Education SEC Grade 6-7)
 - MQF Level 1 (General Education School Leaving Certificate)
4. I have knowledge about the research area of “Opinion Mining” or “Sentiment Analysis” * (Please mark only one)
- No
 - A litte
 - Yes
5. I have worked on “Opinion Mining” or “Sentiment Analysis” * (Please mark only one)
- Never
 - Sometimes
 - Often
6. I am familiar with Malta’s annual Government Budget * (Please mark only one)
- No
 - A litte
 - Yes
7. The Government of Malta annual budget is announced by the Ministry for Finance * (Please mark only one)

- Never
- Sometimes
- Often

8. I follow the Malta Government Budget consistently each year * (Please mark only one)

- No
- A litte
- Yes

9. Each year I read about the main updates of the Malta Government Budget * (Please mark only one)

- Never
- Sometimes
- Often

Appendix F

Human Evaluation - Overview of Opinion Search and Summarisation System

Opinion Search and Summarisation System

Background

Purpose of Opinion Summaries

Multidimensional opinion summarisation provides a **nuanced voice to the general public** i.e., citizens and residents of Malta, on their opinions about a topic that is related to the annual Malta Government Budget. These opinions are summarised for the Government of Malta in terms of different opinion dimensions i.e., subjectivity, sentiment polarity, emotion, irony, and sarcasm, **to help in any decision-making and/or policy-making processes.**

The definition for each opinion dimension is as follows:

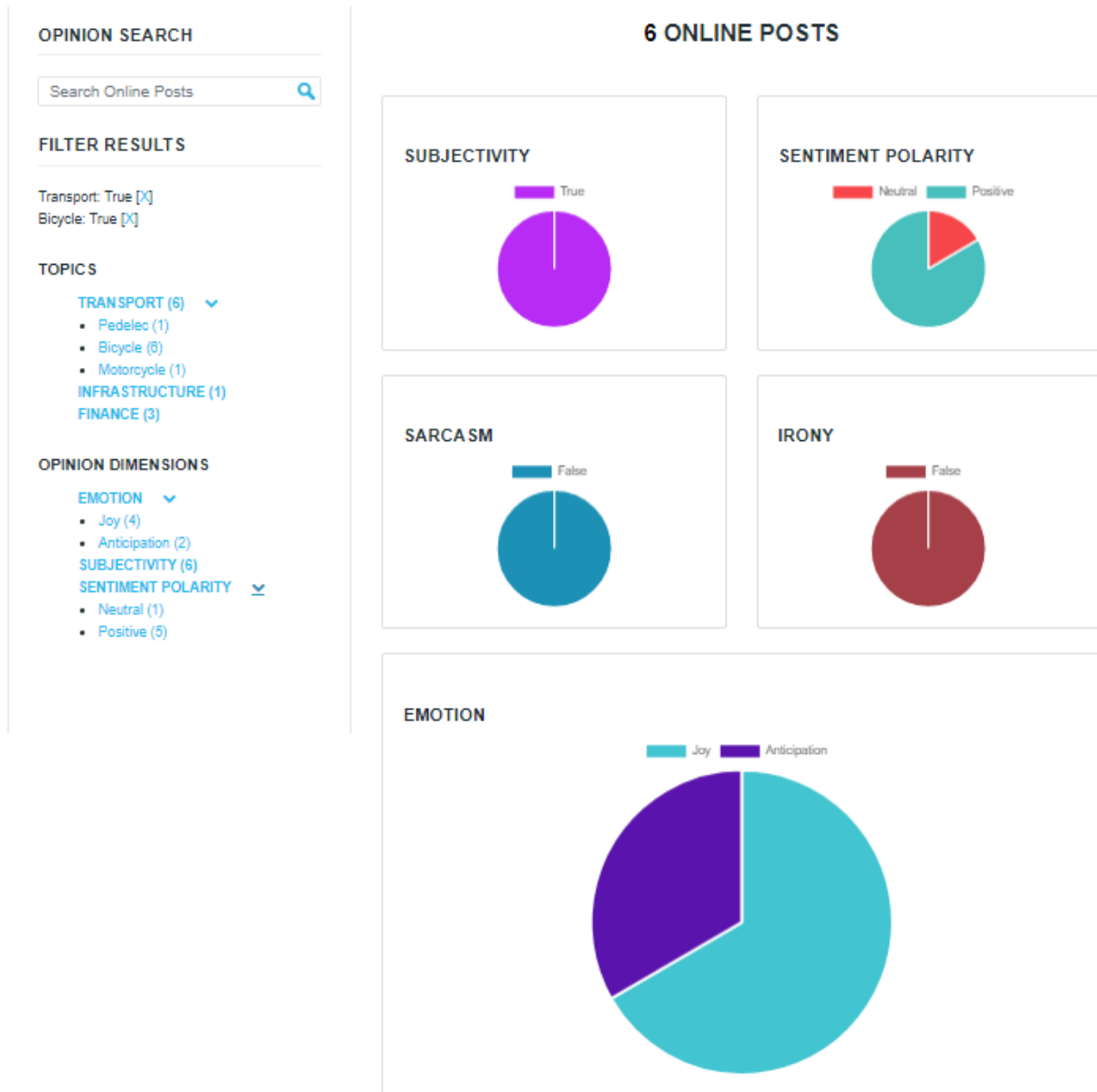
- **Subjectivity:** determines whether a sentence expresses an opinion– in terms of personal feelings or beliefs;
- **Sentiment Polarity:** determines the polarity i.e., positive, negative, or neutral, of an expressed opinion;
- **Emotion:** refers to a person’s subjective feelings and thoughts, such as love, joy, surprise, anger, sadness, and fear;
- **Sarcasm:** holds the “characteristic” of meaning the opposite of what one says, but unlike irony, it is used to hurt the other person towards whom the sarcasm is directed;
- **Irony:** used to convey the opposite meaning of the actual things one says, but its purpose is not intended to hurt another person.

Context behind Visualisations of Opinion Summaries

The opinion summaries visualised through pie charts shall provide a basic and easy-to-understand visualisation of the mentioned opinions posed by the general public. These can be used by the Government of Malta for further **social, economic and/or political considerations**, as previously mentioned.

The example provided below shows an opinion summary about the “**Transport**” topic and “**Bicycle**” sub-topic. The opinion summary visualisation is based on a total of 6 online posts. In summary, this shows that all the online posts focused on the mentioned topic and respective sub-topic consist of the following:

- **Subjectivity:** All of the online posts are *subjective* (100%);
- **Sentiment Polarity:** 83.33% of the online posts are *positive* and 16.67% are *neutral*;
- **Emotion:** 66.67% of the online posts convey a *joy* emotion and 33.33% convey an *anticipation* emotion;
- **Sarcasm:** None of the online posts are *sarcastic*;
- **Irony:** None of the online posts are *ironic*.



Example Use Case

The following is a hypothetical real-world use case of how the opinion search and summarisation tool can be used by the Government of Malta for any decision-making and/or policy-making processes. In Malta’s annual Government Budget of 2019, the Minister for Finance announced that all students shall be offered free public transport. The Government of Malta used this tool to analyse the feedback provided by the general public on the said measure. Given that the overall feedback was very positive (see opinion summaries visualised below), the Government of Malta decided to also include the Gozo fast ferry services as part of this scheme. This was announced in Malta’s annual Government Budget of 2022.

OPINION SEARCH



FILTER RESULTS

Transport: True [X]

Students: True [X]

TOPICS

TAXATION (2) ▾

TRANSPORT (8) ▾

PENSIONS (1) ▾

EDUCATION (1)

OPINION DIMENSIONS

EMOTION ▾

SUBJECTIVITY (8)

SENTIMENT POLARITY ▾

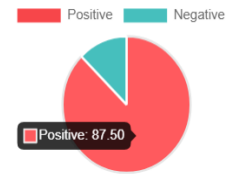
SARCASM (1)

8 ONLINE POSTS

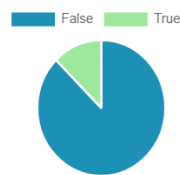
SUBJECTIVITY



SENTIMENT POLARITY



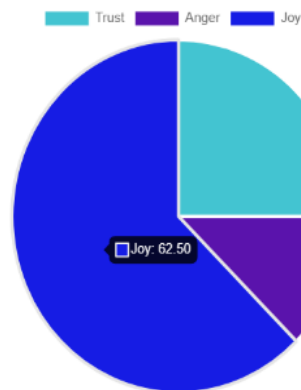
SARCASM



IRONY



EMOTION



Stqsi lil dawk li int ukoll kont tircievi kontijiet tad-dawl u l-ilma gholjien, staqsi lil dawk il-ponsonanti li tahtkom ghal 25 sema ma hadux zieda, staqsi lil dawk l-eluf li kienu jirregistraw ghax-xoghol tahtkom u llum qed jaqighu x'jeklu, staqsi lil studenti li llum mhux qed ihallsu ghal ezamijiet u ghat-transport... Nahseb li m'ghandikx mera id-dar.....

Subjectivity: True

Sentiment Polarity: Positive

Emotion: Trust

Irony: False

Sarcasm: True

Data Source: Times of Malta, Year: 2019

Appendix G

Human Evaluation - Information About Systems

Please read the following information about the Opinion Search and Summarisation System: See Appendix D.

Steps to access the Opinion Search and Summarisation System - Malta Government Budget

- Go to: <https://opinion-miner.azurewebsites.net/>
- Log in with the credentials provided - username and password

In total, the following three systems shall be evaluated for the purposes of this study:

1. System 1 - Opinion Search based on Sentiment Polarity (One Dimensional - Simple Search)
2. System 2 - Opinion Search based on Sentiment Polarity and Emotion (Two Dimensional)
3. System 3 - Opinion Search based on Sentiment Polarity, Emotion, and Irony or Sarcasm (Three Dimensional)

Note: The Opinion Search tasks defined for Systems 1-3 (see below for more information) are based on “subjective” online posts i.e., the ones that reflect an opinion by the

end-user about the Malta Government Budget. For this reason, “objective” online posts shall not be part of this study.

Appendix H

Human Evaluation - Topic-based Tasks

Group 1 - Topic 1: Transport

System 1 - Opinion Search: Sentiment Polarity

Steps to follow to carry out the task below:

- Refresh the **Opinion Search and Summarisation System - Malta Government Budget** web page.
- Filter online posts by “SUBJECTIVITY” i.e., to search on subjective posts only.
- Click on the “TRANSPORT” topic and then also click on the “Students” sub-topic (see image below) under the “TRANSPORT” topic sub-menu.

TOPICS

ENVIRONMENT (1) ▾

TAXATION (12) ▾

TRANSPORT (45) ▾

- Alternative Methods (1)
- Traffic Pollution (1)
- Elderly (3)
- Private Cars (1)
- Youths (8)
- Ferry Transport (1)
- Gozo Link (2)
- School Transport (12)
- Pensioners (1)
- Commercial Vehicles (1)
- Imports (1)
- Pedelec (1)
- Public Transport (15)
- People With Disability (1)
- Students (8)
- Bicycle (6)

- View the “SENTIMENT POLARITY” pie chart (opinion visualisation). You can see the percentage of online posts broken down by sentiment polarity values (i.e., Positive, Negative) by hovering the mouse on the pie chart.
- Filter online posts by the **Positive** “SENTIMENT POLARITY” value in the left side bar (see image below). This will return search results with online posts having a *positive* sentiment polarity only.

OPINION DIMENSIONS

EMOTION ▾

SUBJECTIVITY (8)

SENTIMENT POLARITY ▾

- Positive (7)
- Negative (1)

SARCASM (1)

System 2 - Opinion Search: Sentiment and Emotion

Steps to follow to carry out the task below:

- Continue from your previous task.
- Deselect (press on the “[X]” text/button) the “Positive: True” filter (see image below).

FILTER RESULTS

Subjectivity: True [X]
Transport: True [X]
Students: True [X]
Positive: True [X]

- View the “SENTIMENT POLARITY” and “EMOTION” pie charts (opinion visualisation). You can see the percentage of online posts broken down by sentiment polarity values (i.e., Positive, Negative) and emotion values (i.e., Trust, Anger, Joy) by hovering the mouse on each respective pie chart.
- Filter online posts by the **Positive** “SENTIMENT POLARITY” value and **Trust** “EMOTION” value in the left side bar (see image below). This will return search results with online posts having a *positive* sentiment polarity and *trust* emotion only.

OPINION DIMENSIONS

EMOTION ▾
• Trust (2)
• Anger (1)
• Joy (5)
SUBJECTIVITY (8)
SENTIMENT POLARITY ▾
• Positive (7)
• Negative (1)
SARCASM (1)

System 3 - Opinion Search: Sentiment, Emotion, and Irony or Sarcasm

Steps to follow to carry out the task below:

- Continue from your previous task.
- Deselect (press on the “[X]” text/button) the “Positive: True” and “Trust: True” filters (see image below).

FILTER RESULTS

Subjectivity: True [X]
Transport: True [X]
Students: True [X]
Positive: True [X]
Trust: True [X]

- View the “SENTIMENT POLARITY”, “EMOTION” and “SARCASM” pie charts (opinion visualisation). You can see the percentage of online posts broken down by sentiment polarity values (i.e., Positive, Negative), emotion values (i.e., Trust, Anger, Joy), and sarcasm by hovering the mouse on each respective pie chart.

- Filter online posts by the **Positive** “SENTIMENT POLARITY” value, **Trust** “EMOTION” value, and “SARCASM” (**True**) value in the left side bar (see image below). This will return search results with online posts having a *positive* sentiment polarity, *trust* emotion, and sarcasm only.

OPINION DIMENSIONS

EMOTION ▼

- Trust (2)
- Anger (1)
- Joy (5)

SUBJECTIVITY (8)

SENTIMENT POLARITY ▼

- Positive (7)
- Negative (1)

SARCASM (1)

Group 2 - Topic 2: Environment

System 1 - Opinion Search: Sentiment Polarity

Steps to follow to carry out the task below:

- Refresh the **Opinion Search and Summarisation System - Malta Government Budget** web page.
- Filter online posts by “SUBJECTIVITY” i.e., to search on subjective posts only.
- Click on the “ENVIRONMENT” topic and then also click on the “Electric Cars” sub-topic (see image below) under the “ENVIRONMENT” topic sub-menu.

TOPICS

ENVIRONMENT (73) ▼

- Recycling (1)
- Environmental Sustainability (5)
- Bottle Refund (3)
- Nature (1)
- Waste Technology (1)
- Park (1)
- Urban Garden (2)
- Machinery (5)
- Bottles (1)
- Plastic (1)
- Development (4)
- Zero Emission Vehicles (1)
- Trees (2)
- Culture Change (1)
- Waste (3)
- Hydrogen Cars (1)
- Green Corners (1)
- Biodiversity (1)
- Carbon (1)
- Concrete (1)
- Electric Cars (4)
- Pollution (1)

- View the “SENTIMENT POLARITY” pie chart (opinion visualisation). You can see the percentage of online posts broken down by sentiment polarity values (i.e., Neutral, Negative) by hovering the mouse on the pie chart.
- Filter online posts by the **Neutral** “SENTIMENT POLARITY” value in the left side bar (see image below). This will return search results with online posts having a *neutral* sentiment polarity only.

OPINION DIMENSIONS

EMOTION ▼

SUBJECTIVITY (4)

SENTIMENT POLARITY ▼

- Neutral (2)
- Negative (2)

SARCASM (1)

System 3 - Opinion Search: Sentiment, Emotion, and Irony or Sarcasm

Steps to follow to carry out the task below:

- Continue from your previous task.
- Deselect (press on the “[X]” text/button) the “Neutral: True” filter (see image below).

FILTER RESULTS

Environment: True [X]

Subjectivity: True [X]

Electric cars: True [X]

Neutral: True [X]

- View the “SENTIMENT POLARITY”, “EMOTION” and “SARCASM” pie charts (opinion visualisation). You can see the percentage of online posts broken down by sentiment polarity values (i.e., Neutral, Negative), emotion values (i.e., Disgust, Fear, Sadness, Anticipation), and sarcasm by hovering the mouse on each respective pie chart.
- Filter online posts by the **Neutral** “SENTIMENT POLARITY” value, **Anticipation** “EMOTION” value, and “SARCASM” (**True**) value in the left side bar (see image below). This will return search results with online posts having a *neutral* sentiment polarity, *anticipation* emotion, and sarcasm only.

OPINION DIMENSIONS

EMOTION ▼

- Disgust (1)
- Fear (1)
- Sadness (1)
- Anticipation (1)

SUBJECTIVITY (4)

SENTIMENT POLARITY ▼

- Neutral (2)
- Negative (2)

SARCASM (1)

System 2 - Opinion Search: Sentiment and Emotion

Steps to follow to carry out the task below:

- Continue from your previous task.
- Deselect (press on the “[X]” text/button) the “Neutral: True”, “Anticipation: True” and “Sarcasm: True” filters (see image below).

FILTER RESULTS

Subjectivity: True [X]
Environment: True [X]
Electric cars: True [X]
Neutral: True [X]
Anticipation: True [X]
Sarcasm: True [X]

- View the “SENTIMENT POLARITY” and “EMOTION” pie charts (opinion visualisation). You can see the percentage of online posts broken down by sentiment polarity values (i.e., Neutral, Negative) and emotion values (i.e., Disgust, Fear, Sadness, Anticipation) by hovering the mouse on each respective pie chart.
- Filter online posts by the **Neutral** “SENTIMENT POLARITY” value and **Anticipation** “EMOTION” value in the left side bar (see image below). This will return search results with online posts having a *neutral* sentiment polarity and *anticipation* emotion only.

OPINION DIMENSIONS

EMOTION ▼
• Disgust (1)
• Fear (1)
• Sadness (1)
• Anticipation (1)
SUBJECTIVITY (4)
SENTIMENT POLARITY ▼
• Neutral (2)
• Negative (2)
SARCASM (1)

Group 3 - Topic 3: Taxation

System 2 - Opinion Search: Sentiment and Emotion

Steps to follow to carry out the task below:

- Refresh the **Opinion Search and Summarisation System - Malta Government Budget** web page.
Filter online posts by “SUBJECTIVITY” i.e., to search on subjective posts only.
- Click on the “TAXATION” topic and then also click on the “Income Tax” sub-topic (see image below) under the “TAXATION” topic sub-menu.

TAXATION (567) 

- Cars (7)
- Married Rate (1)
- Milk (5)
- Sports (1)
- Overtime (4)
- Cigarettes (10)
- Indirect Tax (1)
- Rebate (3)
- Self Employed (1)
- Foreign Tax (1)
- Utility Bills (9)
- Rental (2)
- Tax Credit (2)
- Pension (12)
- Bread (3)
- Grants (3)
- Single Rate (1)
- Wages (1)
- Congestion (1)
- Alcohol (5)
- VAT (7)
- Tax Evasion (3)
- Grant (4)
- Music (2)
- Exemption (2)
- Shares (1)
- Fuel (12)
- **Income Tax (32)**

- View the “SENTIMENT POLARITY” and “EMOTION” pie charts (opinion visualisation). You can see the percentage of online posts broken down by sentiment polarity values (i.e., Positive, Negative, Neutral) and emotion values (i.e., Disgust, Trust, Fear, Surprise, Anger, Joy, Anticipation) by hovering the mouse on each respective pie chart.
- Filter online posts by the **Negative** “SENTIMENT POLARITY” value and **Disgust** “EMOTION” value in the left side bar (see image below). This will return search results with online posts having a *negative* sentiment polarity and *disgust* emotion only.

OPINION DIMENSIONS

EMOTION 

- **Disgust (3)**
- Trust (4)
- Fear (1)
- Surprise (2)
- Anger (7)
- Joy (7)
- Anticipation (8)

SUBJECTIVITY (32)

IRONY (1)

SENTIMENT POLARITY 

- Neutral (9)
- Positive (14)
- **Negative (9)**

SARCASM (2)

System 1 - Opinion Search: Sentiment Polarity

Steps to follow to carry out the task below:

- Continue from your previous task.
- Deselect (press on the “[X]” text/button) the “Negative: True” and “Disgust: True” filters (see image below).

FILTER RESULTS

Subjectivity: True [X]
Taxation: True [X]
Income tax: True [X]
Negative: True [X]
Disgust: True [X]

- View the “SENTIMENT POLARITY” pie chart (opinion visualisation). You can see the percentage of online posts broken down by sentiment polarity values (i.e., Positive, Negative, Neutral) by hovering the mouse on the pie chart.
- Filter online posts by the **Negative** “SENTIMENT POLARITY” value in the left side bar (see image below). This will return search results with online posts having a *negative* sentiment polarity only.

OPINION DIMENSIONS

EMOTION ▾
SUBJECTIVITY (32)
IRONY (1)
SENTIMENT POLARITY ▾

- Neutral (9)
- Positive (14)
- Negative (9)

SARCASM (2)

System 3 - Opinion Search: Sentiment, Emotion, and Irony or Sarcasm

Steps to follow to carry out the task below:

- Continue from your previous task.
- Deselect (press on the “[X]” text/button) the “Negative: True” filter (see image below).

FILTER RESULTS

Subjectivity: True [X]
Taxation: True [X]
Income tax: True [X]
Negative: True [X]

- View the “SENTIMENT POLARITY”, “EMOTION” and “IRONY” pie charts (opinion visualisation). You can see the percentage of online posts broken down by sentiment polarity values (i.e., Positive, Neutral, Negative), emotion values (i.e., Disgust, Trust, Fear, Surprise, Anger, Joy, Anticipation), and irony by hovering the mouse on each respective pie chart.

- Filter online posts by the **Negative** “SENTIMENT POLARITY” value, **Disgust** “EMOTION” value, and “IRONY” (**True**) value in the left side bar (see image below). This will return search results with online posts having a *negative* sentiment polarity, *disgust* emotion, and irony only.

OPINION DIMENSIONS

EMOTION ▾

- Disgust (3)
- Trust (4)
- Fear (1)
- Surprise (2)
- Anger (7)
- Joy (7)
- Anticipation (8)

SUBJECTIVITY (32)

IRONY (1)

SENTIMENT POLARITY ▾

- Neutral (9)
- Positive (14)
- Negative (9)

SARCASM (2)

Group 4 - Topic 1: Transport

System 2 - Opinion Search: Sentiment and Emotion

Steps to follow to carry out the task below:

- Refresh the **Opinion Search and Summarisation System - Malta Government Budget** web page.
- Filter online posts by “SUBJECTIVITY” i.e., to search on subjective posts only.
- Click on the “TRANSPORT” topic and then also click on the “Students” sub-topic (see image below) under the “TRANSPORT” topic sub-menu.

TOPICS

ENVIRONMENT (1) ▾

TAXATION (12) ▾

TRANSPORT (45) ▾

- Alternative Methods (1)
- Traffic Pollution (1)
- Elderly (3)
- Private Cars (1)
- Youths (8)
- Ferry Transport (1)
- Gozo Link (2)
- School Transport (12)
- Pensioners (1)
- Commercial Vehicles (1)
- Imports (1)
- Pedelec (1)
- Public Transport (15)
- People With Disability (1)
- Students (8)
- Bicycle (6)

- View the “SENTIMENT POLARITY” and “EMOTION” pie charts (opinion visualisation). You can see the percentage of online posts broken down by sentiment polarity values (i.e., Positive, Negative) and emotion values (i.e., Trust, Anger, Joy) by hovering the mouse on each respective pie chart.
- Filter online posts by the **Positive** “SENTIMENT POLARITY” value and **Trust** “EMOTION” value in the left side bar (see image below). This will return search results with online posts having a *positive* sentiment polarity and *trust* emotion only.

OPINION DIMENSIONS

EMOTION ▾

- Trust (2)
- Anger (1)
- Joy (5)

SUBJECTIVITY (8)

SENTIMENT POLARITY ▾

- Positive (7)
- Negative (1)

SARCASM (1)

System 3 - Opinion Search: Sentiment, Emotion, and Irony or Sarcasm

Steps to follow to carry out the task below:

- Continue from your previous task.
- Deselect (press on the “[X]” text/button) the “Positive: True” and “Trust: True” filters (see image below).

FILTER RESULTS

Subjectivity: True [X]

Transport: True [X]

Students: True [X]

Positive: True [X]

Trust: True [X]

- View the “SENTIMENT POLARITY”, “EMOTION” and “SARCASM” pie charts (opinion visualisation). You can see the percentage of online posts broken down by sentiment polarity values (i.e., Positive, Negative), emotion values (i.e., Trust, Anger, Joy), and sarcasm by hovering the mouse on each respective pie chart.
- Filter online posts by the **Positive** “SENTIMENT POLARITY” value, **Trust** “EMOTION” value, and “SARCASM” (**True**) value in the left side bar (see image below). This will return search results with online posts having a *positive* sentiment polarity, *trust* emotion, and sarcasm only.

OPINION DIMENSIONS

EMOTION ▾

- Trust (2)
- Anger (1)
- Joy (5)

SUBJECTIVITY (8)

SENTIMENT POLARITY ▾

- Positive (7)
- Negative (1)

SARCASM (1)

System 1 - Opinion Search: Sentiment Polarity

Steps to follow to carry out the task below:

- Continue from your previous task.
- Deselect (press on the “[X]” text/button) the “Positive: True”, “Trust: True”, and “Sarcasm: True” filters (see image below).

FILTER RESULTS

Subjectivity: True [X]

Transport: True [X]

Students: True [X]

Positive: True [X]

Trust: True [X]

Sarcasm: True [X]

- View the “SENTIMENT POLARITY” pie chart (opinion visualisation). You can see the percentage of online posts broken down by sentiment polarity values (i.e., Positive, Negative) by hovering the mouse on the pie chart.
- Filter online posts by the **Positive** “SENTIMENT POLARITY” value in the left side bar (see image below). This will return search results with online posts having a *positive* sentiment polarity only.

OPINION DIMENSIONS

EMOTION ▼

SUBJECTIVITY (8)

SENTIMENT POLARITY ▼

• Positive (7)

• Negative (1)

SARCASM (1)

Group 5 - Topic 2: Environment

System 3 - Opinion Search: Sentiment, Emotion, and Irony or Sarcasm

Steps to follow to carry out the task below:

- Refresh the **Opinion Search and Summarisation System - Malta Government Budget** web page.
- Filter online posts by “SUBJECTIVITY” i.e., to search on subjective posts only.
- Click on the “ENVIRONMENT” topic and then also click on the “Electric Cars” sub-topic (see image below) under the “ENVIRONMENT” topic sub-menu.

TOPICS

ENVIRONMENT (73)

- Recycling (1)
- Environmental Sustainability (5)
- Bottle Refund (3)
- Nature (1)
- Waste Technology (1)
- Park (1)
- Urban Garden (2)
- Machinery (5)
- Bottles (1)
- Plastic (1)
- Development (4)
- Zero Emission Vehicles (1)
- Trees (2)
- Culture Change (1)
- Waste (3)
- Hydrogen Cars (1)
- Green Corners (1)
- Biodiversity (1)
- Carbon (1)
- Concrete (1)
- **Electric Cars (4)**
- Pollution (1)

- View the “SENTIMENT POLARITY”, “EMOTION” and “SARCASM” pie charts (opinion visualisation). You can see the percentage of online posts broken down by sentiment polarity values (i.e., Neutral, Negative), emotion values (i.e., Disgust, Fear, Sadness, Anticipation), and sarcasm by hovering the mouse on each respective pie chart.
- Filter online posts by the **Neutral** “SENTIMENT POLARITY” value, **Anticipation** “EMOTION” value, and “SARCASM” (**True**) value in the left side bar (see image below). This will return search results with online posts having a *neutral* sentiment polarity, *anticipation* emotion, and sarcasm only.

OPINION DIMENSIONS

EMOTION

- Disgust (1)
- Fear (1)
- Sadness (1)
- **Anticipation (1)**

SUBJECTIVITY (4)

SENTIMENT POLARITY

- **Neutral (2)**
- Negative (2)

SARCASM (1)

System 1 - Opinion Search: Sentiment Polarity

Steps to follow to carry out the task below:

- Continue from your previous task.
- Deselect (press on the “[X]” text/button) the “Neutral: True”, “Anticipation: True” and “Sarcasm: True” filters (see image below).

FILTER RESULTS

Subjectivity: True [X]
Environment: True [X]
Electric cars: True [X]
Neutral: True [X]
Anticipation: True [X]
Sarcasm: True [X]

- View the “SENTIMENT POLARITY” pie chart (opinion visualisation). You can see the percentage of online posts broken down by sentiment polarity values (i.e., Neutral, Negative) by hovering the mouse on the pie chart.
- Filter online posts by the **Neutral** “SENTIMENT POLARITY” value in the left side bar (see image below). This will return search results with online posts having a *neutral* sentiment polarity only.

OPINION DIMENSIONS

EMOTION ▾
SUBJECTIVITY (4)
SENTIMENT POLARITY ▾
• Neutral (2)
• Negative (2)
SARCASM (1)

System 2 - Opinion Search: Sentiment and Emotion

Steps to follow to carry out the task below:

- Continue from your previous task.
- Deselect (press on the “[X]” text/button) the “Neutral: True” filter (see image below).

FILTER RESULTS

Environment: True [X]
Subjectivity: True [X]
Electric cars: True [X]
Neutral: True [X]

- View the “SENTIMENT POLARITY” and “EMOTION” pie charts (opinion visualisation). You can see the percentage of online posts broken down by sentiment polarity values (i.e., Neutral, Negative) and emotion values (i.e., Disgust, Fear, Sadness, Anticipation) by hovering the mouse on each respective pie chart.
- Filter online posts by the **Neutral** “SENTIMENT POLARITY” value and **Anticipation** “EMOTION” value in the left side bar (see image below). This will return search results with online posts having a *neutral* sentiment polarity and *anticipation* emotion only.

OPINION DIMENSIONS

EMOTION ▾
• Disgust (1)
• Fear (1)
• Sadness (1)
• Anticipation (1)
SUBJECTIVITY (4)
SENTIMENT POLARITY ▾
• Neutral (2)
• Negative (2)
SARCASM (1)

Group 6 - Topic 3: Taxation

System 3 - Opinion Search: Sentiment, Emotion, and Irony or Sarcasm

Steps to follow to carry out the task below:

- Refresh the **Opinion Search and Summarisation System - Malta Government Budget** web page.
Filter online posts by “SUBJECTIVITY” i.e., to search on subjective posts only.
- Click on the “TAXATION” topic and then also click on the “Income Tax” sub-topic (see image below) under the “TAXATION” topic sub-menu.

TAXATION (567) 

- Cars (7)
- Married Rate (1)
- Milk (5)
- Sports (1)
- Overtime (4)
- Cigarettes (10)
- Indirect Tax (1)
- Rebate (3)
- Self Employed (1)
- Foreign Tax (1)
- Utility Bills (9)
- Rental (2)
- Tax Credit (2)
- Pension (12)
- Bread (3)
- Grants (3)
- Single Rate (1)
- Wages (1)
- Congestion (1)
- Alcohol (5)
- VAT (7)
- Tax Evasion (3)
- Grant (4)
- Music (2)
- Exemption (2)
- Shares (1)
- Fuel (12)
- **Income Tax (32)**

- View the “SENTIMENT POLARITY”, “EMOTION” and “IRONY” pie charts (opinion visualisation). You can see the percentage of online posts broken down by sentiment polarity values (i.e., Positive, Neutral, Negative), emotion values (i.e., Disgust, Trust, Fear, Surprise, Anger, Joy, Anticipation), and irony by hovering the mouse on each respective pie chart.
- Filter online posts by the **Negative** “SENTIMENT POLARITY” value, **Disgust** “EMOTION” value, and “IRONY” (**True**) value in the left side bar (see image below). This will return search results with online posts having a *negative* sentiment polarity, *disgust* emotion, and irony only.

OPINION DIMENSIONS

EMOTION

- Disgust (3)
- Trust (4)
- Fear (1)
- Surprise (2)
- Anger (7)
- Joy (7)
- Anticipation (8)

SUBJECTIVITY (32)

IRONY (1)

SENTIMENT POLARITY

- Neutral (9)
- Positive (14)
- Negative (9)

SARCASM (2)

System 2 - Opinion Search: Sentiment and Emotion

Steps to follow to carry out the task below:

- Continue from your previous task.
- Deselect (press on the “[X]” text/button) the “Negative: True”, “Disgust: True” and “Irony: True” filters (see image below).

FILTER RESULTS

Subjectivity: True

Taxation: True

Income tax: True

Negative: True

Disgust: True

Irony: True

- View the “SENTIMENT POLARITY” and “EMOTION” pie charts (opinion visualisation). You can see the percentage of online posts broken down by sentiment polarity values (i.e., Positive, Negative, Neutral) and emotion values (i.e., Disgust, Trust, Fear, Surprise, Anger, Joy, Anticipation) by hovering the mouse on each respective pie chart.
- Filter online posts by the **Negative** “SENTIMENT POLARITY” value and **Disgust** “EMOTION” value in the left side bar (see image below). This will return search results with online posts having a *negative* sentiment polarity and *disgust* emotion only.

OPINION DIMENSIONS

EMOTION

- Disgust (3)
- Trust (4)
- Fear (1)
- Surprise (2)
- Anger (7)
- Joy (7)
- Anticipation (8)

SUBJECTIVITY (32)

IRONY (1)

SENTIMENT POLARITY

- Neutral (9)
- Positive (14)
- Negative (9)

SARCASM (2)

System 1 - Opinion Search: Sentiment Polarity

Steps to follow to carry out the task below:

- Continue from your previous task.
- Deselect (press on the “[X]” text/button) the “Negative: True” and “Disgust: True” filters (see image below).

FILTER RESULTS

Subjectivity: True [X]

Taxation: True [X]

Income tax: True [X]

Negative: True [X]

Disgust: True [X]

- View the “SENTIMENT POLARITY” pie chart (opinion visualisation). You can see the percentage of online posts broken down by sentiment polarity values (i.e., Positive, Negative, Neutral) by hovering the mouse on the pie chart.
- Filter online posts by the **Negative** “SENTIMENT POLARITY” value in the left side bar (see image below). This will return search results with online posts having a *negative* sentiment polarity only.

OPINION DIMENSIONS

EMOTION

SUBJECTIVITY (32)

IRONY (1)

SENTIMENT POLARITY

- Neutral (9)
- Positive (14)
- Negative (9)

SARCASM (2)

Appendix I

Human Evaluation - Opinion

Summarisation Results'

Questionnaire

The objective of the questions below is to determine whether the results of [System 1 / System 2 / System 3] to be used for policy formulation, policy-making, decision-making, and decision-taking, reflect the summary of opinions of the general public.

***Required**

1. Did you manage to complete all the steps for [System 1 / System 2 / System 3]? *

(Please mark only one)

Yes

No

2. Is the opinion summary based on SENTIMENT POLARITY (System 1) / SENTIMENT POLARITY and EMOTION (System 2) / SENTIMENT POLARITY, EMOTION, and IRONY/SARCASM (System 3) useful? * (Please mark only one)

Excellent

Good

- Fair
- Poor
- Bad

3. Are the online posts provided as part of the SENTIMENT POLARITY (System 1) / SENTIMENT POLARITY and EMOTION (System 2) / SENTIMENT POLARITY, EMOTION, and IRONY/SARCASM (System 3) opinion summary useful? * (Please mark only one)

- Excellent
- Good
- Fair
- Poor
- Bad

4. Are the data sources of the online posts provided as part of the SENTIMENT POLARITY (System 1) / SENTIMENT POLARITY and EMOTION (System 2) / SENTIMENT POLARITY, EMOTION, and IRONY/SARCASM (System 3) opinion summary useful? * (Please mark only one)

- Excellent
- Good
- Fair
- Poor
- Bad

5. Is the opinion summary based on SENTIMENT POLARITY (System 1) / SENTIMENT POLARITY and EMOTION (System 2) / SENTIMENT POLARITY, EMOTION, and IRONY/SARCASM (System 3) informative? * (Please mark only one)

- Excellent

Good

Fair

Poor

Bad

Appendix J

Human Evaluation - Generic Opinion Search and Summarisation System Usability

Please fill in the following system usability questionnaire to evaluate the generic usability of the Opinion Search and Summarisation tool across System 1, System 2, and System 3, to be used for purposes of policy formulation, policy-making, decision-making, and decision-taking.

***Required** (Please mark only one for each question)

*

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I think that I would like to use this system frequently.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I found the system unnecessarily complex.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I thought the system was easy to use.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I think that I would need the support of a technical person to be able to use this system.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I found the various functions in this system were well integrated.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I thought there was too much inconsistency in this system.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would imagine that most people would learn to use this system very quickly.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I found the system very cumbersome to use.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt very confident using the system.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I needed to learn a lot of things before I could get going with this system.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Appendix K

Human Evaluation - Other Questions

- Do you have any comments on any of the systems?

- Do you have any feedback or suggestions?

Appendix L

Human Evaluation - Subject Matter Expert Interview Questions

1. In your opinion, from Systems 1-3 which one best reflects the summary of opinions of the general public for purposes of policy formulation, policy-making, decision-making, and decision-taking?
2. As a potential end-user of the Opinion Search and Summarisation system, would you prefer using the full-text search or the faceted search (i.e., filter by topic/topic and sub-topic based search) feature to obtain the relevant opinion summaries? Please state why?
3. As a Subject Matter Expert in the Government domain (policy-maker, Chief Information Officer, Chief Technology Officer, etc.), in which cases/projects/initiatives would you find such a system useful, and in terms of potential use cases/entities who would benefit from using this system/integration with existing systems?
4. Do you have any feedback or suggestions on the functionality of the system?