How Emotions Can Help Detect Synthetic Text

Alan Cowap
M.Sc., B.Sc(Hons), Dip.App.Sc., Grad.Cert.ICT

A thesis submitted for the Award of Doctor of Philosophy
(Ph.D.)

School of Computing
Dublin City University

Supervisors:
Dr. Jennifer Foster
Dr. Yvette Graham (Trinity College Dublin, Ireland)

April 2025

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 (Ph.D.) 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: Alan Cowap

ID No.: 19213953

Date: 11-April-2025

Dedication

I dedicate this work to those who provide positive inspiration - especially those artists who create inspirational science fiction. In particular from Star Trek: Gene Roddenberry, Leonard Nimoy (Mr Spock), Brent Spiner (Data); and from Terminator: James Cameron, Arnold Schwarzenegger (T-101), Robert Patrick (T-1000), and Linda Hamilton (Sarah Connor).

Acknowledgements

Firstly, I thank my supervisors Jennifer Foster and Yvette Graham. Thank you for your support and guidance throughout my PhD journey. I am extremely fortunate to have benefitted from your expertise and experience. I will miss our conversations and your insightful perspectives.

Thanks to Mark Keane and Yalemisew Abgaz for their thorough and challenging examination in the viva-voce, which was tough and enjoyable. Thanks to Silvana MacMahon for chairing the viva and ensuring there was "no eye-gouging". Thank you to Alan Smeaton and Anya Belz for their examination and valuable feedback in the PhD Transfer.

Thanks to the DCU NLP and NLG groups where we shared many interesting and informative discussions and presentations. I extend my thanks to my co-authors and collaborators and am delighted our work on gaBERT was presented at LREC 2022. I am grateful for the opportunity to present my work at EMNLP 2023, and meet colleagues at ACL 2022 and AICS 2019. I extend my appreciation to all the authors whose work I cite in this thesis, some of whom I shared conversations in person and some virtually. Particular thanks to Rowan Zellers for permitting the use of (human-authored) articles from his *RealNews-test* dataset in our *NEWSsynth* dataset. Thank you to the DCU Graduate Studies Office, particularly Jonny Hobson and Joseph Stokes, for their support, encouragement and understanding. Thanks to ADAPT TechOps for keeping the high-performance computing cluster running and for quality tech-support.

Thank you to the Research Ireland Centre for Research Training in Machine Learning (ML-Labs) who funded this research. Thank you to the Executive of ML-Labs for running an exceptional PhD programme which was run professionally and never lost sight of the people at it's heart. Particular thanks to Noel O'Connor and Angela Lally whose support went above and beyond what anyone could reasonably expect. Thank you to my friends and colleagues in ML-Labs and the other Centres for Research Training, particularly Stephen Carroll (CRT d-real) for organising the Writing Retreat which was the starting point for writing this dissertation. Heartfelt thanks to my friends and colleagues in the "Never-ending writing group" (especially Anastasia, Bahavathy and Agatha) for your friendship, support, and encouragement.

Thanks to my siblings and parents for your early influence - especially Mum for supporting and believing in me, and brother Tom for our thought-provoking discussions and allowing 10 year-old me to program your Sinclair ZX Spectrum.

Finally I would like to thank my family for their love and support: Kim for our many discussions on research challenges and sharing your experience on how to successfully navigate a research degree; Ryan, Aisling, and Roisin for being curious, challenging my thinking, and just for being you; I love you.

Contents

Glossary **x**

List of Tables **xv**

List of Figures **xxi**

1 Introduction **1**
- 1.1 Thesis Topic . . . 1
- 1.2 Research Questions . . . 5
- 1.3 Thesis Structure . . . 8
- 1.4 Publications . . . 13

2 Related Work **15**
- 2.1 Introduction . . . 15
- 2.2 Language Models . . . 16
  - 2.2.1 Language Models Continue to Improve, but Still Pose Challenges . . . 16
  - 2.2.2 Language Models are Dual-Use and Biased . . . 20
  - 2.2.3 Why do we Want to Detect Synthetic Text? . . . 22
  - 2.2.4 Detecting Synthetic Text is Difficult . . . 24
- 2.3 Human Factors . . . 27
  - 2.3.1 Human Cognition is Dual Process, Bimodal, and Biased . 27
  - 2.3.2 Emotions are a Particularly Human/Animal Trait . . . 30
  - 2.3.3 Can Emotions Help Detect Synthetic Text? . . . 33
  - 2.3.4 Classifying Emotions is Difficult. . . . 35
- 2.4 Measuring and Metrics . . . 39
- 2.5 Conclusion . . . 41

3 Improving Neural Detection of Synthetic Text via Emotion Recognition **43**
- 3.1 Method: Equipping PLMs with Emotional Intelligence . . . 43
- 3.2 Experiments in the News Domain . . . 45
  - 3.2.1 Generator and Detector Models . . . 45
  - 3.2.2 Datasets . . . 45
  - 3.2.3 Fine-Tuning `BERTsynth` . . . 47
  - 3.2.4 Fine-Tuning `emoBERT` . . . 47
  - 3.2.5 Fine-Tuning `emoBERTsynth` . . . 49
- 3.3 Results . . . 49
- 3.4 Analysis . . . 53

3.4.1 Length of Human vs Synthetic articles 53
3.4.2 Size of Fine-Tuning Splits 54
3.4.3 Alternative Forms of `emoBERT` 57
3.4.4 A Larger Detector Model 59
3.5 ChatGPT Experiments 60
3.5.1 *ChatGPT100* 60
3.5.2 Detection Task 61
3.5.3 Results 64
3.6 Conclusion 65

4 $\boldsymbol{\psi}$ - a Novel Evaluation Framework (for Dimensional Emotion Classifiers) 67
4.1 Overview 68
4.2 A Novel Evaluation Framework for Emotion Classification 69
4.2.1 Defining $\boldsymbol{\delta}$ - a Measure of Emotion Similarity 69
4.2.2 Defining $\phi(\delta)$ - a Flexible Loss Function 70
4.2.3 Defining $\boldsymbol{\psi}$ - a Novel Metric for Emotion Classification Performance 71
4.3 Toy Example 73
4.3.1 Demonstrating $\boldsymbol{\psi}$ 73
4.3.2 Comparing Metrics 76
4.4 Extending $\boldsymbol{\phi(\delta)}$ - a Flexible Loss Function 79
4.5 Conclusion 84

5 Developing and Selecting a Dimensional Emotion Classification Model 86
5.1 Method 87
5.2 Experimental Setup 87
5.2.1 Emotion Model 87
5.2.2 Dataset for Fine-Tuning Emotion Classifiers 88
5.2.3 Fine-Tuning Emotion Classifiers 91
5.3 Results 92
5.4 Analysis 94
5.4.1 Choosing an Emotion Detection Model 94
5.4.2 Correlating $\boldsymbol{\psi}$ with $\boldsymbol{F1_{\mu}}$, MAE, and RMSE 96
5.4.3 Zero-Shot NLI for Emotion Classification 100
5.5 Conclusion 103

6 Affective Profiling of Human and Synthetic Text 105
6.1 Method 106
6.2 Experiments 106
6.3 Results 109
6.4 Analysis 114
6.4.1 Affective Coherence (in Emotion N-grams) for Human and Synthetic News Articles 115
6.4.2 Affective Coherence (in Emotion Affective Profiles) for Human and Synthetic News Articles 117
6.4.3 Measuring Affective Coherence with $\boldsymbol{\psi}$ and $\boldsymbol{\delta}$-Profile . . . 119
6.5 Sentiment Experiments 123

6.5.1 Selecting a Sentiment Classifier 123
6.5.2 Affective Coherence (in Sentiment N-grams) for Human and Synthetic News Articles 127
6.5.3 Affective Coherence (in Sentiment Affective Profiles) for Human and Synthetic News Articles 127
6.6 Conclusion 133

7 Conclusion 136
7.1 Contributions 136
7.2 Research Questions Revisited 138
7.2.1 RQ1 - Can the performance of automatic synthetic text detectors be improved by making them 'emotionally-aware' compared to standard detectors? 138
7.2.2 RQ2 - Is it possible to develop a framework for measuring and evaluating classifiers of Plutchik's 8 emotions which takes account of their dimensional relationship? 139
7.2.3 RQ3 - Can we exploit the dimensional relationship between Plutchik's 8 emotions to enable optimal selection of an automatic emotion classifier? 141
7.2.4 RQ4 - Can we generate and compare the affective profile of human text and synthetic text? 142
7.2.5 Overarching Hypothesis Revisited 145
7.3 Limitations 145
7.4 Future Work 148
7.4.1 RQ1 149
7.4.2 RQ2 149
7.4.3 RQ3 150
7.4.4 RQ4 152
7.5 Concluding Remarks 153

A Reproducibility 1

B Hyperparameters Used for Fine-Tuning 4
B.1 `BERTsynth`, `emoBERTsynth` 4
B.2 `emoBERT` 5
B.3 BLOOM 5

C Emotions and Content in News Articles and *NEWSsynth* 6

D Static Word Embeddings and Dimensional Emotions 7

E Human Cognition and Emotion & Sentiment Distribution in News Headlines 10

F Affective Profile - Statistical Significance Tests 12

Glossary

affect heuristic "*Describes the importance of affect in guiding judgements and decisions*" (Slovic et al., 2007). 28, 35

affective coherence We introduce the term *affective coherence* to refer to the level of emotional consistency in a document. The greater the similarity between the emotions in a document, the greater the affective coherence of the document. See also *affective incoherence*. x, 26

affective deficit We introduce the term *affective deficit* to refer to the shortcomings of language models in relation to emotion. For example, language models are not driven by emotion (in the way humans are), and language models can have difficulty representing the relations between words which represent emotions. xxv, 33

affective incoherence We introduce the term *affective incoherence* to refer to the level of emotional inconsistency in a document. The greater the difference between the emotions in a document, the greater the affective incoherence of the document. See also *affective coherence*. x, xxv

affective profile We introduce the term *affective profile*. A representation of the affective content of a document (sequence of text). The affective profile could be, for example, a sequence of emotions within a document, a sequence of n-grams of emotions within a document, or combinations of such representations of the document. Thus, an affective profile could allow comparison, and potentially discrimination, between documents i.e. affective profiling. xxv

associative coherence Where the world behaves as System 1 expects it to behave. Interruptions to associative coherence, e.g. due to something surprising or unexpected, will cause System 2 to act. 28, 29

attention residue "*Attention residue reflects the persistence of cognitive activity about a Task A even though one stopped working on Task A and currently performs a Task B.*" (Leroy, 2009). See also *zeigarnik effect*. xiv, 30

attention restoration theory Concerned with the extent to which our reservoir of *directed attention* is consumed and restored in different environments e.g. urban, rural. (Kaplan, 2001). 30

availability bias Availability bias describes how our judgement of the probability of an event or the frequency of classes are evaluated by the ease with which examples are recalled (Tversky and Kahneman, 1973). Not to be confused with *availability cascade*. xi, 23, 29

availability cascade "*A self-reinforcing process of collective belief formation by which an expressed perception triggers a chain reaction that gives the perception increasing plausibility through its rising availability in public discourse.*" (Kuran and Sunstein, 1999). See also *endogeneity*. Not to be confused with *availability bias*. xi, xii

bad actor An actor (whether individual, state, or other) who intentionally spreads disinformation or malinformation. 20

cognition "*All forms of knowing and awareness, such as perceiving, conceiving, remembering, reasoning, judging, imagining, and problem solving. Along with affect and conation, it is one of the three traditionally identified components of mind*" - American Psychological Association [1]. 27

deliberation without attention effect Derived from *unconscious thought theory*, this effect shows that simple choices are better done with deliberation, but more complex choices are better done with unconscious thought (assuming you have gathered all relevant information). (Dijksterhuis et al., 2006). 29

directed attention Attention directed by cognitive control processes, it requires effort and a limited resource, for example doing a maths calculation. Also known as voluntary attention. Contrasts with involuntary attention. x

disinformation "*Fabricated or deliberately manipulated content. Intentionally created conspiracy theories or rumours*" (Wardle and Derakhshan, 2018). xi, 20, 21

distributional hypothesis Words which are synonyms (like "couch" or "sofa") tend to occur in the same context (like "sat" or "lay"), with the amount of difference in the words roughly corresponding to the difference in the context. Probably best summarised by Philosopher Ludwig Wittgenstein "The meaning of a word is its use in the language" (Wittgenstein, 1953, PI 43). This means we can create embeddings for a word based on the words surrounding that word. For example, two embeddings for "couch" might be "a man sat on the X" and "the dog lay on the X". Similar embeddings might exist for "sofa" which then suggests that "couch" and "sofa" have similar meaning. 17

dual process "*The dual process hypothesis postulates that performance and introspection reflect different underlying processes. It makes two fundamental assumptions: (1) The processes underlying the reasoning performance, e.g., matching bias, are not generally available for introspective report. (2) Introspective accounts of performance reflect a tendency for the subject to*

---

[1] `https://dictionary.apa.org/cognition/`

construct a justification for his own behavior consistent with his knowledge of the situation." (Wason and Evans, 1974). xiv, 27

endogeneity Also known as "reciprocal causation", the phenomenon whereby media reporting of an event impacts the event itself (Miller and Albert, 2015). See also *availability cascade*. xi

engagement based ranking Technique used by Facebook and other social media platforms to determine which content is most relevant to users interests. Highly problematic according to Facebook whistle-blower Haugen (2021). 23, 29

exposure bias The discrepancy where a generative language model is trained on the words from the training data distribution (human text) but not on the words from the model distribution (synthetic text), but at prediction time it is the model distribution that is used. Exposure bias was coined by (Ranzato et al., 2016). 21

human factors "*Ergonomics (or human factors) is the scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data, and methods to design in order to optimise human well-being and overall system performance. The terms ergonomics and human factors are often used interchangeably or as a unit (e.g., human factors and ergonomics – HFE or EHF), a practice that is adopted by the IEA* [International Ergonomics Association]"[2]. 22

inductive bias The set of assumptions that a language model makes in order to make predictions from inputs, unseen during the training phase, which can include the model parameters and hyper-parameters. Induction (in logic) is the process of inferring the general from specific examples. 21

information "*Verifiable information, in the public interest, which is what authentic journalism gives rise to*" (Wardle and Derakhshan, 2018). 21

involuntary attention Attention captured by stimuli which are intrinsically intriguing or important, for example a sudden noise or quickly moving object. Contrasts with voluntary or directed attention. xi

knowledge effect We tend to assume people know the same things we know, consequently we do not state the obvious (Hayes et al., 1986). This could lead to language models which are not trained on things which we consider obvious and do not state in writing (i.e. writing which is used to train language models), for example, that fire does not normally occur underwater. 21

malinformation "*Deliberate publication of private information for personal or corporate rather than public interest such as revenge porn. Deliberate change*

---

[2] `https://iea.cc/what-is-ergonomics/`

of context, date or time of genuine content" (Wardle and Derakhshan, 2018). xi, 20, 21

mere exposure effect "*Mere repeated exposure of the individual to a stimulus object enhances his attitude toward it*" (Zajonc, 1968). The original paper refers to perceptible stimuli, but a later paper includes subconscious or subliminal stimuli (Zajonc, 2001). 29

misinformation "*Unintentional mistakes such as inaccurate photo captions, dates, statistics, translations, or when satire is taken seriously*" (Wardle and Derakhshan, 2018). 21

negativity bias Negative events have a far greater effect on our psychological state than similar strength neutral or positive events. 23, 28

positivity effect A developmental pattern: "*Whereas negative material appears to be especially salient in youth, a shift evident in middle age and extending into old age appears to occur such that positive material is increasingly favored.*" Evidenced in both recollection, and fMRI scans showing amygdala activation. (Carstensen and Mikels, 2005). 29

publication bias "*Studies with significant or positive results were more likely to be published than those with non-significant or negative results.*" in research (Song et al., 2010). However, in news media negative news is more prevalent (Galtung and Ruge, 1965). 21

sampling bias The sample used is not representative of the population due to non-random means. 21

sampling strategy Given a prompt, an LM produces a probability for each token in its vocabulary as the output, the sampling strategy is the method to determine which token should be chosen. ArgMax chooses the token with the highest probability. Top-k chooses a token from the top-k highest probabilities, e.g. k=20 choose from the 20 highest probability tokens. Top-p chooses a token from the cumulative top-p probabilities. Random chooses a token at random, etc. 18, 33

somatic marker hypothesis Emotional processes guide decision making. We associate bodily feelings with emotions from specific situations that we were in. When we are subsequently in the same or similar situation, the recalled feelings from previous experience guide our present decision making. Patients with damage to the ventromedial pFC (prefrontal cortex) lack these markers and exhibit poor personal and social decision making (Damasio et al., 1996). 28

Surprise Negativity Anger Paradigm (SNAP) refers to the disproportionate use of surprise, negativity, and anger in communication. See preliminary experimental results in Appendix E. 23, 30

system 1 The intuitive process of human reasoning in *dual process theory*, also called Type 1, which is fast, instinctive and emotional. System 1 is automatic, and does not lend itself to introspection (Wason and Evans, 1974). 28, 35

system 2 The reasoned process of human reasoning in *dual process theory*, also called Type 2, which is slower, more deliberative, and more logical. System 2 is used for effortful mental activities, for example complex calculations (Wason and Evans, 1974). 28, 35

unconscious thought theory UTT distinguishes between two modes of thought, namely conscious thought and unconscious thought, which have different characteristics that make each mode preferable in different circumstances. (Dijksterhuis and Nordgren, 2006). xi, 28, 29

WEIRD An acronym for people from Western, Educated, Industrialised, Rich and Democratic societies (who tend to be over-represented in many studies but are not representative of the population (< 12%). (Henrich et al., 2010). 21

zeigarnik effect The ability of interrupted and incomplete tasks (that we want to complete) to dominate our attention, thus decreasing performance on the current task. See also *attention residue*. x, 30

List of Tables

2.1 An example of Natural Language Inference from the MultiNLI dataset (Williams et al., 2018). A *premise* and *hypothesis* are used as prompt to the NLI model, which responds with either *entailment, contradiction, or neutral.* In this example, `'entailment'` is the gold label, meaning that the hypothesis can be inferred from the premise. 37

2.2 An example of zero-shot NLI (§5.4.3). A *premise* (sentence) and *hypotheses* (emotions) are used as prompt to the model, which returns an *entailment* (prediction) i.e. a soft-maxed probability for each *hypothesis*. The *hypothesis* (emotion) with the highest probability is *entailed* (predicted) from the *premise* (sentence). In this example, `'joy'` is the prediction of the zero-shot NLI model. 38

2.3 Table showing Stevens (1946) 4 levels of Measurement. Ekmans 6 basic emotions are `Nominal` - simply named emotions, almost like a "*bag of emotions*". Plutchik's 8 emotions are `Ordinal` - named and ordered around the Plutchik Wheel. Our 'Enhanced Plutchik' is `Interval` - by including intervals between emotions - which we introduce in Chapter 4. Future Work could elevate our 'Enhanced Plutchik' to `Ratio` by including 'Neutral' emotion as absolute zero. 40

2.4 Table showing absolute value of ρ and the strength of relationship. 42

3.1 Emotion Mapping Schema: GoodNewsEveryone (15 emotions) to Ekman 6 basic emotions. % shows the emotion label distribution in the dataset. 46

3.2 Comparison of `BERTsynth` (Bs) and `emoBERTsynth` (emoBs) against the *NEWSsynth* test set. (Variance is shown in brackets under the mean). emoBs outperforms Bs in head-to-head for all 5 runs in Accuracy, F1, and Precision; while Bs outperforms emoBs in head-to-head for all 5 runs in Recall. 49

3.3 `emoBERTsynth` outperforms other model architectures and sizes detecting human and Grover$_{\text{mega}}$ (1.5B) synthetic text from the RealNews-Test dataset. Detector model sizes include 11M and 124M parameters and architectures include `FastText`, `GPT-2`$_{\text{base}}$, and `BERT`$_{\text{base}}$. The `FastText, GPT-2`$_{\text{base}}$ `and BERT`$_{\text{base}}$ results are reported by Zellers et al. (2019). 51

3.4 Comparison of Human and synthetic text in the *NEWSsynth* dataset showing the mean ($\overline{x}$) and standard deviation (σ) for Words Per Article, Sentences Per Article, and Words Per Sentence. Human articles are longer overall, but have slightly shorter sentences than synthetic text; and Human articles have more Sentences Per Article - which accounts for their longer mean length. 54

3.5 `BERTsynth` metrics for different split sizes, using the *NEWSsynth* dataset averaged over 5 runs (with variance shown in brackets). . 54

3.6 Ablation experiments, using different emotion datasets for fine-tuning `emoBERT`, comparing `emoBERTsynth` (eBs) detectors on the task of synthetic text detection on the *NEWSsynth* dataset. GNE is the GoodNewsEveryone dataset which is used in the main experiments. $\mathrm{GNE_r}$ is GNE with randomised labels. AT is AffectiveText. GA is GNE and AT combined. SST-2 is the SST-2 sentiment dataset. GAS is the combined GNE, AT, and SST-2 datasets. S-GA is first fine-tuned on sentiment using the SST-2 dataset, and then fine-tuned on emotion using the GNE and AT datasets, and finally fine-tuned on synthetic text detection. GAS+- is GAS but mapped to positive and negative sentiment. 59

3.7 Comparison of `BLOOMsynth` and `emoBLOOMsynth` against the *NEWSsynth* test set averaged over 5 runs (with variance in brackets). `emoBLOOMsynth` outperforms `BLOOMsynth` in Accuracy, F1, Recall, and Precision. 60

3.8 Human and ChatGPT articles from the *ChatGPT100* dataset. The Human article was retrieved on 20-Mar-2023 from the URL shown. The ChatGPT article was generated (using gpt-3.5-turbo-0314) on the same day using the prompt shown. 62

3.9 ChatGPT (gpt-3.5-turbo-0314) detecting Human and ChatGPT articles from *ChatGPT100*. The aricles were previously shown in Table 3.8 and are partially reproduced here. The classification was conducted on 21-Mar-2023. ChatGPT incorrectly predicts that the Human article is Synthetic, and that the Synthetic article is Human. 63

3.10 Our emotionally aware PLM (`emoBERTsynth`) outperforms ChatGPT and `BERTsynth` at detecting synthetic text in the *ChatGPT100* dataset. Note that ChatGPT is performing the task zero-shot. 64

4.1 Distance, δ, between Plutchik's Emotions. The closer two emotions are (according to Plutchik's Wheel) the lower their δ value. Note that, for Plutchik's 8 emotions, $\delta_{max} = 4$. 69

4.2 The score awarded to a prediction is determined by δ i.e. how close the predicted emotion is to the correct emotion on Plutchik's Wheel. We use a linear loss which is calculated as the quotient of the distance between the two emotions and the maximum possible distance between two emotions i.e. $\phi(\delta) = \frac{\delta}{\delta_{max}}$ 71

4.3 Toy example illustrating the difference between our novel ψ metric and existing metrics including Accuracy, MAE, and RMSE. Note that MAE is the mean δ value, RMSE is the square root of the mean δ^2 value. Accuracy is the quotient of the number of correct predictions and the total number of predictions, N. ψ is calculated per Equation 4.1. Most notable is that ψ is higher than Accuracy (62.50% vs 33.33%) - showing that ψ rewards predictions which may be "incorrect" but close to the correct emotion. This is evident in the $\frac{\delta}{\delta_{max}}$ column - which is effectively the loss of the prediction - because the losses are not all 1 when incorrect, they also include losses of only 0.75, 0.5, and 0.25. 73

4.4 Distribution of n_δ counts from the toy example in Table 4.3 as defined by ψ in Equation 4.1. The table shows that 4 examples had $\delta = 0$ (n_0), 2 examples had $\delta = 1$ (n_1) etc. This distribution of δ counts provides additional insight into the performance of the classifier. 79

5.1 Emotion Mapping Schema: GoodNewsEveryone (15 emotions) to Plutchik 8 emotions. % shows the emotion label distribution in the dataset. 89

5.2 Table showing the emotion counts for the GNE and AT datasets, combing those datasets (GA), and upsampling and downsampling to balance across all 8 Plutchik Emotions (GAB). 90

5.3 Table showing results for 15 emotion-classification PLMs and ensemble (best of 6 ensembles on Validation, see Table 5.4 for ensemble results) against the GAB test set, which were fine-tuned with X-Entropy loss, and validated on $F1_\mu$. Note that for MAE and RMSE lower scores are better. The table is sorted by $F1_\mu$ ascending. Best score in bold, second best score underlined. The mean and standard deviation (σ) are shown for each metric. . . . 93

5.4 Table showing ensembles of models against the validation set. The best performing ensemble, Ensemble-top5, was subsequently run against the test set as shown in Table 5.3. 94

5.5 Pearson correlation ρ between four metrics ψ, $F1_\mu$, MAE, and RMSE. 97

5.6 Spearman correlation ρ between four metrics ψ, $F1_\mu$, MAE, and RMSE. 99

5.7 An example of zero-shot NLI from DeBERTa-base on the GAB dataset (training split), using the "EMO-Name" prompt type from Plaza-del Arco et al. (2022). A *premise* (sentence) and *hypotheses* (emotions) are used as prompt to the model, which returns an *entailment* (prediction) i.e. a soft-maxed probability for each *hypothesis*. The *hypothesis* (emotion) with the highest probability is *entailed* (predicted) from the *premise* (sentence). In this example, `'joy'` is the prediction of the zero-shot NLI model, which agrees with the gold label from GAB. 101

5.8 Table (sorted by ψ ascending) comparing 15 emotion-classification zero-shot NLI models against the GAB Test set. Best score in bold, second best score underlined. Models with an asterisk (*) are provided by Laurer et al. (2024). The mean and standard deviation (σ) are shown for ψ and $F1_{\mu}$. 102

6.1 An example *affective profile*, 3-grams, and number of emotion changes; for a human article from the *NEWSsynth* dataset. The affective profile is parsed into 3-grams of emotion. The number of changes in each 3-gram is counted. We analyse the distribution of emotion changes in 3-grams and compare human versus synthetic, and random chance. We repeat the above process for 2, 3, 4, and 5-grams of emotion on the *NEWSsynth* (train, 10k) dataset. . . . 108

6.2 Statistics for percent emotion changes for human and synthetic articles in NEWSsynth (train 10k). The human and synthetic distributions are similar for mean, median, and σ (standard deviation), as illustrated in Figure 6.5. Nonetheless, the differences are statistically significant ($p < .001$). 117

6.3 An example *affective profile* and resulting δ-profile for a human article from the *NEWSsynth* dataset. The δ-profile corresponds to the δ between successive emotions. For example, 'fear'-'fear' in the affective profile results in '0' in the δ-profile; while 'fear'-'anticipation' results in '3' in the δ-profile. See Table 4.1 for the δ between Plutchik's emotions. 120

6.4 Table showing ψ and the strength of affective coherence. 121

6.5 Statistics for ψ (affective coherence) for human and synthetic articles in NEWSsynth (train 10k). The human and synthetic distributions are similar for mean, median, and σ (standard deviation), as illustrated in Figure 6.8. 122

6.6 Sentiment Mapping Schema: GoodNewsEveryone (15 emotions), and Affective text (6 emotions) mapped to binary sentiment. Count shows the sentiment label distribution in the dataset after balancing. * Surprise was removed from AT because we previously used DistilBERT-base to map surprise to sentiment, so testing DistilBERT-base on 'surprise' from AT would bias the results in favour of DistilBERT-base. 126

6.7 Results for sentiment classifiers tested against the (mapped to sentiment) GNE and AT-s (Affective Text minus 'surprise') datasets. TPR is True Positive Rate (aka Sensitivity, Recall), and TNR is True Negative Rate (aka Specificity). TPR, TNR, and Accuracy are reported. SiEBERT has a lower TNR, but outperforms $\text{DistilBERT}_{\text{base}}$ for Accuracy and TPR. 126

6.8 An example *affective profile* for binary sentiment of a human article and the corresponding synthetic article. The synthetic article was generated by Grover (using the human article as prompt). Both articles are taken from the *NEWSsynth* dataset. 'N' is negative sentiment, 'P' is positive sentiment. 130

6.9 Statistics for percent sentiment changes for human and synthetic articles in NEWSsynth (train 10k). The human and synthetic distributions are similar for mean, median, and σ (standard deviation), as illustrated in Figure 6.14. Nonetheless, the differences are statistically significant ($p < .001$). 130

6.10 Two patterns of results arise consistently across all RQ4 experiments. We refer to these patterns of results as PR1 and PR2. . . 134

6.11 Table showing our RQ4 experiments and their corresponding pattern or results. Each pattern of results (PR1 and PR2) is produced in three separate experiments, covering different classifiers, emotion models (Plutchik, sentiment), and experimental methodologies (n-grams, affective profile, δ-profile, ψ metric). 135

A.1 Domains used for human text in *ChatGPT100* dataset released with this paper. Articles were gathered between 16-24 March 2023. 3

D.1 Analogy predictions of Plutchik's Model and Word Embedding. Given the analogy on the left, what does Plutchik's model specify, and what do the Word Embedding vector calculations return e.g. $X = argmin\{D_c(sadness + fear - joy)\}$. 7

F.1 Chi-squared results for the distribution of number of emotion changes for emotion n-grams calculated from the *NEWSsynth* (train, 10k) dataset. n is the size of the n-gram. χ^2 is the chi-squared value. df is the degrees of freedom. p-value is the probability that the null hypothesis is true, a p-value $p < .05$ is considered statistically significant. All of the results are well below the .05 threshold, indicating that the difference between human and synthetic distributions of emotion changes in emotion n-grams is statistically significant. 13

F.2 Chi-squared results for the distribution of number of sentiment changes for sentiment n-grams calculated from the *NEWSsynth* (train, 10k) dataset. n is the size of the n-gram. χ^2 is the chi-squared value. df is the degrees of freedom. p-value is the probability that the null hypothesis is true, a p-value $p < .05$ is considered statistically significant. All of the results are well below the .05 threshold, indicating that the difference between human and synthetic distributions of sentiment changes in sentiment n-grams is statistically significant. 13

F.3 Kolmogorov-Smirnov Test results for the distribution of 'percent of emotion/sentiment maximum changes per article' calculated from the *NEWSsynth* (train, 10k) dataset. Both of the results are well below the $p < .05$ threshold, indicating that the difference between human and synthetic distributions is statistically significant - for both emotion and sentiment. 14

F.4 Distribution of number of emotion changes for emotion 2-grams calculated from the *NEWSsynth* (train, 10k) dataset. These are the raw figures which were shown as percentage values in Figure 6.1 15

F.5 Normalized distribution of number of emotion changes for emotion 2-grams calculated from the *NEWSsynth* (train, 10k) dataset. These are the values from Table F.4 above, but normalized to the approximate mid-point (117,115) of the Human and Synthetic totals. This normalization ensures the sample size variation does not impact the chi-squared results, and consequently normalized tables were used to calculate χ^2 and p-values shown in Figure F.1. 15
F.6 Distribution of number of emotion changes for emotion n-grams calculated from the *NEWSsynth* (train, 10k) dataset. These are the raw figures which were shown as percentage values in Figures 6.1 - 6.4 15
F.7 Distribution of number of sentiment changes for sentiment n-grams calculated from the *NEWSsynth* (train, 10k) dataset. These are the raw figures which were shown as percentage values in Figures 6.10 - 6.13 16

List of Figures

1.1 Two dimensional representation of the Truth/Falsehood of Human/Synthetic text. Our research focuses on the X-axis i.e. human versus synthetic text detection. 2

2.1 Plutchik Wheel of Emotion. The middle ring contains the 8 primary emotions, which includes Ekman's 6 emotions ('joy', 'sadness', 'fear', 'anger', 'disgust', 'surprise'.) plus 'trust' & 'anticipation'. Similar emotions lie on adjacent spokes e.g. anger-disgust, while opposing emotions are placed on opposing spokes e.g. joy-sadness. Image from (Plutchik, 2001), downloaded from `https://en.m.wikipedia.org/wiki/File:Plutchik-wheel.svg`. 32

3.1 The emotionally-aware PLM (`emoPLMsynth`) takes advantage of its prior fine-tuning on emotion to improve performance on the task of synthetic text detection. In contrast, the standard PLM fine-tuned only on synthetic text detection (`PLMsynth`) has no training on emotion. Our experiments show the emotionally-aware PLM (`emoPLMLsynth`) outperforms the standard PLM (`PLMsynth`) in multiple scenarios. 44

3.2 Combined Confusion Matrix for Emotion Classification on GoodNewsEveryone validation set. 48

3.3 Test results for `BERTsynth` and `emoBERTsynth` on the *NEWSsynth* dataset. `emoBERTsynth` is higher for Accuracy, Precision and F1, while `BERTsynth` is higher for Recall. 50

3.4 Scatter plot of number of words per article pair of synthetic text vs. human text in *NEWSsynth* (Pearson $r = 0.20$). 55

3.5 Scatter plot of number of sentences per article pair of synthetic text vs. human text in *NEWSsynth* (Pearson $r = 0.19$). 55

3.6 Histogram showing the number of words per article for human (green) and synthetic (orange) text in *NEWSsynth* (train and validation splits). Human articles tend to be longer ($\overline{x}$ = 595) than synthetic articles ($\overline{x}$=418), as shown by the 'right-shift' of the human (green) distribution compared to the synthetic (orange) distribution. Human articles have a much wider range of lengths (σ = 503) compared to synthetic articles (σ = 162), as shown by the longer tail of the human (green) distribution compared to the narrower synthetic (orange) distribution. 56

3.7 Histogram showing the number of sentences per article for human (green) and synthetic (orange) text in *NEWSsynth* (train and validation splits). Similar to Figure 3.6, human articles tend to be longer ($\overline{x} = 27$) and have a much wider range of lengths ($\sigma = 25$) than synthetic articles ($\overline{x} = 18$) and ($\sigma = 9$) respectively. 56
3.8 Histogram showing the number of words per sentence for human (green) and synthetic (orange) text in *NEWSsynth*. Human articles tend to have slightly fewer words per sentence ($\overline{x} = 22$) and a slightly narrower range of values ($\sigma = 16$) than synthetic articles ($\overline{x} = 23$) and ($\sigma = 17$) respectively. However, the human and synthetic distributions here are more similar than those in Figures 3.6 and 3.7. 57

4.1 Plutchik Wheel of Emotion (in blue) showing, as an example, the δ (in the green octagon) between 'Joy' and the other emotions. Rotating Plutchik's wheel of emotions so that a given emotion aligns with 0 on the δ-octagon will give the δ between that emotion and the other Plutchik emotions. 70
4.2 Plot of a linear loss function $\phi(\delta)$ vs δ for our toy example from Equation 4.7. This is the same loss function we use in later experiments and is suitable for Plutchik's 8 emotions. 80
4.3 Plot of a logistic loss function $\phi(\delta)$ vs δ using Equation 4.8. 82
4.4 Plot of a quadratic loss function $\phi(\delta)$ vs δ using Equation 4.9. . . 83

5.1 Scatter Plot of $F1_{\mu}$ vs ψ for 15 models and 1 ensemble on the GAB test set. XLM is the outlier. There is strong positive correlation between the metrics. 96
5.2 Scatter Plot of MAE vs ψ for 15 models and 1 ensemble on the GAB test set. XLM is the outlier. There is very strong negative correlation between the metrics. 97
5.3 Scatter Plot of RMSE vs ψ for 15 models and 1 ensemble on the GAB test set. XLM is the outlier. There is very strong negative correlation between the metrics. 98

6.1 The distribution of emotion changes for emotion 2-grams. Human text exhibits affective coherence (emo changes = 0) well above chance (40.07% vs 12.50%). Human text exhibits affective incoherence (emo changes = 1) well below chance (59.93% vs 87.50%). The results indicate human text is affectively coherent. However, synthetic text exhibits similar affective coherence to human text. Nonetheless, human text is more affectively coherent than synthetic text, and human text is less affectively incoherent than synthetic text, and this difference is statistically significant ($p < .001$). 110
6.2 The distribution of emotion changes for emotion 3-grams. Human text exhibits affective coherence (emo changes = 0) well above chance (20.90% vs 1.56%). Human text exhibits affective incoherence (emo changes = 2) well below chance (40.79% vs 76.56%). The human and synthetic distribution difference is significant ($p < .001$). The results are consistent with 2-grams in Figure 6.1. . . . 110

6.3 The distribution of emotion changes for emotion 4-grams. Human text exhibits affective coherence (emo changes = 0) well above chance (12.41% vs 0.20%). Human text exhibits affective incoherence (emo changes = 3) well below chance (22.17% vs 66.99%). The human and synthetic distribution difference is significant ($p < .001$). The results are consistent with 2-grams and 3-grams in Figures 6.1 and 6.2. 111

6.4 The distribution of emotion changes for emotion 5-grams. Human text exhibits affective coherence (emo changes = 0) well above chance (7.98% vs 0.02%). Human text exhibits affective incoherence (emo changes = 4) well below chance (18.73% vs 58.62%). The human and synthetic distribution difference is significant ($p < .001$). The results are consistent with 2-grams, 3-grams, and 4-grams in Figures 6.1 to 6.3. 111

6.5 Distribution of emotion changes per article for human and synthetic text in NEWSsynth (train 10k). The distributions are similar for human and synthetic text, peaking at just over 60%, as shown in Table 6.2. Nonetheless, the differences are statistically significant ($p < .001$) 118

6.6 Scatter plot of synthetic vs human percentage emotion changes per article - for a human article (used as prompt) and the corresponding article. There is weak correlation, with Pearson $\rho = 25.92$ and Spearman $\rho = 24.84$. 118

6.7 Distribution of δ values for human and synthetic articles in NEWSsynth (train 10k). The greater values for affective coherence ($\delta = 0$) and lower values for affective incoherence ($\delta = 4$), in addition to the similarity between human and synthetic distributions, are consistent with our earlier findings on affective profiles. 120

6.8 Distribution of ψ score per article for human and synthetic text in NEWSsynth (train 10k). The distributions are similar for human and synthetic text, with mean and median around 70%, and σ around 12, as in Table 6.5. 124

6.9 Scatter plot of synthetic vs human ψ score per article - for a human article (used as prompt) and the corresponding article. There is weak correlation, with Pearson $\rho = 31.19$ and Spearman $\rho = 31.65$. 124

6.10 The distribution of sentiment changes for sentiment 2-grams. Human text exhibits affective coherence (sentiment changes = 0) well above chance (66.51% vs 50.00%). Human text exhibits affective incoherence (sentiment changes = 1) well below chance (33.49% vs 50.00%). The results indicate human text is affectively coherent. However, synthetic text exhibits similar affective coherence to human text. Nonetheless, human text is more affectively coherent than synthetic text, and human text is less affectively incoherent than synthetic text, and this difference is statistically significant ($p < .001$). 128

6.11 The distribution of sentiment changes for sentiment 3-grams. Human text exhibits affective coherence (sentiment changes = 0) well above chance (48.48% vs 25.00%). Human text exhibits affective incoherence (sentiment changes = 2) well below chance (15.63% vs 25.00%). The human and synthetic distribution difference is significant ($p < .001$). The results are consistent with 2-grams in Figure 6.10. 128
6.12 The distribution of sentiment changes for sentiment 4-grams. Human text exhibits affective coherence (sentiment changes = 0) well above chance (37.07% vs 12.50%). Human text exhibits affective incoherence (sentiment changes = 3) well below chance (6.16% vs 12.50%). The human and synthetic distribution difference is significant ($p < .001$). The results are consistent with 2-grams and 3-grams in Figures 6.10 and 6.11. 129
6.13 The distribution of sentiment changes for sentiment 5-grams. Human text exhibits affective coherence (sentiment changes = 0) well above chance (29.19% vs 6.25%). Human text exhibits affective incoherence (sentiment changes = 4) well below chance (2.90% vs 6.25%). The human and synthetic distribution difference is significant ($p < .001$). The results are consistent with 2-grams, 3-grams, and 4-grams in Figures 6.10 to 6.12. 129
6.14 Distribution of 'percent sentiment change' per article for human and synthetic text in NEWSsynth (train 10k). The distributions are similar for human and synthetic text, with mean and median around 34%, and σ around 17%, as in Table 6.9. Nonetheless, the differences are statistically significant ($p < .001$). 132
6.15 Scatter plot of synthetic vs human 'percent sentiment change' per article - for a human article (used as prompt) and the corresponding synthetic article. There is very weak correlation, with Pearson $\rho = 28.10$. 132

D.1 Plutchik Wheel of Emotion. The middle ring includes Ekman's 6 emotions plus trust & anticipation. Similar emotions lie on adjacent spokes e.g. anger-disgust, while opposing emotions are placed on opposing spokes e.g. joy-sadness. Image from (Plutchik, 2001), downloaded from `https://en.m.wikipedia.org/wiki/File:Plutchik-wheel.svg`. 8
D.2 Visualisation of the word embedding space for Plutchik's 8 Emotions indicating the semantic relationship between words. Note this is 300D reduced to 2D for illustration. 8

E.1 The distribution of valence and emotions in online news headlines (GoodNewsEveryone (Bostan et al., 2020)) and media in general (UnifiedEmotions (Bostan and Klinger, 2018)) 11

How Emotions Can Help Detect Synthetic Text

Alan Cowap

Abstract

Question: Can you tell whether any of this thesis was written by AI? Recent developments in generative AI have shone a spotlight on high performance synthetic text generation technologies. The wide availability and ease of use of such models highlights the urgent need to provide equally powerful technologies capable of identifying synthetic text. With this in mind, we draw inspiration from psychological studies which suggest that people can be driven by emotion and encode emotion in the text they compose. We hypothesise that pretrained language models (PLMs) have an *affective deficit* because they lack such an emotional driver when generating text and consequently may generate synthetic text which has *affective incoherence* i.e. lacking the kind of emotional coherence present in human-authored text. We subsequently develop an emotionally aware detector by fine-tuning a PLM on emotion. Experiment results indicate that our emotionally-aware detector achieves improvements across a range of synthetic text generators, various sized models, datasets, and domains. We compare our emotionally-aware synthetic text detector to ChatGPT in the task of identification of its own output and show substantial gains, reinforcing the potential of emotion as a signal to identify synthetic text. These findings support the hypothesis that PLMs may have an *affective deficit.* Next, we investigate the hypothesis that synthetic text may be *affectively incoherent.* We create a novel flexible evaluation framework and use it to select an emotion classifier to generate an *affective profile* for 10k human and synthetic news articles. Our analysis of the human and synthetic affective profiles show that they are similar, but synthetic text is more *affectively incoherent* and less *affectively coherent*, than human text. Answer: AI wrote none of this thesis, but how do *you* know for certain? This lack of certainty motivates the task of synthetic text detection.

Chapter 1

Introduction

1.1 Thesis Topic

Synthetic text, i.e. text generated by Artificial Intelligence (AI), is revolutionising many aspects of how we work. However, synthetic text raises serious concerns including attribution, accuracy, authenticity, ethical issues, and data security (Brown et al., 2020). Attribution of text, as human or synthetic, motivates the task of synthetic text detection and is the focus of this thesis.

In this thesis we hypothesise that language models, which generate synthetic text, have an *affective deficit* i.e. an inability to properly represent emotions; and this *affective deficit* manifests itself in the generated synthetic text which can be *affectively incoherent.* In our first set of experiments, we exploit this *affective deficit* by creating an "emotionally-aware" language model and compare it against a standard language model in the task of synthetic text detection. The remainder of this thesis addresses the *affective incoherence* of synthetic text. We develop a flexible and extensible metric that allows us to measure and evaluate, at a fine-grained level, the performance of emotion classifiers. We use this metric to evaluate and choose an emotion classifier tailored to our needs. We then use this emotion classifier to compare and contrast the *affective coherence* of human and synthetic text.

Modern Pretrained Language Models (PLMs) can surpass human-level baselines across several tasks in general language understanding (Wang et al., 2018, 2019) and can produce synthetic text that can exceed human level quality, such as synthetic propaganda thought to be more plausible than human written propaganda (Zellers et al., 2019). PLMs have been used to generate disinformation (Zellers et al., 2019; Brown et al., 2020), left- or right-biased news (Gupta et al., 2020), fake comments (Weiss, 2019), fake reviews (Adelani et al., 2019), and plagiarism (Gao et al., 2022) and can generate synthetic text at scale, across domains, and across languages. The increasing high quality of synthetic text from larger and larger PLMs (and more recently LLMs - Large LMs, such as GPT 3 (Brown et al., 2020)) brings with it an increasing risk of negative impact due to potential misuses. ‘

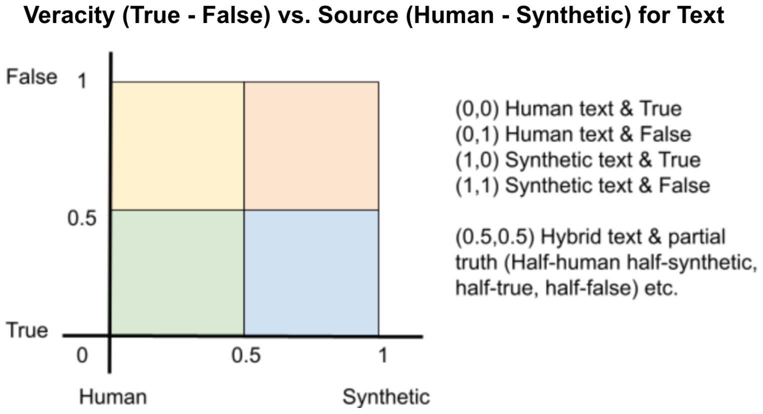

Veracity (True - False) vs. Source (Human - Synthetic) for Text
False
1
0.5
True
0
0.5
1
Human
Synthetic
(0,0) Human text & True
(0,1) Human text & False
(1,0) Synthetic text & True
(1,1) Synthetic text & False
(0.5,0.5) Hybrid text & partial truth (Half-human half-synthetic, half-true, half-false) etc.

Figure 1.1: Two dimensional representation of the Truth/Falsehood of Human/Synthetic text. Our research focuses on the X-axis i.e. human versus synthetic text detection.

In this thesis, we focus on the task of synthetic text detection. As illustrated in Figure 1.1, it is not simply the case that human text is true, and synthetic text is false, there are a continuum of values across the two dimensions. Truthful human journalism would reside at grid location (0,0), human fake news is at (0,1), while half-truths may be at (0,0.5). Synthetic text like Summarization, Machine Translation would reside at (1,0) given a true source and a non-hallucinating generator, while Grover prompted with propaganda might be at (1,1). This research focuses on the X-axis - detecting synthetic text.

Due to the potentially profound consequences of global synthetic disinformation we focus mainly, but not exclusively, on the detection of synthetic text in the news domain.[1] Synthetic news has already been published on one highly reputable media website, only later to be withdrawn and apologies issued for the “breach of trust” (Crowley, 2023a,b).

Current approaches to synthetic text detection tend to focus on learning artefacts from the output distribution of PLMs (Gehrmann et al., 2019; Pillutla et al., 2021; Mitchell et al., 2023), e.g. increased perplexity caused by nucleus sampling (Zellers et al., 2019). However, PLM distributions are dependent on training data and numerous hyperparameter choices including model architecture and sampling strategy. This gives rise to a combinatorial explosion of possible distributions and makes the task of synthetic text detection very difficult. Furthermore, it is not unexpected that performance decreases when classifying out-of-distribution instances, and there is a growing field of work investigating this shortcoming (Yang et al., 2023).

In this thesis, we consider not only the PLM output distribution, but also the other side of the synthetic text detection coin – human factors. We present a novel approach to the task of synthetic text detection which aims to exploit any difference between the expression of emotion in human-authored and synthetic text. We introduce the term ***affective coherence*** to refer to text comprising consistent (i.e. the same or similar) emotions. We introduce the term ***affective incoherence*** to refer to text comprising inconsistent (i.e. different or dissimilar) emotions. (We provide an example of each further below.) In Chapter 6 we introduce several ways to measure affective coherence and affective incoherence.

Neural word representations can have difficulty with emotion words, and PLM sampling strategies (i.e. how the next word or token is selected) are stochastic rather than driven by emotion – we use the term ***affective deficit*** to refer to these shortcomings of language models. As a result of the language

---

[1] The news domain is recognised as having high emotional content (Strapparava and Mihalcea, 2007; Bostan et al., 2020). See also Appendix C.

models affective deficit, the generated synthetic text can express emotion in an incoherent way, i.e. synthetic text can be affectively incoherent. To be clear, we do not contend that synthetic text is devoid of emotion, rather that the emotional content of synthetic text may be affectively incoherent, and that this affective incoherence stems from the underlying affective deficit of the PLM.

Our hypothesis that human text is affectively coherent and synthetic text is affectively incoherent is based upon the differences between human text generation and synthetic text generation (which we describe in §2.3 and §2.2 respectively). We point out that, despite the hypothesis, it is possible for synthetic text to exhibit affective coherence (e.g. this would be expected even in a random scenario, or where the language model parrots human text), and for human text to exhibit affective incoherence (e.g. for plot development, or a happy ending to a sad story). That said, the current work is primarily focused in the news domain rather than creative writing. Furthermore, Wang et al. (2020b) report that human dialogue exhibits emotional consistency whereby similar emotions tend to appear adjacent to one another, while dissimilar emotions seldom do.

For the purpose of demonstration of the affective deficit that we believe to be characteristic of text produced by PLMs, we provide the following simple example of human- versus machine-authored text with positive emotion words highlighted in orange and negative emotion words in pink. One shows coherent emotion expected of human-authored text, while the other demonstrates affective incoherence (see footnote[2] to reveal which was synthetic/human-authored text).

1. *Roberts chuckled when asked if he was happy to be on the other team now when Puig's name comes up. "Yeah, I am happy," he said, smiling.*
2. *I'm really happy for him. Over the course of those three seasons, the 25-year-old has gone from rolling to poor to worse and old.*

---

[2] (1) is human-authored while (2) is synthetic text. Both are from the *NEWSsynth* dataset (see §3.2.2).

In this simple example, we have demonstrated one kind of affective incoherence present in synthetic text but we suspect that there are additional and more complex emotional patterns that might go undetected by humans. To reiterate, we hypothesise that the *affective deficit* of PLMs could result in synthetic text which is *affectively incoherent*, which could be useful in distinguishing it from human text.

1.2 Research Questions

Given the context of synthetic text detection as an important task, and the potential *affective deficit* of language models manifesting as *affective incoherence* in synthetic text, the overarching hypothesis of this research is that **the task of synthetic text detection can be improved by exploiting emotion.** To address this hypothesis, we explore the following four research questions.

RQ1 - Can the performance of automatic synthetic text detectors be improved by making them 'emotionally-aware' compared to standard detectors? We investigate this hypothesis in RQ1 and provide examples of the improved performance of emotionally-aware detectors in §3.3.

We use a transfer learning (Pan and Yang, 2010) method to fine-tune an "emotionally-aware" detector model. By fine-tuning a PLM first on emotion classification and then on our target task of synthetic text detection, we demonstrate improvements across a range of synthetic text generators, various sized models, datasets and domains. Furthermore, our emotionally-aware detector proves to be more accurate at distinguishing between human and ChatGPT text than (zero-shot) ChatGPT itself.

Finally, we create two new datasets: *NEWSsynth*, a dataset of 20k human and synthetic news articles, and *ChatGPT100*, a testset of 100 human and ChatGPT texts on a range of topics. All code, models and datasets are publicly available to aid future research.[3]

Having demonstrated the *affective deficit* of PLMs in RQ1, we next investigate the *affective coherence* of human and synthetic text. We do this by generating an *affective profile* (RQ4), a term we introduce to refer to the emotional content and characteristics of human and synthetic text, for human and synthetic articles. For this purpose, an *affective profile* refers to the sequence of emotions in an article, or portion of an article, at the sentence level. Before we can generate an affective profile we need an appropriate emotion classifier (RQ3), and in order evaluate and choose an emotion classifier we must use an appropriate metric (RQ2).

RQ2 - Is it possible to develop a framework for measuring and evaluating classifiers of Plutchik's 8 emotions which takes account of their dimensional relationship? Emotion classifiers are most frequently evaluated in NLP emotion literature using F-score (Alswaidan and Menai, 2020), along with Accuracy, and Precision and Recall (Kusal et al., 2023). A feature of these metrics is that they are binary, because a prediction is either correct or incorrect, and they give an output in the range 0..100%. In choosing a metric for evaluating emotion classifiers we first consider which emotion model to use. In RQ1 we use Ekman's 6 basic emotions which represent 6 independent emotions (Ekman, 1999). However, there are dimensional models which place an order and relationship between emotions. The most commonly used dimensional emotion model in NLP literature is Plutchik's Wheel of Emotion (Plutchik, 1980, 2001).

The obvious natural choice is to use the commonly used metrics for measuring and evaluating the emotion classifiers. However, if we exploit the relationship between Plutchik's 8 emotions and specify a numerical difference between each pair of emotions - then we can use it to measure a more fine-grained and nuanced difference between two emotions - rather than a simple binary mea-

---

[3] `https://github.com/alanagiasi/emoPLMsynth`

surement of correct or incorrect. Furthermore, we can define a mathematical equation to calculate this difference and output a value in the range 0..100% - which is familiar to the affective NLP community. We address this task in RQ2 and contrast our novel metric with Accuracy and two commonly used error-based metrics i.e. mean absolute error (MAE) and root mean squared error (RMSE). In addition, we will use this metric in RQ4 when we measure affective coherence in human and synthetic articles, but first we use it in RQ3 to evaluate emotion classifiers.

RQ3 - Can we exploit the dimensional relationship between Plutchik's 8 emotions to enable optimal selection of an automatic emotion classifier? We can now use our novel evaluation metric from RQ2 to evaluate emotion classifiers. However, there is no emotion classifier for Plutchik's 8 emotions in the news domain, nor is there a dataset for Plutchik's 8 emotions in the news domain. Nonetheless, there are many PLMs that we can fine-tune on this task by modifying existing datasets. As such, we take two datasets in the news domain labelled with emotions, GoodNewsEveryone (Bostan et al., 2020) and AffectiveText (Strapparava and Mihalcea, 2008), and map the emotions defined in the datasets to Plutchik's 8 emotions. We use this combined and customised dataset to fine-tune PLMs in the task of emotion detection. We evaluate the PLMs using our novel metric, in addition to F-score, MAE, and RMSE, and select the highest performing model to generate affective profiles in RQ4.

RQ4 - Can we generate and compare the affective profile of human text and synthetic text? Equipped with a suitable emotion classifier from RQ3, we generate affective profiles and investigate the affective coherence of human and synthetic articles. We take inspiration from the use of n-grams in classical NLP, which refers to sequences of adjacent tokens, and generate n-grams of emotions at the sentence level. We define *affective coherence* as a sequence of unchanging emotions, in contrast to *affective incoherence* which is a sequence of changing emotions. We further investigate human and synthetic text by comparing the affective coherence in each article. We also take advantage of our custom metric, from RQ2, and use it to measure *affective coherence* - thereby demonstrating its flexibility and extensibility as a metric, and allowing us to numerically compare the affective coherence of human and synthetic text. We demonstrate that the affective coherence of human text and synthetic text is similar - but slightly different - human text is slightly more affectively coherent, and slightly less affectively incoherent, than synthetic text. Our statistical analyses show the differences between human and synthetic text are statistically significant ($p < .001$).

We present the following novel contributions:

1. Demonstrated that emotion can help in the task of synthetic text detection.
2. *NEWSsynth* - a dataset of 20k human and synthetic (Grover) news articles.
3. ChatGPT100 - a dataset of 100 human and synthetic articles (ChatGPT) in various domains.
4. Multiple PLMs fine-tuned on emotion classification for Plutchik's 8 emotions, and Ekman's 6 emotions.
5. Multiple datasets customised with emotion mappings to Plutchik's 8 emotions, Ekman's 6 emotions, and sentiment.
6. ψ a novel flexible framework for evaluating classifiers (of dimensional emotions, and more).
7. An analysis of *affective profiles* in human and synthetic text news articles.

1.3 Thesis Structure

The remainder of this thesis is outlined below:

Chapter 2: Related Work This chapter provides the background to two main topics relevant to this thesis, discussing relevant literature and identifying the research gap this thesis addresses. First, we discuss language models in the context of synthetic text generation and detection, noting how they continue to improve but still pose challenges - including their dual-use for good and for bad. We describe the motivation for, and the difficulty with, the task of synthetic text detection. Second, we discuss human factors (particularly emotion) and how emotion plays a pivotal role in human cognition, and this may be absent from language models (*affective deficit*) causing synthetic text to be *affectively incoherent* - which we attempt to exploit in the task of synthetic text detection. The chapter also briefly describes relevant measurements and metrics used in this thesis.

Chapter 3: Improving Neural Detection of Synthetic Text via Emotion Recognition This chapter aims to answer RQ1: *Can the performance of automatic synthetic text detectors be improved by making them 'emotionally-aware' compared to standard detectors?*. It describes our primary experiments which demonstrate the benefit of exploiting emotion in the task of synthetic text detection, and are indicative of the *affective deficit* of language models.

We first describe how we create *NEWSsynth* - a dataset of 20,000 human and synthetic news articles, used throughout this thesis. We contrast the performance of "emotionally-aware" language models versus otherwise equivalent standard language models, in the task of synthetic text detection. We do this using different size language models for text generation and for detector models, in different domains, and use different language model architectures; obtaining consistent improvements for the "emotionally-aware" models. We suggest these results are indicative of the *affective deficit* of language models.

We conduct additional experiments investigating the relative length of human and synthetic text - in order to rule it out as a proxy for predicting synthetic text. We conduct experiments using different combinations of emotion and sentiment to create the "emotionally-aware" detector and compare their performance; and use the results to inform our choice of emotion datasets in this experiment, and in our later experiments in RQ3. We conducted an experiment comparing our detectors (standard and emotionally-aware) against ChatGPT at detecting ChatGPTs own text (in a dataset we created *ChatGPT100*), the purpose being to see how our technique worked with LLMs like ChatGPT as opposed to the PLMs we primarily used.

The results in this chapter are indicative of the *affective deficit* of language models. The remaining experimental chapters are aimed towards investigating the *affective coherence* of human and synthetic text.

Chapter 4: ψ - a Novel Evaluation Framework (for Dimensional Emotion Classifiers) This chapter aims to answer RQ2: *Is it possible to develop a framework for measuring and evaluating classifiers of Plutchik's 8 emotions which takes account of their dimensional relationship?*. It introduces ψ, our novel evaluation metric, describing the motivation and inspiration behind its development.

We describe the rationale and derivation of the various components of the ψ equation. We demonstrate its use in a toy example and provide a detailed comparison between ψ, Accuracy, MAE and RMSE. The inspiration for ψ is borne, on the one hand, from the dimensional relationship between emotions in Plutchik's Wheel of Emotion (Plutchik, 1980, 2001), and, on the other hand, due to the failure of typical metrics like Accuracy and F-score to measure this dimensional relationship. We introduce δ, a measure of emotion similarity, which we use to measure the 'distance' between emotions, and which allows us to define a 'loss' between, for example, a gold label emotion and a predicted emotion. We generalise this loss into a loss function, $\phi(\delta)$, which can be specified in numerous ways to suit user requirements - thus providing flexibility and extensibility. We then add terms to the ψ equation so the loss is summed and converted to a percentage score. We do this to allow consistent comparison of ψ scores and because it is familiar to the affective NLP research community (Accuracy, F-score, Precision, and Recall are the commonly used metrics and all are percentages). We also use δ as a means to calculate MAE and RMSE scores for emotion classification, and compare and contrast these with ψ and Accuracy. We do this for our toy example in this chapter, and in RQ3 we use real-world examples (when fine-tuning PLMs and LLMs) to correlate ψ with $F1_{\mu}$, MAE, and RMSE.

We employ ψ, in the first instance, to evaluate the performance of emotion classifiers in RQ3, and demonstrate its flexibility and extensibility in RQ4.

Chapter 5: Developing and Selecting a Dimensional Emotion Classification Model This chapter aims to answer RQ3: *Can we exploit the dimensional relationship between Plutchik's 8 emotions to enable optimal selection of an automatic emotion classifier?*. It describes how we fine-tune, and (using ψ from RQ2) evaluate and select a custom emotion classifier.

We describe how we take advantage of our findings from RQ1, relating to which emotion datasets used for fine-tuning resulted in the best performing language models, and combine those datasets to create a custom dataset mapped to Plutchik's 8 emotions in the target domain of news. We use our custom dataset to fine-tune 15 language models in emotion classification, and we create 6 ensembles of models. We use ψ to evaluate and select the best performing model or ensemble overall. We describe related work using zero-shot Natural Language Inference (NLI) for emotion classification, and we report our results using this technique on 15 diverse language models. Based on the ψ score, we select the best performing model from our fine-tuning and zero-shot NLI experiments. This model will be used in RQ4 to classify emotions and generate affective profiles for human and synthetic text.

Chapter 6: Affective Profiling of Human and Synthetic Text This chapter aims to answer RQ4: *Can we generate and compare the affective profile of human text and synthetic text?*. It describes how we generate and compare the affective profiles of human and synthetic text. It brings together aspects of the previous three chapters, and is the culmination of the previous two chapters.

We use our *NEWSsynth* dataset from RQ1, our ψ metric and δ-profile from RQ2, and our top-performing emotion classifier from RQ3. We use the emotion classifier to predict emotion at the sentence level. We define the *affective profile* of an article as the sequence of emotions in that article, and we describe how we use the *affective profile* to measure *affective coherence*. Recall that in RQ1 our results were indicative of the *affective deficit* of language models, and we are now investigating if that affective deficit results in synthetic text which is *affectively incoherent*. Therefore, we use the *affective profile* of an article to investigate its *affective coherence* so we can compare human and synthetic text. We compare human and synthetic text - at the article level, and at the n-gram of emotion level (using n-grams of emotion of length n = 2, 3, 4, and 5). Additionally, we use the δ-profile of an article to compare human and synthetic text, and use the ψ metric to measure the *affective coherence* of an article. We repeat some of these experiments using sentiment, by mapping our custom dataset to binary sentiment. We see a consistent pattern of results repeat itself across all of the experiments in this chapter, which inform our conclusions on the affective coherence of human and synthetic text. Our results consistently demonstrate that human and synthetic text display similar, but slightly different, *affective coherence* and *affective incoherence*. Our statistical analyses show the differences between human and synthetic text are statistically significant ($p < .001$). The results also show a very weak, or weak, correlation between the *affective coherence* of a human article (used as prompt to a PLM) and the corresponding generated synthetic article.

Chapter 7: Conclusion This chapter summarises the original contributions of this thesis, and revisits each of the Research Questions. It addresses the limitations of this work, and identifies directions for future work.

1.4 Publications

During this research the work described mainly in RQ1 (Chapter 3) was published in the following paper:

Cowap, A., Graham, Y. and Foster, J. (2023) 'Do Stochastic Parrots have Feelings Too? Improving Neural Detection of Synthetic Text via Emotion Recognition', in H. Bouamor, J. Pino, and K. Bali (eds) Findings of the Association for Computational Linguistics: EMNLP 2023. Findings 2023, Singapore: Association for Computational Linguistics, pp. 9928–9946. Available at: `https://aclanthology.org/2023.findings-emnlp.665` (Accessed: 12 December 2023).

A related collaborative project involving pretraining and fine-tuning a PLM (gaBERT - the first published Irish language model), which informed our experimental method for RQ1 and RQ3, was published in the following paper:

Barry, J., Wager, J., Cassidy, L., **Cowap, A.**, Lynn, T., Walsh, A., O'Meachair, MJ., Foster, J. (2022) 'gaBERT — an Irish Language Model', in Proceedings of the Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, pp. 4774–4788. Available at: `https://aclanthology.org/2022.lrec-1.511`.

Papers describing work from RQ2-4 (Chapters 4-6) are being prepared for submission to the ACL journal 'Transactions of the Association for Computational Linguistics' (TACL), and ACL Rolling Review (for EACL, NAACL, or ACL):

Cowap, A., Graham, Y. and Foster, J. (2024) 'ψ - a Flexible Framework for Evaluating Classifiers: Demonstrated on emoRoBERTa - a Dimensional Emotion Classifier for Plutchik's 8 Emotions', TACL Submission expected September - October 2024.

Cowap, A., Graham, Y. and Foster, J. (2024) 'Human or Machine? - Contrasting Human and Synthetic Text using Affective Profiles', ARR Submission expected August - October 2024.

Chapter 2

Related Work

2.1 Introduction

In this chapter we consider two distinct areas of research which our work brings together. We consider Language Models (§2.2) in the context of synthetic text generation and detection. We consider Human Factors (§2.3) such as human cognition and the key role played by emotion. We highlight emotion as a key difference between humans and language models, and how our work focuses on exploiting emotion to help detect synthetic text. We draw the readers attention to the Glossary, located after the Table of Contents, which serves as a convenient reminder of some key phrases (typically printed *in italics*) used in this thesis and from psychology literature.

First, we consider language models (LM) and the body of research concerning synthetic text generation and detection. We identify research gaps, in the task of synthetic text detection, which this work attempts to address. We will see how language models are continually improving and pose many challenges (§2.2.1) such as bias, and of particular interest to this work - their dual use (§2.2.2). This dual use motivates the task of synthetic text detection (§2.2.3) which remains a difficult task (§2.2.4).

Secondly, we consider the other side of synthetic text detection - human factors (particularly emotion) - and their effect on human authored text. We discuss human cognition and its inherent biases (§2.3.1), and how emotions appear specific to humans and animals (§2.3.2). We address the question of whether we can exploit emotions in the task of synthetic text detection (§2.3.3) and discuss how identifying emotions remains a difficult task (§2.3.4). We finish with a brief look at measuring and metrics (2.4).

2.2 Language Models

2.2.1 Language Models Continue to Improve, but Still Pose Challenges

In this subsection we discuss word counts vs vectors, sparse vs dense vectors, static vs dynamic embeddings, and causal vs masked LMs. An LM is a probabilistic representation of a language. Early LMs counted the occurrence of a single word, or sequence of words called n-grams, to calculate the probability of a sequence of words and predict which word is most likely to come next. These models could not properly handle words unseen during training, i.e out-of-vocabulary words, although this is mitigated using smoothing techniques. They generated synthetic text that was semantically incoherent, and while increasing n-gram size improved readability, it tended to reproduce training text rather than generate new text.

Words can also be represented as vectors. A vector is a series of numbers which can be represented as dimensions in hyperspace, and also allows for words to be compared using vector distance. The more similar two vectors, measured by dot product or cosine similarity, the more similar the words.

Sparse vector embeddings such as one-hot-encoding can grow unwieldy (e.g. 50,000 dimensions) and this comes at a memory and compute cost. Dense vectors like word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and fastText (Bojanowski et al., 2017) are typically 50-1,000 dimensions. Most vectors produce distributed representations of words which consider the context (i.e. surrounding words) in which a word occurs based on the *distributional hypothesis* (Wittgenstein, 1953, PI 43)(Firth, 1957). This means out-of-vocabulary words are easily and accurately handled based on the context. Dense vectors can capture semantic relationships between words e.g. "man is to king", as "woman is to _____" and the LM correctly predicts "queen" (Pennington et al., 2014). Semantic relationships between words are not without their biases (Bolukbasi et al., 2016). In addition, word representations do not differentiate antonyms from synonyms (Santus et al., 2014). GloVE and word2vec embeddings do not perform well at distinguishing opposing emotions (e.g. joy-sadness) (Seyeditabari and Zadrozny, 2017).

Another property of embeddings is whether they are static or dynamic. Static embeddings have a single representation for each word which is problematic for polysemous words e.g. "trust" is a financial instrument and an emotion.[1] Dynamic embeddings solve this by allowing multiple senses of a word to be represented by multiple embeddings.

Recurrent Neural Networks (RNNs) bring together dynamic embeddings, dense vectors, and the distributional hypothesis, by taking a whole sentence as input and encoding word embeddings for each word in the sentence by including information from nearby words. The Long Short Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1996) is one type of RNN which gained widespread use in NLP because it handles longer input sequences better than a standard RNN. The Bidirectional LSTM (BiLSTM) (Graves and Schmidhuber, 2005) improves upon the LSTM by considering the context on both sides of a given word from the input sequence. It achieves this by doing a forward pass through the input sequence to a forward LSTM, and a backward pass through the input sequence to a backward LSTM, and then combining the hidden states of the two LSTMs. ELMo (Embeddings from Language Models) uses a bidirectional LSTM to create contextualised word representations based on the whole

---

[1] In a preliminary experiment we use semantic relationships, between static embeddings, and demonstrate this shortcoming. We describe the experiment in Appendix D.

input sequence, which allows it to handle polysemous words and show improvements in several tasks including sentiment analysis (Peters et al., 2018). ELMo is a Pretrained Language Model (PLM) which is subsequently fine-tuned on specific downstream tasks, with relatively small amounts of data, thereby taking advantage of the ELMo word representations learned from pretraining on a large dataset.

Deep learning models advanced from RNN, LSTM, BiLSTM, to the Transformer (Vaswani et al., 2017). Transformers transformed the field of NLP by doing away with recurrence and using attention mechanisms. Self-attention is an attention mechanism which computes representations for an input sequence by relating different positions of the input sequence. In other words, self-attention measures how useful the current input word is to another word in the input sequence, based on some arbitrary task e.g. language modelling. Crucially, self-attention calculations are parallelisable, making transformer-based models much quicker to train. Transformers use an encoder and/or decoder architecture with attention mechanisms. In simple terms, the encoder converts a sequence of tokens into a vector, while the decoder converts a vector into a sequence of tokens. Transformer based LMs can be categorised as Causal LMs (CLM) or Masked LMs (MLM). CLMs use a stacked Decoder architecture that masks future tokens and this makes them suitable for generating text. For example, given the prompt “The man sat on ” the CLM will generate a probability for each token in its vocabulary and choose the next token using a *sampling strategy* (e.g. argmax, beam search, random, top-k, top-p) to return say “a”. This new token is appended to the original prompt to create a new prompt “The man sat on a” and the cycle continues generating more synthetic text. MLMs use a stacked Encoder architecture which can mask tokens anywhere in the text and this makes them particularly suitable for classifying an input sequence. The prompt for an MLM will be like a cloze-test, for example, given the prompt “The man [MASK] on a chair” the model will assign probabilities and use a *sampling strategy* to return say “sat” or “stood”.

Pretrained Transformer-based models can be finetuned for user-specified tasks using relatively small amounts of data and compute. Fine-tuning is a very efficient way to take advantage of the learned word embeddings of a PLM and modify those embeddings (particularly the final layers) to improve performance on a specific task (Rogers et al., 2021). PLMs perform well on multiple tasks including the General Language Understanding Evaluation (GLUE) benchmark, which consists of nine diverse linguistic tasks across different domains (Wang et al., 2018) and the more difficult SuperGLUE benchmark (Wang et al., 2019). LMs currently occupy all the top spots while non-expert human baselines currently lie in 23rd position on the GLUE leaderboard[2] and 8th position on the SuperGLUE leaderboard.[3] We will use fine-tuned PLMs and their dynamic dense vectors to both generate and detect synthetic text. For our primary experiments in RQ1, we use a CLM for synthetic text generation and an MLM for synthetic text detection and emotion classification. The CLM used is Grover (Zellers et al., 2019) which has the same architecture as GPT-2 (Radford et al., 2019), and the MLM used is BERT (Devlin et al., 2019).

The last couple of years have seen the introduction, and global impact, of Large Language Models (LLM). LLMs are much larger (billions and tens-of-billions of parameters) and more human-aligned than PLMs. LLMs undergo a second training or “alignment” phase using Reinforcement Learning from Human Feedback on top of the underlying LLM - for example GPT 3.5 (Ouyang et al., 2022; OpenAI, 2022). Nonetheless, LLMs have issues which remain unresolved, such as hallucination, toxicity, and data privacy. We will use LLMs such as GPT 3.5 and BLOOM (Scao et al., 2023) in RQ1 and RQ3.

LMs have improved from count-based n-grams, to word embeddings, to PLMs, and the current state-of-the-art LLMs. While LMs are improving, they continue to pose challenges. We now turn to some further issues with LMs -

---

[2] `https://gluebenchmark.com/leaderboard` accessed on 04-June-2024.

[3] `https://super.gluebenchmark.com/leaderboard` accessed on 04-June-2024.

their dual use and bias.

2.2.2 Language Models are Dual-Use and Biased

LMs have been successfully deployed in a wide variety of applications including: information retrieval e.g. Google Search uses BERT;[4] Microsoft uses GPT-3 for text-to-SQL generation;[5] Woebot Health deployed a chatbot which creates "therapeutic bonds" with users while adapting to their emotional state and cognitive ability.[6] Other applications include: rewriting a food recipe given dietary restrictions (e.g. vegan) (Hegel et al., 2020); Machine Translation; Autocompletion; Summarization; supporting endangered and low resource languages (Barry et al., 2022) etc.

LLMs continue to have a global impact, which started with the public launch of ChatGPT (OpenAI, 2022). Very quickly, concerns about ChatGPT plagiarism arose (Gao et al., 2022). Microsoft launched Copilot[7] a chat-based assistant intended to replace standard search, with Google using Gemini for AI powered search[8].

However, concerns about the misuse of powerful LMs led to the staged release of GPT-2[9] and contributed to the restricted release of GPT-3.[10] LMs can be finetuned at little to no cost using freely available resources and their misuse has been demonstrated successfully, for example, submitting fake comments to a US Federal website (Weiss, 2019). Furthermore, *bad actors* could pretrain LMs to generate synthetic *disinformation* or *malinformation.* Zellers et al. (2019) demonstrated this potential for misuse with Grover, an LM with the same architecture as GPT-2 but pretrained on a dataset of 32M news articles.

---

[4] `https://blog.google/products/search/search-language-understanding-bert/`

[5] `https://blogs.microsoft.com/ai/from-conversation-to-code-microsoft-introduces-its-first-product-features-powered-by-gpt-3/`

[6] `https://woebothealth.com/`

[7] `https://blogs.microsoft.com/blog/2023/09/21/announcing-microsoft-copilot-your-everyday-ai-companion/`

[8] `https://blog.google/products/search/generative-ai-search/`

[9] `https://openai.com/blog/better-language-models/`

[10] `https://openai.com/blog/openai-api/`

Prompted with human written propaganda, Grover generated synthetic propaganda that was more plausible than the human written propaganda - rated by human subjects (Zellers et al., 2019). In addition to propaganda, other misuses of LMs include generating: fake news (Zellers et al., 2019; Brown et al., 2020); fake reviews (Adelani et al., 2019); fake comments (Weiss, 2019); other possible misuses are limited only by imagination.

LMs typically have *exposure bias* as a result of being trained on (certain types of) human text only (Ranzato et al., 2016) and *inductive bias* as a result of learning by generalising from specific training examples. Indeed, this inductive capability differentiates LMs, and machine learning in general, from traditional declarative programming with data structures and data models. The biases in datasets used to train LMs can be compounded by *sampling biases* such as: sampling online social media which is more likely to be *WEIRD*[11] (representing less than 12% of the population (Henrich et al., 2010)), young and male (as in the case of outbound reddit links[12] used to train GPT-2); *publication bias* (Song et al., 2010); and finally the *knowledge effect* (Hayes et al., 1986) whereby we tend not to state the obvious, so LMs may not learn obvious relationships such as that fire does not normally occur underwater. LMs have biases inherent in their design which can be compounded by human biases embedded in training data.

It is important to distinguish between *information*, *misinformation*, *disinformation*, and *malinformation*, as defined by Wardle and Derakhshan (2018). Using these definitions a third dimension, Purpose, could allow for the differentiation of satire (fake news is not always intended to deceive, it can be satire e.g. Waterford Whispers,[13] The Onion[14]), hoax, propaganda, and fake

---

[11] WEIRD is an acronym for people from Western, Educated, Industrialised, Rich and Democratic societies (who tend to be over-represented in many studies but are not representative of the population i.e. $< 12\%$).

[12] https://www.pewresearch.org/journalism/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

[13] https://waterfordwhispersnews.com/

[14] https://www.theonion.com/

news from trusted news; a differentiation which has been demonstrated using linguistic characteristics (Rashkin et al., 2017; Horne and Adali, 2017). However, while linguistic characteristics can help differentiate the purpose of human text, they are limited against synthetic text because the style of synthetic text is consistent regardless of purpose (Schuster et al., 2020).

The potential dangers for misuse of LMs are amplified by their ability to generate synthetic text at massive scale and speed across domains and languages, exploiting the pervasiveness of online fora and social media. They are further exacerbated by *human factors* including the multitude of biases inherent in human cognition and the resultant impact on human decision making and action. This human susceptibility to LM dual-use is a primary motivation for the task of synthetic text detection; and we discuss this topic next.

2.2.3 Why do we Want to Detect Synthetic Text?

We want to detect synthetic text for several reasons, including misinformation (§1), dual-use issues (§2.2.2), human biases (§2.3.1) and their effect on decision making and behaviour. Add to that the social impact of NLP including potential and actual harms (Hovy and Spruit, 2016; Bender et al., 2021). For example, when consumers purchase goods or services online they often rely on other people's reviews to guide their decision, but fake reviews undermine consumer confidence and represent a growing business (Hovy, 2016), and they can be hard for people to detect (Adelani et al., 2019). Cambridge Analytica and Facebook arguably played a decisive role in the outcome of the UK "Brexit" referendum (Hall, 2021) and US Presidential election (Isaak and Hanna, 2018), both held in 2016. These are huge impacts. The information harvested for these targeted advertising campaigns could be combined with user profile data to create a prompt for generating synthetic text - customised and targeted individually. We already saw that LMs can generate propaganda that is more plausible than human written propaganda (Zellers et al., 2019). There is also a danger that online misinformation in the form of synthetic text will begin to pollute the very datasets we use for training LMs. Redefining truth and undermining trust are dangerous, and the potential misuse of LMs could contribute to that harm.

Another reason to detect synthetic text is the potential for exploitation of the effect of emotion on humans, and our resulting susceptibility to manipulation. More extreme and surprising news is overreported compared to its statistical occurrence, while less extreme and surprising news is underreported (Combs and Slovic, 1979). For example, Combs & Slovic found that violent and catastrophic causes of death are overreported, whereas deaths from stroke, all diseases, cancers etc. were underreported. Most notably, Combs & Slovic found that the estimated frequency of causes of death reported by test subjects matched the overreporting and underreporting of the causes of death in the news coverage. Even when instructed to avoid such biases they persisted (Lichtenstein et al., 1978) in what appear to be examples of *availability bias* (Tversky and Kahneman, 1973). It has long been the case that negative news is reported more often than positive news (Galtung and Ruge, 1965) and we will see the significant effect of *negativity bias* in §2.3.1. Online platforms use *engagement based ranking* particularly using anger to increase engagement (Berger and Milkman, 2012; Haugen, 2021) and target specific content at a specific individual (Hall, 2021). We describe the triumvirate of surprise, negativity, and anger as the *Surprise Negativity Anger Paradigm* (SNAP).[15] Now that we understand some of the motivating factors for synthetic text detection, next we discuss how it remains a difficult task.

---

[15] In a preliminary experiment we investigate the distribution of emotions and sentiment in news headlines compared to general media. We find that Surprise, Negativity, and Anger are more prevalent in news headlines than in general media. See Appendix E.

2.2.4 Detecting Synthetic Text is Difficult

People are relatively poor at detecting synthetic text, and have been shown to score just above random chance (Gehrmann et al., 2019; Uchendu et al., 2021). Hybrid systems, such as GLTR (Gehrmann et al., 2019) for example, use automation to provide information to aid human classification, highlighting a text sequence using colours to represent likeness to the PLM output distribution such as GPT-2 (Radford et al., 2019). Gehrmann et al. (2019) reported an increase in detection accuracy of approximately 18% (from 54% to 72%) using GLTR, while Uchendu et al. (2021) report an F1 score of 46% using GLTR with a heuristic based on an analysis of human text.

Both human and hybrid approaches involve human decisions, which can be slow, expensive, susceptible to bias, and inconsistent. Automatic detection produces the best results for synthetic text detection. This usually involves training PLMs to detect other PLMs, but zero-shot detection methods also exist, e.g. DetectGPT (Mitchell et al., 2023). Potentially the best supervised detector, BERT, can detect synthetic text from 19 different generators with a mean F1 of 87.99%, compared to 56.81% for hybrid, and worst of all humans at 53.58% (Uchendu et al., 2021).

Performance of SOTA detectors can however be inconsistent and unpredictable due to several factors specific to both the detector and generator, including: model size and architecture, training data and domain thereof, sampling strategy, hyperparameter selection, and sentence length. As mentioned above, Uchendu et al. (2021) showed the best of these models (BERT) achieves a mean F1 of 87.99% on 19 different synthetic text generators. However, the mean score hides the wide range ($\approx$53%) of F1's, ranging from as low as 47.01% to as high as 99.97%, for distinct synthetic text generators. This volatility may be due in part to the detector simply learning artefacts of the generator distribution. RoBERTa is the second best performing model with a mean F1 of 82.80%. Consequently, the task of synthetic text detection is somewhat of an arms race with detectors playing catch-up, forced to learn ever-changing distributions due to the numerous factors that can potentially change those distributions.

Existing approaches to synthetic text detection exploit properties of synthetic text. Synthetic text can be incoherent and degrade as the length of generated text increases (Holtzman et al., 2020), perplexity increases with increasing length unlike human text (Zellers et al., 2019), and PLMs are susceptible to sampling bias, induction bias, and exposure bias (Ranzato et al., 2016). For example, exposure bias can contribute to brittle text which is repetitive, incoherent, even containing hallucinations (Arora et al., 2022). Synthetic text can have an inconsistent factual structure, such as mentioning irrelevant entities (Zhong et al., 2020). Perhaps unsurprisingly, synthetic text detection is less difficult with longer excerpts of generated text, for both humans and machines (Ippolito et al., 2020).

One aspect of writing that has not, up to now, been a focus of synthetic text detection efforts is the expression of emotion. The problem of encoding emotion was first identified in neural NLP with static embeddings such as word2vec (Mikolov et al., 2013; Wang et al., 2020a). Static word embeddings have difficulty distinguishing antonyms from synonyms (Santus et al., 2014). This deficit is present in embeddings for words which represent opposing emotions (e.g. joy-sadness) (Seyeditabari and Zadrozny, 2017). Furthermore, words representing opposing emotions can have closer embeddings relative to words representing similar emotions (Agrawal et al., 2018). There have been various approaches to address this affective deficit in embeddings, such as transfer learning from sentiment analysis (Kratzwald et al., 2018), an additional training phase using an emotional lexicon and psychological model of emotions (Seyeditabari et al., 2019), and combining separately-learned semantic and sentiment embedding spaces (Wang et al., 2020a).

Addressing potential affective deficits of PLMs is also the goal of work aiming to make dialogue systems more empathetic. For example Huang et al. (2018) force dialogue generation to express emotion based on the emotion detected in an utterance, while Rashkin et al. (2019) follow a similar approach with a transformer architecture to make the system more empathetic.[16] In contrast, Wang et al. (2020b) report that human dialogue tends to display consistency in emotional content whereby similar emotions tend to occur adjacent to each other while dissimilar emotions seldom do.

In this context, the task of emotion classification can benefit from considering emotions in concert rather than in isolation. Wang et al. (2020b) leverage what they call "*emotional consistency*" i.e. the relations between preceding and succeeding emotions to determine the globally best emotion labels. Labat et al. (2022) describe "*emotion trajectories*" i.e. recurring emotion transitions which they use to improve emotion classification. Wemmer et al. (2024) describe "*emotion progression*" which considers the emotions up to the current sequence when predicting an emotion label. We introduce the term ***affective coherence*** to describe the similarity between consecutive emotions in text.

We hypothesise that, while human text can be *affectively coherent* - as indicated by the work of Wang et al. (2020b); Labat et al. (2022); Wemmer et al. (2024), in contrast, synthetic text may be *affectively incoherent* as indicated by the work to improve or control emotion in synthetic text by Huang et al. (2018); Rashkin et al. (2019) and surveyed for sentiment by Lorandi and Belz (2023). We first test this hypothesis by fine-tuning emotionally-aware PLMs which show improvement over non emotionally-aware PLMs in the task of synthetic text detection described in Chapter 3, and more directly in Chapter 6 where we compare the *affective profile* of human and synthetic text for both emotions and sentiment.

Past work in synthetic text detection has focused on the properties of syn-

---

[16] For a comprehensive survey of sentiment control in synthetic text see (Lorandi and Belz, 2023) and for studies of emotion in human writing, see (Brand, 1985, 1987, 1991; Bohn-Gettler and Rapp, 2014; Knaller, 2017).

thetic text generators and is yet to take advantage of the factors that potentially influence human-authored text, such as the emotions humans express in the text they write. Our work exploits this PLM affective deficit to improve synthetic text detection. Having discussed LMs in detail, including their ongoing improvement and the challenges they pose, and presented the motivation and the difficulty of detecting synthetic text; we now turn to human factors.

2.3 Human Factors

Our work aims to exploit emotion as a means to detect synthetic text. Therefore, we discuss human factors and emotions specifically, in contrast to LMs which we have already discussed. Recognising the differences between human cognition (and the key role of emotion) and LM stochastic operation can shed light on the potential for exploiting emotion in detecting synthetic text.

2.3.1 Human Cognition is Dual Process, Bimodal, and Biased

We now examine human *cognition* and find it is *dual process* (Wason and Evans, 1974; Kahneman, 2003), bimodal (Dijksterhuis et al., 2006) and biased (Kanouse and Hanson, 1972; Tversky and Kahneman, 1973; Carstensen and Mikels, 2005; Slovic et al., 2007). We will see that certain aspects of cognition are influenced by emotion which can be susceptible to manipulation and can influence our judgements and the decisions we make. Cognitive biases also present problems for accurate dataset annotation, and thus for collators and annotators of datasets, and the developers and users of systems (including LMs) built upon those datasets. In order to differentiate human text from synthetic text it may be instructive to differentiate human cognition from the mechanics of LMs.

The *dual process theory* of human reasoning describes human thought as, not one, but two processes (or sets of processes) (Wason and Evans, 1974). These two processes are called by the generic names *System 1* and *System 2* (Stanovich and West, 2000). Dual process theory describes System 1 as being intuitive and System 2 as being reasoned (Kahneman, 2003). System 1 is fast, emotional, instinctive and does not lend itself to introspection - it happens automatically. System 2 is slower, more deliberative and gives a sense of agency - you make choices. Moral judgements are intuitive (System 1) and moral reasoning (System 2) occurs post-hoc after a judgement has already been reached (Haidt, 2001) suggesting that it is emotion, not reason or logic, which drives our judgements. This is a stunning proposition and turns on its head the notion that we rationally deliberate and make decisions, but rather we are driven by intuitive emotional responses and subsequently rationalise these instinctive choices. Kahneman (2003) describes *intuitive judgements* as those judgements generated by System 1 and not modified by System 2. Our judgements can have an obvious impact on our decision-making.[17] Planning is heavily influenced by emotion (Camerer et al., 2005); and according to the *somatic marker hypothesis* emotional processes guide behaviour particularly decision making (Damasio et al., 1996). The *affect heuristic* describes how our emotional state can drive our judgements and decision making, and that these affective responses are rapid and automatic (Slovic et al., 2007) typical of System 1. Thus, our cognitive reliance on emotion could expose us to manipulation via emotion and emotive language. Furthermore, our cognition and resulting perception of the world can be influenced by biases such as negativity bias, availability bias, and our attention (e.g. *associative incoherence* (as opposed to *associative coherence*), *unconscious thought theory*). We briefly consider some of these biases next.

Negativity bias causes negative events to have a greater effect on our psychological state (Kanouse and Hanson, 1972), to be more contagious (Rozin and

---

[17] For a detailed treatment of judgement and decision-making see Eysenck and Keane (2020, Ch.13).

Royzman, 2001), and to form quicker and be less resistant to disconfirmation than similar strength positive or neutral events (Baumeister et al., 2001). In addition, the *positivity effect* is a developmental pattern whereby negative material is particularly salient in youth but is disfavoured with age (Carstensen and Mikels, 2005). This means younger people can be disproportionately affected by negative news and imagery.[18]

Availability bias skews our judgement regarding the frequency of events, the easier we find it to recall instances of an event the more frequent we judge those events to be (Tversky and Kahneman, 1973). This bias has been shown to persist even when specifically instructed to avoid it, and possible sources include disproportionate exposure (e.g. *mere exposure effect*) and memorability (Lichtenstein et al., 1978). This can be a problem when we are repeatedly exposed to a subset of information based on *engagement based ranking* and recommender algorithms as described by Haugen (2021).

When the *associative coherence* of System 1 fails we are confounded and look for coherence by invoking System 2 resulting in a cognitive load that significantly diminishes performance on subsequent tasks (Afiki and Bar, 2020). Errors in judgement and decision making can occur because System 1 generates a faulty intuition, perhaps due to bias, which System 2 subsequently fails to detect and correct (Morewedge and Kahneman, 2010). Our cognitive process has flaws which can contribute to believing misinformation.

Cognition is also bimodal. *Unconscious thought theory* distinguishes between two modes of thought, namely conscious thought and unconscious thought, which have different characteristics making each mode preferable in different circumstances (Dijksterhuis and Nordgren, 2006). One such example is demonstrated in the *deliberation without attention effect* where simple choices (e.g. buying an oven glove) are better done with conscious thought and deliberation, but more complex choices (e.g. buying a car) are better done with

---

[18] In Appendix E we contrast the prevalence of negative and positive news.

unconscious thought (assuming you have gathered all relevant information) (Dijksterhuis et al., 2006).

Attention restoration theory asserts we have a reservoir of directed attention that requires conservation and restoration to allow for proper cognitive function; and our environment plays an important role (Kaplan, 2001). The *Zeigarnik effect* describes how interrupted tasks can dominate our attention (Zeigarnik, 1927) and negatively impact performance on a new task. Furthermore our attention reservoir is also diminished by *attention residue*. A compulsive internet user spending hours in the very engaging and distracting online world is likely to suffer from an attention deficit, and this can reduce their ability to identify misinformation.

Having considered some qualities, and vulnerabilities, of human cognition we turn to a crucial and primary quality of human cognition - emotion.

2.3.2 Emotions are a Particularly Human/Animal Trait

Emotions and their effect on humans have been discussed for millennia. Plato described people as attempting to drive a chariot drawn by two horses - emotion and reason (Plato, 370BCE, 246a-b). Aristotle said emotions change us and affect our judgements (Aristotle, 350BCE, II.1, 1378a). Aristotle espoused the importance of emotions to persuade a crowd (Aristotle, 350BCE, I.2, 1356a) which resonates with the *Surprise Negativity Anger Paradigm* described earlier (§2.2.3). Hsun Tzu (Xunzi) believed man's nature is evil and our intrinsic emotions must be controlled (Tzu, 250BCE, 23).

In the nineteenth century different systems of emotions were proposed, including: discrete emotions such as fear, anger, disgust etc. (Darwin, 1872); and dimensions of emotions such as pleasantness and intensity (Wundt, 1896). Both discrete emotions and dimensional emotions are used in NLP. Discrete emotions in NLP are now most commonly associated with Ekman's 6 basic emotions - joy, sadness, fear, anger, disgust and surprise (Ekman, 1992, 1999). Dimensional emotions in NLP are now most commonly associated with the Plutchik Wheel of Emotion (Plutchik, 1980, 2001), as illustrated in Figure 2.1. The circumplex (i.e. emotions arranged in a circle) represents degrees of similarity between emotions, the concentric rings represent emotion intensity, while the emotions in the blank spaces (on the outermost circle) are the primary dyads i.e. emotions that are a mixture of the two adjacent emotions. The 8 emotions in the middle ring of the three concentric rings, are the 'primary emotions' and are used in this thesis and by other affective NLP researchers using Plutchik's model. The closer two emotions are on Plutchik's Wheel - the more closely related those two emotions are. For example, the primary emotions 'anger' and 'disgust' are adjacent to one another and are therefore closely related; while 'joy' and 'sadness' are on opposing spokes and are therefore opposing emotions, according to Plutchik (1980, 2001). Our work uses both systems of emotion.

LMs can create embeddings for words which represent emotions, but that is not to say LMs have "knowledge" or "understanding" of emotion. We discussed in §2.2.1 the inability of GloVe and word2vec embeddings to distinguish opposing emotions, and attempts to increase the affective information in non-contextualised embeddings (Seyeditabari and Zadrozny, 2017; Agrawal et al., 2018; Kratzwald et al., 2018; Seyeditabari et al., 2019; Wang et al., 2020a). Other work addresses the human expectation of affective knowledge by trying to bolt it on to dialogue systems (Huang et al., 2018; Rashkin et al., 2019; Wang et al., 2020b). All of this work is an implicit acknowledgement that LMs lack knowledge and understanding of emotion. RQ1 attempts to address the affective deficit of LMs by finetuning on emotion classification.

Having examined emotions and their role in human cognition, we now pose the question - can emotions help in the difficult task of synthetic text detection?

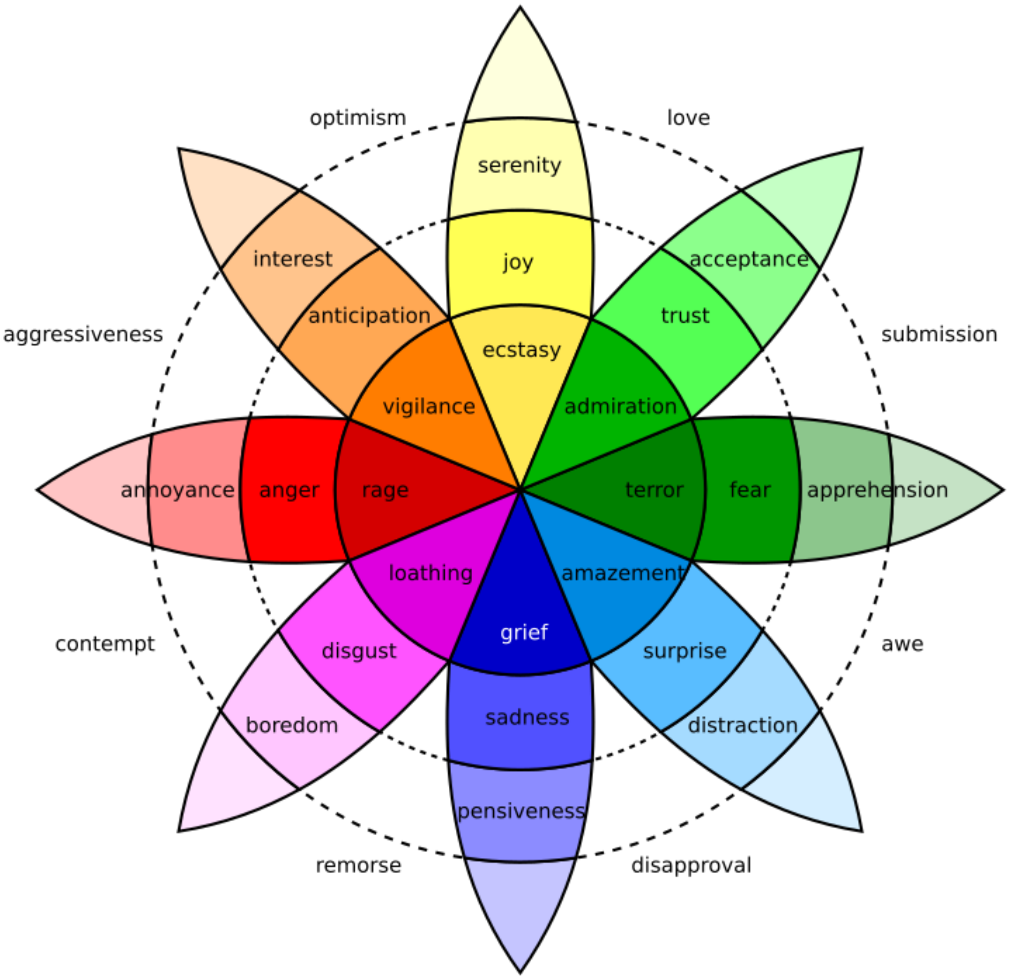


Figure 2.1: Plutchik Wheel of Emotion. The middle ring contains the 8 primary emotions, which includes Ekman's 6 emotions ('joy', 'sadness', 'fear', 'anger', 'disgust', 'surprise'.) plus 'trust' & 'anticipation'. Similar emotions lie on adjacent spokes e.g. anger-disgust, while opposing emotions are placed on opposing spokes e.g. joy-sadness. Image from (Plutchik, 2001), downloaded from `https://en.m.wikipedia.org/wiki/File:Plutchik-wheel.svg`.

2.3.3 Can Emotions Help Detect Synthetic Text?

We previously highlighted deficiencies in LMs with respect to words representing emotions, and highlighted affective deficiencies in the synthetic text generated by LMs - as evidenced by attempts to control emotion in LM generated synthetic text. We reiterate our hypothesis that LMs have an *affective deficit*, which can result in generating synthetic text that is affectively incoherent. Now, we discuss some differences in the production of synthetic text and human text, differences which our work attempts to exploit in the task of synthetic text detection. As with all aspects of cognitive psychology, human language production, particularly writing, is an active area of research which is complex and still evolving. As Eysenck and Keane (2020) state: "*We know more about language comprehension than language production*".

The writing model of Hayes (2012) begins with "motivation" followed by "goal setting". Hayes (2012) included "motivation", despite originally omitting it, because he was convinced it was not possible to adequately account for how people write without considering motivation. We suggest that LMs lack both "motivation" and "goal setting". One could argue that "goal setting" for a human is akin to "prompting" an LM. We disagree because of the lack of intent on the part of the LM versus the human. But, even if one were to concede that point, there is still no LM equivalent for the human "motivation" in writing. Furthermore, according to Knaller (2017): "*One can say that writing processes are inherently linked to emotions. As a psychophysical process, writing relates to emotions as a prerequisite for self- and other-experiences, judgments, evaluations, understanding and perception*". We have already seen the primacy of emotion in human cognition, in contrast to the affective deficit of LMs; now we can add human motivation and the lack of motivation for LMs as another contrast in text generation.

The choice of statistical *sampling strategy* determines the synthetic text generated by LMs, with no reference to meaning (Bender et al., 2021). In contrast, human intuitions and decisions are not statistical (Tversky and Kahneman, 1971) humans are driven by emotion and can be irrational (Neisser, 1963; Kahneman and Tversky, 1979; Thaler, 1980; Damasio et al., 1996; Haidt, 2001; Camerer et al., 2005; Slovic et al., 2007). This suggests that any distribution of human text is likely to be far more complex than that represented by LMs or sampling strategies, and this is evident in the lack of variance of some synthetic text compared to human text (Holtzman et al., 2020). In addition, the perplexity of synthetic text increases with sequence length compared to human text (Zellers et al., 2019) i.e. perplexity increases at a faster rate than the increase in synthetic text length. Sampling strategies can mitigate these variances but they come at the cost of verifiability (Massarelli et al., 2020) and add artefacts that make the synthetic text easier to detect (Ippolito et al., 2020). The intent of human authors (e.g. truth, hoax, propaganda) can be detected using lexical features (71% Accuracy for fake vs real) (Rashkin et al., 2017; Horne and Adali, 2017) however synthetic text tends to be stylistically consistent (Schuster et al., 2020). Emotions are the core of human cognition and, according to Knaller (2017) inherently linked to writing, in contrast, they are absent from the mechanics of LMs.

These crucial differences between human cognition and LMs may present an opportunity for synthetic text detection. Humans do not deliberately choose words based on statistics - LMs are probabilistic and stochastic. LMs attempt to mimic human text using mathematics, statistics and stochasticity; and as Chomsky (1956) noted, statistical techniques could generate grammatically correct but meaningless sentences like "*colourless green ideas sleep furiously*". Sixty five years later, and in reference to the latest LMs, a similar criticism is stated by Bender et al. (2021) who refer to LMs as "*stochastic parrots*". Human cognition is dual-process and bimodal - current LMs are not capable of cognition. Human writing is borne from motivation - LMs are devoid of motivation. Humans are driven by emotion - LMs are effectively devoid of emotion.

Given the hypothesis that emotions can help in the task of synthetic text detection, we next consider how to classify emotions.

2.3.4 Classifying Emotions is Difficult.

Emotion classification is a difficult task for humans. The difficulty inherent in attempting a subjective assessment of emotions is evident from our previous discussion of human cognition and bias. In particular, we recall that *System 1* does not lend itself to introspection and *System 2* attempts a post-hoc justification of our feelings and intuitions (Wason and Evans, 1974; Kahneman and Tversky, 1979; Thaler, 1980; Stanovich and West, 2000; Evans and Stanovich, 2013). In practice, the difficulty of emotion classification is reflected in the low level of inter-annotator agreement in emotion datasets (Strapparava and Mihalcea, 2007, 2008; Bostan et al., 2020; Troiano et al., 2021) for example a Pearson correlation ranging from 0.36 to 0.68 for different emotions within the same dataset (Strapparava and Mihalcea, 2007). This also presents a difficulty for automated classification of emotions, in part because the "gold" annotations in datasets may not be reliable or consistent (Clark et al., 2021), can vary across cultures (An et al., 2017), be affected by annotator mood (another example of the *affect heuristic*) and the preceding examples (Shen et al., 2019), and are affected by annotator confidence and emotion intensity (Troiano et al., 2021). Another difficulty is the relatively small size of emotion datasets. Small datasets can result in more volatile outcomes (Tversky and Kahneman, 1971), for example finetuning is more susceptible to hyperparameter selection for small datasets than larger datasets (Devlin et al., 2019). Other difficulties include imbalanced datasets, different domains (news, blogs, social media, etc.), and the different emotion models such as (Plutchik, 1980, 2001; Ekman, 1992, 1999) used in dataset annotation (Bostan and Klinger, 2018), and sentiment (Ortony et al., 1988).

Datasets are crucial for (pre-training and) fine-tuning language models. We choose two commonly used datasets which are most relevant to our work, and modify them where necessary to match our requirements. Those datasets are AffectiveText (AT) (Strapparava and Mihalcea, 2008) and GoodNewsEveryone (GNE) (Bostan et al., 2020). These datasets are in the target domain (news) and language (English) and are labelled with appropriate emotions. AT is labelled with Ekman's 6 emotions, while GNE is labelled with 15 emotions which we map to Plutchik's 8 emotions, Ekman's 6 emotions, and binary sentiment, as required. In some experiments we combine the datasets, balance them across labels (emotions), and map the emotions as described above for our experiments.

Emotion classification shares a similar development history to some other NLP classification tasks. From keyword identification using lexicons like `EmoLex` (Mohammad and Turney, 2013) and `EmoWordNet` (Badaro et al., 2018), to feature-based classifiers - as surveyed extensively by Alswaidan and Menai (2020) - like Support Vector Machines, to state-of-the-art transformer based (e.g. BERT and RoBERTa) PLMs (Chen et al., 2023; Shah et al., 2023), and prompt based techniques including Natural Language Inference (NLI) (Basile et al., 2021; Plaza-del Arco et al., 2022; Bareiß et al., 2024). We use the latter two techniques, i.e. state-of-the-art fine-tuned PLMs and prompt based NLI, for emotion classification.

In addition to emotion classification, we also use sentiment analysis. Specifically, we use binary sentiment as the topmost division of positive and negative valence as described by Ortony et al. (1988). Sentiment analysis attempts to analyse text by, as Acheampong et al. (2020) put it "*extracting opinions, ideas and thoughts*" and assigning them a valence/polarity that is either positive or negative (or neutral).

There are no suitable emotion classification models which satisfy our specific requirements (news domain, English language, Plutchik's 8 emotions and/or

Ekman's 6 emotions). Consequently, we will fine-tune several LMs in the task of emotion classification for RQ1 in Chapter 3, and RQ3 in Chapter 5. In contrast, in §5.4.3 we utilise 15 existing zero-shot Natural Language Inference LMs for the task of emotion classification with no need for fine-tuning. Next, we discuss Natural Language Inference, zero-shot, and the importance of prompting.

Premise	`‘How do you know? All this is their information again.’`
Hypothesis	`‘This information belongs to them.’`
Label	`entailment`

Table 2.1: An example of Natural Language Inference from the MultiNLI dataset (Williams et al., 2018). A *premise* and *hypothesis* are used as prompt to the NLI model, which responds with either *entailment, contradiction, or neutral.* In this example, `‘entailment’` is the gold label, meaning that the hypothesis can be inferred from the premise.

Another approach to emotion classification which we use is zero-shot Natural Language Inference (NLI). First we will explain NLI, followed by the meaning of zero-shot NLI, and explain the importance of prompting. In NLI the prompt to the LM consists of a *premise* and *hypothesis* (or *hypotheses*), and the LM response will imply *entailment*, *contradiction* or *neutral.* The NLI task can be stated as follows: Given a *premise*, does the *hypothesis* logically follow (*entailment*), or logically unfollow (*contradiction*), or is undetermined (*neutral*). An example from the MultiNLI dataset (Williams et al., 2018) is shown in Table 2.1. In the example, `‘entailment’` is the gold label, meaning that the hypothesis logically follows (i.e. can be inferred from) the premise. If, however, we change the *hypothesis* to `‘This is not their information.’` then the NLI model might respond with *‘contradiction’* because the new hypothesis appears to contradict the premise. If, instead, we change the hypothesis to `‘The window is closed.’`, then the NLI model might respond with *‘neutral’* because the hypothesis cannot be determined from the premise.

Table 2.2 shows another example, this time for emotion classification which we use in §5.4.3. The prompt consists of: the *premise* `‘Ukraine` `election:Comedian Volodymyr Zelensky wins’`; and the *hypotheses* `[‘joy’, ‘sadness’, ‘fear’, ‘anger’, ‘disgust’, ‘surprise’, ‘trust’, ‘anticipation’]`. The response from the LM consists of a SoftMax’ed probability for each hypothesis - we take the hypothesis with the highest probability as *entailment* i.e. the prediction of the model. NLI provides an alternative approach to fine-tuning for the task of emotion classification.

Premise	`‘Ukraine election:Comedian Volodymyr Zelensky wins’`	
Hypotheses	`[‘joy’, ‘sadness’, ‘fear’, ‘anger’, ‘disgust’, ‘surprise’, ‘trust’, ‘anticipation’]`	
Entailment	`disgust`	`0.0039077484980225560`
	`anger`	`0.0044322255998849870`
	`fear`	`0.0047090123407542706`
	`sadness`	`0.0051015573553740980`
	`trust`	`0.0416014716029167200`
	`anticipation`	`0.0712974146008491500`
	`surprise`	`0.1438306123018264800`
	`joy`	`0.7251200079917908000`

Table 2.2: An example of zero-shot NLI (§5.4.3). A *premise* (sentence) and *hypotheses* (emotions) are used as prompt to the model, which returns an *entailment* (prediction) i.e. a soft-maxed probability for each *hypothesis*. The *hypothesis* (emotion) with the highest probability is *entailed* (predicted) from the *premise* (sentence). In this example, `‘joy’` is the prediction of the zero-shot NLI model.

Zero-shot means the LM has not received any training or fine-tuning examples from which to learn. This is in contrast to fine-tuning, where we typically give 1,000’s of labelled examples to the LM so that it can ‘learn’ the correct prediction. Performing a task zero-shot has a number of advantages over fine-tuning (Laurer et al., 2024), for example:

1. There is no need to fine-tune or validate the LM, saving resources such as time, compute, and data.
2. Hypotheses (emotions in our case) are specified at inference time, so it is more flexible - users can specify Plutchik, Ekman, sentiment etc. without having to fine-tune the model again.
3. There is no requirement for labelled data (we already discussed the difficulty with emotion datasets and annotator agreement for labels).

4. There is no need for training and validation data, we can use the test data only, thus reducing the data requirement.

Prompting plays a pivotal role in the performance of LMs. We use a prompt type `EMO-Name` as defined by Plaza-del Arco et al. (2022) for zero-shot NLI emotion classification. The `EMO-Name` prompt is the prompt shown in our example above and Table 5.7. We chose this prompt for its simplicity and potential usefulness as a suitable baseline for future prompt-engineering experiments. While we use some multilingual models for zero-shot NLI, we continue to use English prompts because Bareiß et al. (2024) show that prompting in English performs better than target language prompts for NLI-based zero shot emotion classification across a range of scenarios (e.g. different data languages, different LMs, different prompt types).

We considered human factors, particularly emotion, in cognition, hypothesised about the use of emotion in synthetic text detection, and considered the difficulty posed in emotion classification. We now turn our attention to commonly used metrics, and a novel metric we will introduce for RQ2 in Chapter 4.

2.4 Measuring and Metrics

We now consider suitable metrics for measuring emotions, and evaluating classifier performance. F-score is the most frequently used metric in NLP emotion classification literature (Alswaidan and Menai, 2020), often along with its component Precision and Recall scores, and Accuracy (Kusal et al., 2023). The measurements for these metrics are binary, either a prediction is correct or incorrect. This makes sense in a typical classification scenario where classes are nominal and discrete - either you predicted the correct class or you didn't. In Chapter 3 we will use these metrics when classifying Ekman's 6 basic emotions, and when classifying news article authors as human or synthetic.

However, for affective profiling in Chapter 6 we will use Plutchik's 8 emotions (Plutchik, 1980, 2001) as illustrated in Plutchik's Wheel of Emotion in Figure 2.1. In contrast to Ekman, who essentially treats the 6 basic emotions as a *bag-of-emotions*, Plutchik includes a dimensional relationship between the 8 emotions. To use the language of Stevens (1946) "*theory of scales of measurement*", as shown in Table 2.3, Ekman's 6 basic emotions use a "*nominal*" scale, whereas Plutchik's 8 emotions use an "*ordinal*" scale. Plutchik's Wheel of Emotion orders the 8 emotions such that similar emotions are adjacent to each other while dissimilar emotions are further apart. Thus, according to Plutchik, there is a dimensional relationship between the emotions.

Level	Name	Example	Description
1	`Nominal`	Ekman 6 Basic Emotions	Named items
2	`Ordinal`	Plutchik Wheel of Emotion	+ ordered items
3	`Interval`	Our 'Enhanced' Plutchik	+ intervals between items
4	`Ratio`	Enhanced Plutchik + neutral?	+ absolute zero

Table 2.3: Table showing Stevens (1946) 4 levels of Measurement. Ekmans 6 basic emotions are `Nominal` - simply named emotions, almost like a "*bag of emotions*". Plutchik's 8 emotions are `Ordinal` - named and ordered around the Plutchik Wheel. Our 'Enhanced Plutchik' is `Interval` - by including intervals between emotions - which we introduce in Chapter 4. Future Work could elevate our 'Enhanced Plutchik' to `Ratio` by including 'Neutral' emotion as absolute zero.

The dimensional relationship between Plutchik's 8 emotions gives rise to the question - is it possible to exploit the dimensional relationship between Plutchik's 8 emotions and employ metrics other than the binary correct-incorrect metrics like F1-score etc.? Non-discrete metrics (for example error-based metrics) such as *Mean Absolute Error* (MAE) and *Root Mean Squared Error* (RMSE) are used to evaluate models which generate continuous predictions, such as linear regression. Continuous metrics are not binary (correct or incorrect), they vary depending on how close the prediction is to the correct value. For example, for a true value of 50, predicting 42 is better than predicting 31 and continuous metrics reflect this (while discrete metrics do not, the predictions are both considered incorrect). The problem when attempting to use these metrics with Plutchik's emotions is the absence of a numerical value assigned to the emotions. Turning to Stevens (1946) we can say that, in order to use continuous metrics for Plutchik's 8 emotions, we need to elevate Plutchik's wheel of Emotions from the Ordinal scale to the Interval scale i.e. we need to find a way to assign meaningful numerical values to each of Plutchik's 8 emotions. We address this problem in RQ2 and describe our proposed solution in Chapter 4. We then employ our solution in Chapters 5 and 6 and exploit the dimensional aspect of Plutchik's 8 emotions in the task of emotion classification and the affective profiling of human and synthetic text respectively.

Correlations: Pearson ρ and Spearman ρ Throughout this thesis we will have a set of measurements, such as Accuracy, that we want to compare: typically human versus synthetic. We want to measure the correlation between the human article and the synthetic article. For this purpose we use Pearson ρ correlation and Spearman ρ correlation, both of which have values in the range -1 to $+1$. Negative values indicate a negative correlation i.e. as one variable increases, the other variable decreases, and vice versa. Positive values indicate a positive correlation i.e. the variables tend to increase together. The magnitude of the value indicates the strength of the correlation between the variables. Pearson ρ measures the linear relationship between two sets of values i.e. how well they increase or decrease together; whereas Spearman ρ is computed on ranks and measures the monotonic relationship between the two sets of values (ranking). Table 2.4 shows the strength of relationship corresponding to various ranges of ρ correlation values.[19]

2.5 Conclusion

We examined the problem of synthetic text detection from the automatic and human perspectives. We discussed how LMs are powerful, biased and dual use,

---

[19] Taken from `https://www.westga.edu/academics/research/vrc/assets/docs/scatterplots_and_correlation_notes.pdf`

Absolute Value of ρ	Strength of Relationship
$r < 0.3$	None or very weak
$0.3 < r < 0.5$	Weak
$0.5 < r < 0.7$	Moderate
$r > 0.7$	Strong

Table 2.4: Table showing absolute value of ρ and the strength of relationship.

and how humans are dual process (driven by emotion), bimodal and biased. We have imperfect datasets, annotated by imperfect people or other imperfect means. We use these imperfect datasets to train imperfect LMs, which are capable of generating synthetic text at scale. Therefore, the task of synthetic text detection is crucial, even though it is difficult. This research attempts to contribute to that task by leveraging the differences between LMs and human cognition - particularly where emotion is concerned. In the next four chapters we answer our four research questions, and describe our experiments to address the task of synthetic text detection.

Chapter 3

Improving Neural Detection of Synthetic Text via Emotion Recognition

In this chapter we answer Research Question 1: "*Can the performance of automatic synthetic text detectors be improved by making them 'emotionally-aware' compared to standard detectors?*". RQ1 was described in §1.2.

It begins with a description of our experimental Method (§3.1) and Experiments (§3.2) followed by our primary Results (§3.3). We provide Analysis (§3.4) with several further experiments, including detecting synthetic text generated by ChatGPT (§3.5). The chapter finishes with our Conclusions (§3.6).

3.1 Method: Equipping PLMs with Emotional Intelligence

Our method is illustrated in Figure 3.1. The process works as follows:

1. PLMsynth: In the leftmost column of Figure 3.1, human articles and synthetic articles are used to fine-tune a PLM to discriminate between the two kinds of text. This is indicated by the blue nodes in the PLM.

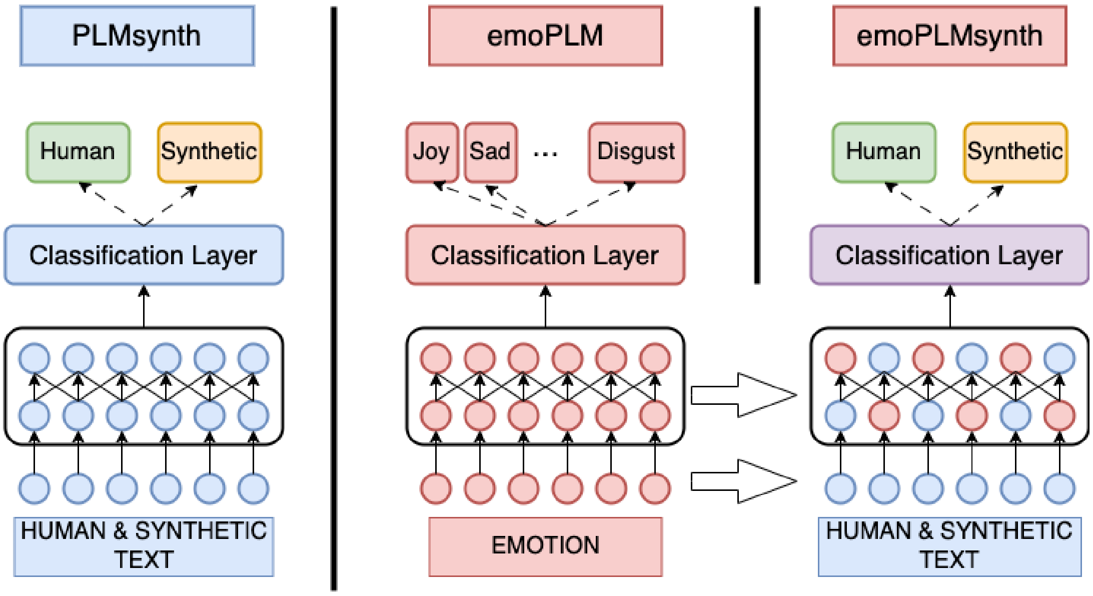


Figure 3.1: The emotionally-aware PLM (`emoPLMsynth`) takes advantage of its prior fine-tuning on emotion to improve performance on the task of synthetic text detection. In contrast, the standard PLM fine-tuned only on synthetic text detection (`PLMsynth`) has no training on emotion. Our experiments show the emotionally-aware PLM (`emoPLMLsynth`) outperforms the standard PLM (`PLMsynth`) in multiple scenarios.

2. EMOPLM: In the middle column of Figure 3.1, a second dataset annotated for emotions with Ekman's 6 emotions (Ekman, 1992, 1999, 2016) is used to fine-tune a PLM on the task of emotion classification. This makes our model emotionally-aware, as indicated by the red nodes in the PLM.

3. EMOPLMSYNTH: The multi-class (6 head) classification layer from `emoPLM` is removed and replaced with a binary classification layer. The emotionally-aware PLM is then fine-tuned on the task of discriminating between human and synthetic articles. The PLM is still emotionally-aware while also being able to detect synthetic text - as indicated by the red and blue nodes respectively in the PLM.

We conduct experiments using various PLM sizes, architectures, datasets, and domains for synthetic text generation (e.g. Grover, ChatGPT) and detection (e.g. BERT, BLOOM, ChatGPT), which we now describe.

3.2 Experiments in the News Domain

3.2.1 Generator and Detector Models

To generate synthetic text, we use the Grover causal PLM (GPT-2 architecture[1]) pretrained on 32M news articles from the RealNews dataset (Zellers et al., 2019). Heretofore, we referred to our detector models using the general term PLM, e.g. `emoPLMsynth` and `PLMsynth`, but we must choose a specific PLM for our experiments. We choose BERT (Devlin et al., 2019) as the PLM for our main detector model since it is freely available and performs well in several tasks including sequence classification. Specifically, we choose $\text{BERT}_{\text{base}}$-cased as the PLM for our detector model. [2] A baseline BERT model (we call this `BERTsynth`) is fine-tuned on the task of synthetic text detection, while our proposed model is the same BERT model, firstly fine-tuned on emotion classification (we call this intermediate model `emoBERT`) before further fine-tuning for synthetic text detection. This final proposed model is referred to as `emoBERTsynth`.

3.2.2 Datasets

We create and release *NEWSsynth*, a dataset containing 10k human and 10k synthetic news articles. 10k human-authored news articles were taken from the RealNews-Test dataset (Zellers et al., 2019) and used as a prompt to $\texttt{Grover}_{\texttt{base}}$ to generate a corresponding 10k synthetic articles. The prompt includes the news article, headline, date, author, web domain etc. as described by Zellers et al. (2019). The dataset was split 10k-2k-8k for train, validation, and test respectively, the same ratio used by Zellers et al. (2019) with 50:50 human:synthetic text in each split, see Appendix A for details. An investigation

---

[1] $\text{GPT-2}_{\text{small}}$ is a decoder based transformer architecture with 12 stacked decoders and 12 attention heads, totalling 117M parameters (Radford et al., 2019).

[2] $\text{BERT}_{\text{base}}$-cased uses a transformer architecture with 12 stacked encoders and 12 bidirectional attention heads, totalling 110M parameters. It is ‘cased’ because it is pretrained on data which includes upper and lower case text (Devlin et al., 2019) which aligns with our datasets.

of the length of human vs synthetic text is provided in §3.4.1.

In a second experiment, we also use the full RealNews-Test dataset itself, which comprises the same 10k human news articles used in *NEWSsynth* and 10k synthetic articles generated by $\texttt{Grover}_{\texttt{mega}}$. The use of synthetic text generated by $\texttt{Grover}_{\texttt{mega}}$ instead of $\texttt{Grover}_{\texttt{base}}$ facilitates comparison of `BERTsynth` and `emoBERTsynth` on text generated by a larger generator model, and against results reported for other models on this dataset.

We use the GoodNewsEveryone dataset (Bostan et al., 2020) to train `emoBERT`. This dataset contains 5k news headlines, and was chosen since it is within the target domain (news) and language (English) and is annotated with categorical emotions. The 15 emotion labels from GoodNewsEveryone were reduced to 11 emotions using the mapping schema of Bostan and Klinger (2018), and further reduced to 6 emotions based on the Plutchik Wheel of Emotion (Plutchik, 1980, 2001) – see Table 3.1, and Figure 2.1 in Chapter 2.3.2 – resulting in 5k news headlines labelled with Ekman's 6 basic emotions, the most frequently used categorical emotion model in psychology literature (Ekman, 1992, 1999, 2016).

GoodNewsEveryone		Ekman	%
disgust	→	disgust	(8%)
fear	→	fear	(8%)
sadness, guilt, shame	→	sadness	(14%)
joy, trust, pride, love/like, positive anticipation/optimism	→	happiness	(17%)
anger, annoyance, negative anticipation/pessimism	→	anger	(24%)
negative surprise, positive surprise	→	surprise	(30%)

Table 3.1: Emotion Mapping Schema: GoodNewsEveryone (15 emotions) to Ekman 6 basic emotions. % shows the emotion label distribution in the dataset.

3.2.3 Fine-Tuning `BERTsynth`

We fine-tune a `BERT`$_{\texttt{base}}$`-cased` model for synthetic text detection (using the *NEWSsynth* or RealNews-Test dataset) the resulting model is called `BERTsynth`. Input sequence length was maintained at the BERT maximum of 512 tokens ($\approx$ 384 words). Five training runs were conducted. Each training run was 4 epochs – the most possible within GPU time constraints and similar to those of Zellers et al. (2019) who used 5 epochs.[3] For each training run, a unique seed was used for model initialisation, and a unique set of three seeds were used for the dataset shuffle - one seed each for train, validation, and test splits. Furthermore, the HuggingFace library shuffles the training data between epochs. The reproducibility of the training and validation results using seeds was verified by conducting multiple runs of training and validation. Hyperparameter values are listed in Appendix B.

3.2.4 Fine-Tuning `emoBERT`

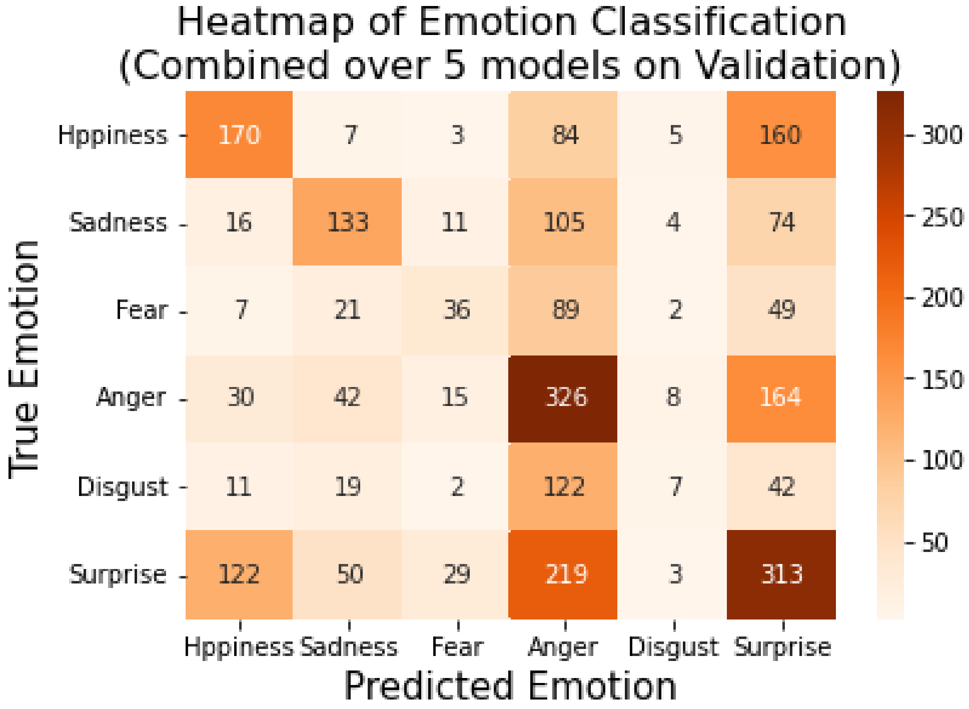
We fine-tune a `BERT`$_{\texttt{base}}$`-cased` model on the single label multiclass task of emotion classification using the GoodNewsEveryone dataset, the resulting model is called `emoBERT`. Fine-tuning `emoBERT` followed a similar process to fine-tuning `BERTsynth` described in §3.2.3. This time, there were 5k examples and fine-tuning was for 10 epochs.

Classification accuracy is not the end goal for `emoBERT`. Its purpose is to reduce the affective deficit of the PLM by modifying the representations of words conveying emotions and to improve performance in the task of synthetic text detection by transfer learning. The mean $F1_\mu$ for `emoBERT` is 39.4% on the validation set - more than double mean chance (16.7%) and within the range 31% to 98% reported by Bostan and Klinger (2018) for within-corpus emotion classification in UnifiedEmotion. Bostan et al. (2020) do not report the news

---

[3] After each epoch the model (checkpoint) was run against the validation set for Accuracy, and the checkpoint and Accuracy results were saved (in addition to F1, Precision and Recall). The checkpoint with the highest Accuracy score was then run on the test set.

headline emotion classification metrics for GoodNewsEveryone.



	Hppiness	Sadness	Fear	Anger	Disgust	Surprise
Hppiness	170	7	3	84	5	160
Sadness	16	133	11	105	4	74
Fear	7	21	36	89	2	49
Anger	30	42	15	326	8	164
Disgust	11	19	2	122	7	42
Surprise	122	50	29	219	3	313

Figure 3.2: Combined Confusion Matrix for Emotion Classification on GoodNewsEveryone validation set.

Figure 3.2 depicts the combined results of the best performing model (in validation) from the 10 epochs, in each of the 5 training runs. The imbalance across emotion labels (shown in the first column of Table 3.1) is reflected in performance in Figure 3.2. Anger and Surprise are the two emotions best classified and best represented in the dataset at 24% and 30% respectively; while Fear and Disgust are the two emotions most poorly classified and least represented in the dataset at 8% each. The 4 emotions Happiness, Sadness, Anger, and Surprise are classified correctly more often than as any of the other 5 emotions. Fear and Disgust are most likely to be misclassified as Anger.

We see a correlation between class size and performance on that class - those classes with more examples performed better than those with fewer examples. To reiterate, outright performance is not the end goal for `emoBERT`. The purpose of `emoBERT` is to reduce the *affective deficit* of the PLM by modifying the word representations of words representing emotions and to improve performance in the task of synthetic text detection by transfer learning.

3.2.5 Fine-Tuning `emoBERTsynth`

We fine-tune an `emoBERT` model for synthetic text detection (using the *NEWSsynth* or RealNews-Test dataset) - the resulting model is called `emoBERTsynth`. The best `emoBERT` model (checkpoint) from each of the 5 training runs had its emotion classification head (6 outputs) replaced with a binary classification head (2 outputs) for human vs synthetic text classification, see Figure 3.1. Each model was then fine-tuned on the synthetic text detection task using the exact same process and set of random seeds (for dataset shuffling) as the 5 best models described in §3.2.3. This provided a direct comparison between the 5 `BERTsynth` models (trained on synthetic text detection only) and the 5 `emoBERTsynth` models (fine-tuned on emotion classification followed by synthetic text detection).

3.3 Results

	Precision		Recall		F1		Accuracy	
Run	Bs	emoBs	Bs	emoBs	Bs	emoBs	Bs	emoBs
1	80.30	81.25	92.40	92.20	85.92	86.38	84.86	85.46
2	82.26	84.30	90.90	89.83	86.37	89.77	85.65	86.55
3	78.01	82.88	92.40	88.20	84.60	85.45	83.18	84.99
4	77.44	85.84	94.85	88.20	85.27	87.00	83.61	86.83
5	86.09	87.14	87.58	86.75	86.83	86.95	86.71	86.98
Mean	80.82	**84.28**	**91.63**	89.04	85.80	**87.11**	84.80	**86.16**
Var.	(9.89)	(4.35)	(5.70)	(3.45)	(0.62)	(2.08)	(1.68)	(0.63)
Δ	+3.46		-2.59		+1.31		+1.36	

Table 3.2: Comparison of `BERTsynth` (Bs) and `emoBERTsynth` (emoBs) against the *NEWSsynth* test set. (Variance is shown in brackets under the mean). emoBs outperforms Bs in head-to-head for all 5 runs in Accuracy, F1, and Precision; while Bs outperforms emoBs in head-to-head for all 5 runs in Recall.

The results in Figure 3.3 and Table 3.2 show the performance of `BERTsynth` and `emoBERTsynth` when fine-tuned on the *NEWSsynth* dataset. The results support the hypothesis that emotion can help detect synthetic text. `emoBERTsynth` outperforms `BERTsynth` in head-to-head for accuracy and F1 in all 5 runs.

Looking at precision and recall, `emoBERTsynth` outperforms `BERTsynth` in precision in all 5 runs, while the opposite is the case for recall. It is worth comparing the relative difference in recall and precision between `emoBERTsynth` and `BERTsynth` models in Table 3.2. `emoBERTsynth` has a difference between the mean recall and mean precision of 4.76 (89.04 - 84.28) while the difference for `BERTsynth` is more than double that at 10.81 (91.63 - 80.82). Thus, we suggest our emotionally-aware PLM, `emoBERTsynth,` is a better performing model than the standard PLM, `BERTsynth`, because it has a better balance between precision and recall.

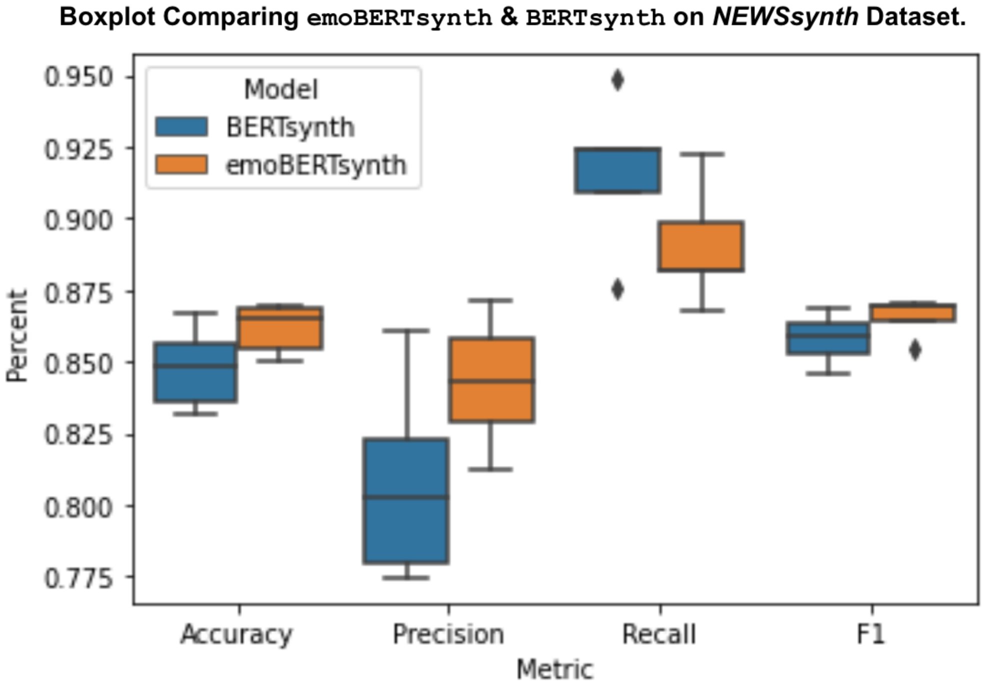


Figure 3.3: Test results for `BERTsynth` and `emoBERTsynth` on the *NEWSsynth* dataset. `emoBERTsynth` is higher for Accuracy, Precision and F1, while `BERTsynth` is higher for Recall.

In Table 3.3 we compare `BERTsynth` and `emoBERTsynth` on the RealNews-Test dataset. Recall that this dataset contains synthetic articles generated by $\texttt{Grover}_{\texttt{mega}}$ instead of the smaller $\texttt{Grover}_{\texttt{base}}$. We also compare against the `FastText`, `GPT-2` and `BERT` detector models reported by Zellers et al. (2019) on this dataset. `emoBERTsynth` has the highest accuracy, outperforming `BERTsynth` by 1.4%, `BERT`$_{\texttt{base}}$ by 9.03%, `GPT-2`$_{\texttt{base}}$ by 10.03%, and `FastText` by 12.43%. These results support the hypothesis that emotion can improve synthetic text detection.

There is a 7.63 point difference between our `BERTsynth` model and the BERT model reported by Zellers et al. (2019), despite both models being `BERT`$_{\texttt{base}}$ and fine-tuned on the same dataset and splits. However, there are differences in how the models were treated before this fine-tuning, and there may be some hyperparameter differences for fine-tuning. We described in §3.2.3 how we fine-tune a randomly initialised BERT model to create `BERTsynth`. Zellers et al. (2019) reported that their BERT models were domain adapted to News (by training on RealNews) at a length of 1024 WordPiece tokens. It is possible that this additional domain-adaptation and extended input sequence length actually harmed the performance of the `BERT`$_{\texttt{base}}$ model on the synthetic detection task. The performance of synthetic text detectors can improve with length (Ippolito et al., 2020) and the longer input sequence length could help in this regard. However, the vast majority of human and synthetic news articles in RealNews-Test are shorter than 1024 tokens. Thus, they may not benefit from that extended input length and the model may in fact be somewhat reliant on those later input tokens for prediction.

Size	Model	Acc.
11M	FastText	63.80
124M	GPT-2$_{\text{base}}$	66.20
	BERT$_{\text{base}}$	67.20
	`BERTsynth`	74.83
	`emoBERTsynth`	**76.23**

Table 3.3: `emoBERTsynth` outperforms other model architectures and sizes detecting human and Grover$_{\text{mega}}$ (1.5B) synthetic text from the RealNews-Test dataset. Detector model sizes include 11M and 124M parameters and architectures include `FastText`, `GPT-2`$_{\texttt{base}}$, and `BERT`$_{\texttt{base}}$. The `FastText, GPT-2`$_{\texttt{base}}$ `and BERT`$_{\texttt{base}}$ results are reported by Zellers et al. (2019).

Two Examples of emoPLMsynth vs PLMsynth In Chapter 1 we saw an example of affective incoherence in synthetic text compared to affective coherence in human text. Now we consider two examples where the emotionally-aware PLM outperforms the standard PLM on affectively incoherent synthetic text.

For the purpose of demonstrating the improved performance of our emotionally-aware PLM (`emoPLMsynth`) we select two examples of affectively incoherent synthetic text from the *NEWSsynth* dataset (see §3.2.2) which `emoPLMsynth` correctly identified as synthetic, but `PLMsynth` incorrectly identified as human.

Positive emotion words are highlighted in orange and negative emotion words in pink.

1. *Marshal Yanda looked "happy and refreshed" before Wednesday's minicamp practice but made some notes. After a disappointing season, the 37-year-old Yanda might be "just getting better and better and better" physically, veteran linebacker Whitney Mercilus said.*;
2. *Bethenny Frankel, who hosted her daughter, Brynn, on 'Bethenny Ever After,' opened up about how she found balance when dealing with two recent heartbreaks. Bethenny Frankel, 39, has had a rocky relationship with her child, Bryn, 11, but all the drama behind the scenes has her engaged in peace and happiness.*

In the first example, the sequence of highlighted texts are "*happy*" - "*disappointing*", and "*better and better and better*". This sequence can be classified[4] as a sequence of emotions: `[joy, sadness, joy]`. We refer to the sequence of emotions in an article as an *affective profile*. This particular affective profile displays *affective incoherence* because the emotions in the profile change substantially,[5] i.e. from joy to sadness and then back to joy.

---

[4] We use RoBERTa-large (as described in §5.4.1) for emotion classification in these two examples.

[5] In §6.4.3 we demonstrate the flexibility and extensibility of the ψ metric, introduced in

In the second example, the sequence of highlighted texts are "*two recent heartbreaks*" - "*rocky relationship*" - "*peace and happiness*". The affective profile is `[sadness, sadness, joy]`. This affective profile displays some affective coherence (sadness followed by sadness) and affective incoherence (sadness followed by joy).

The emotionally aware PLM (`emoPLMsynth`) correctly identified these two examples of synthetic text containing affective incoherence. In contrast, the standard PLM (`PLMsynth`) failed to identify them as synthetic and incorrectly labelled them as human.

3.4 Analysis

In this section, we perform a further set of experiments to aid in interpreting our main results.

3.4.1 Length of Human vs Synthetic articles

We investigate whether PLMs simply learn something about the length of articles as a proxy for discrimination between human and synthetic text. An analysis of *NEWSsynth* articles (train and validation splits) reveals no obvious correlation (Pearson $r = 0.19$) between the number of sentences in a human article and the number of sentences in the resulting synthetic article. Similarly, there is no obvious correlation (Pearson $r = 0.20$) between the number of words in a human article and the number of words in the resulting synthetic article. In terms of words per article - 64% of human articles are longer than their corresponding synthetic article, while 34% of synthetic articles are longer. Human articles are longer overall, but have slightly shorter sentences than synthetic text; and human articles have more sentences per article - which accounts for

---

Chapter 4, to measure and numerically quantify the affective coherence of an article. The first example above has a ψ score of 0% for affective coherence, while the second example has a ψ score of 50%. We leave the details, of calculating ψ for affective coherence, until §6.4.3.

their longer mean length. Similar observations were made for RealNews-Test by Bhat and Parthasarathy (2020). See Table 3.4 and Figs. 3.4 to 3.8. which illustrate the relative lengths of human and synthetic articles and sentences in *NEWSsynth* (train and validation splits). Overall, these results point neither to article length nor sentence length as a reliable discriminator for synthetic text suggesting that detector models are not simply learning length as a proxy for human vs synthetic text.

	Words per Article		Sentences per Article		Words per Sentence	
	$\overline{x}$	σ	$\overline{x}$	σ	$\overline{x}$	σ
Human	594.56	503.07	27.05	25.23	21.98	15.98
Synthetic	417.98	162.09	18.34	8.64	22.79	16.60
	Figure 3.6		Figure 3.7		Figure 3.8	

Table 3.4: Comparison of Human and synthetic text in the *NEWSsynth* dataset showing the mean ($\overline{x}$) and standard deviation (σ) for Words Per Article, Sentences Per Article, and Words Per Sentence. Human articles are longer overall, but have slightly shorter sentences than synthetic text; and Human articles have more Sentences Per Article - which accounts for their longer mean length.

3.4.2 Size of Fine-Tuning Splits

Split	Prec.	Recall	F1	Acc.
5-1-4k	78.39	79.85	78.89	78.58
Var.	(24.10)	(17.33)	(3.17)	(6.51)
10-2-8k	80.82	91.63	85.80	84.80
Var.	(9.89)	(5.70)	(0.62)	(1.68)
Δ	+2.43	+11.78	+6.91	+6.22

Table 3.5: `BERTsynth` metrics for different split sizes, using the *NEWSsynth* dataset averaged over 5 runs (with variance shown in brackets).

The `BERTsynth` fine-tuning regime (§3.2.3) was repeated using all (20k) and half (10k) of *NEWSsynth*. In all 5 runs, the `BERTsynth` model trained on the larger 20k dataset performed better than the equivalent model trained on the smaller 10k dataset – see Table 3.5. There was a modest improvement in precision (+2.43%) with a much larger increase in recall (+11.78%). These results

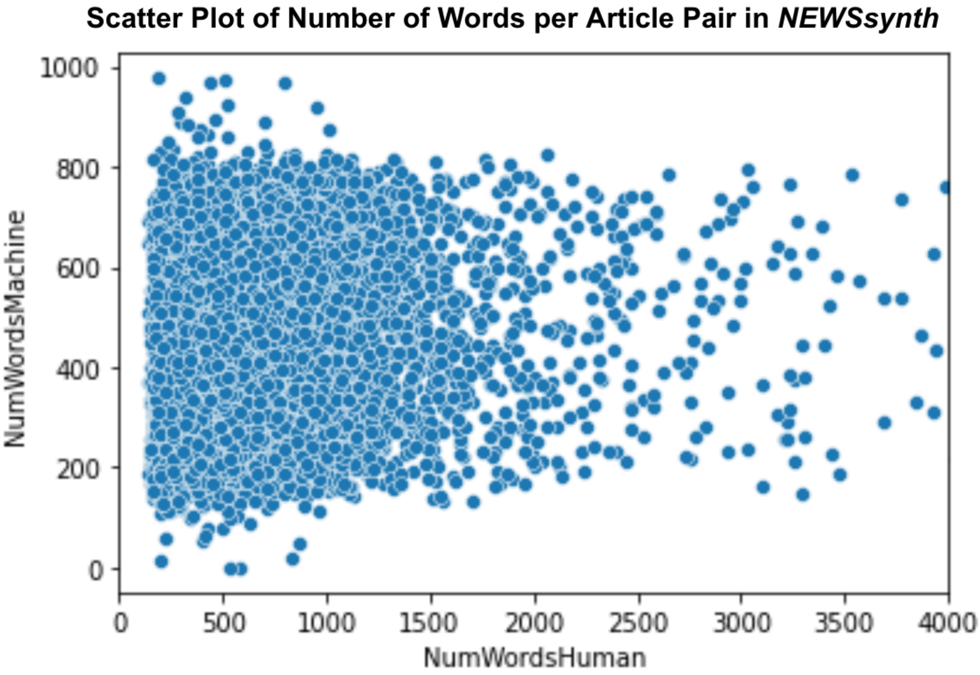


Figure 3.4: Scatter plot of number of words per article pair of synthetic text vs. human text in *NEWSsynth* (Pearson $r = 0.20$).

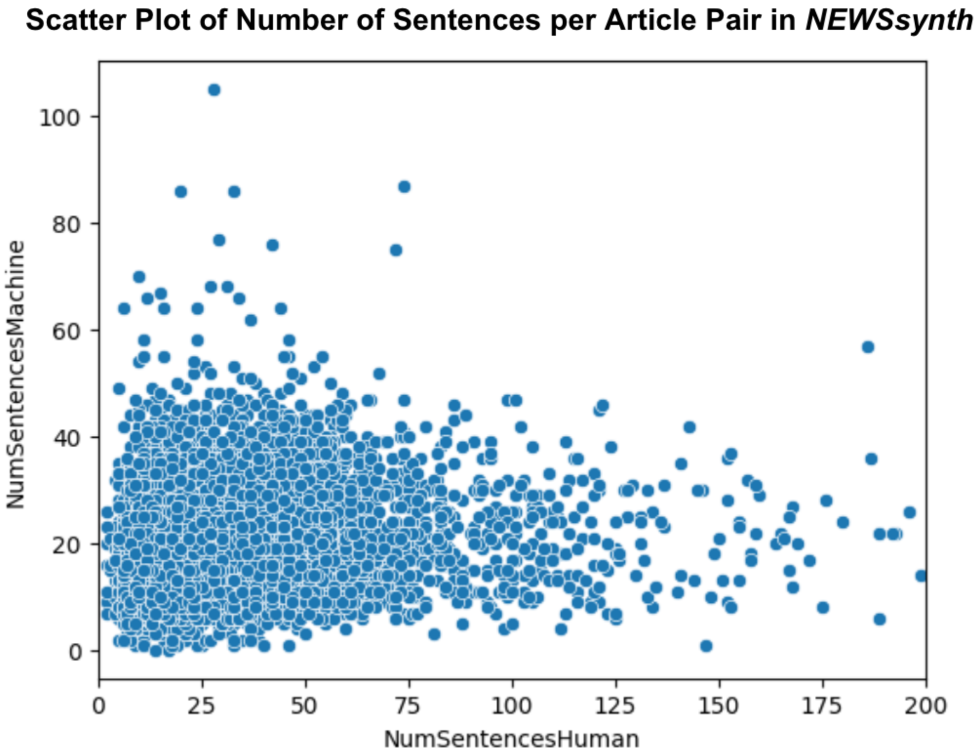


Figure 3.5: Scatter plot of number of sentences per article pair of synthetic text vs. human text in *NEWSsynth* (Pearson $r = 0.19$).

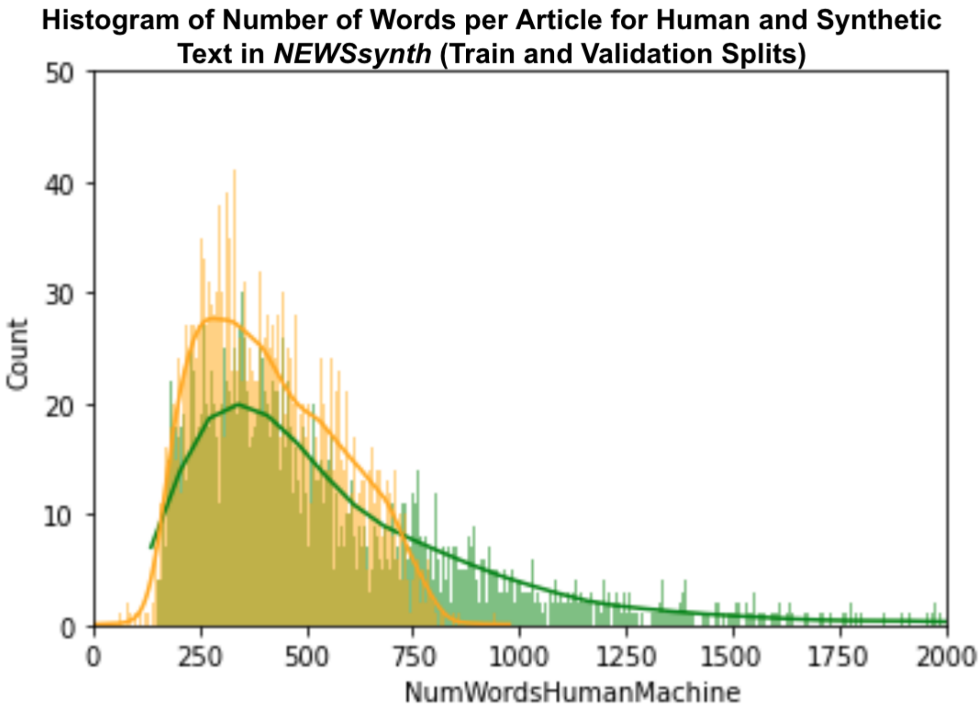


Figure 3.6: Histogram showing the number of words per article for human (green) and synthetic (orange) text in *NEWSsynth* (train and validation splits). Human articles tend to be longer ($\overline{x}$ = 595) than synthetic articles ($\overline{x}$=418), as shown by the 'right-shift' of the human (green) distribution compared to the synthetic (orange) distribution. Human articles have a much wider range of lengths (σ = 503) compared to synthetic articles (σ = 162), as shown by the longer tail of the human (green) distribution compared to the narrower synthetic (orange) distribution.

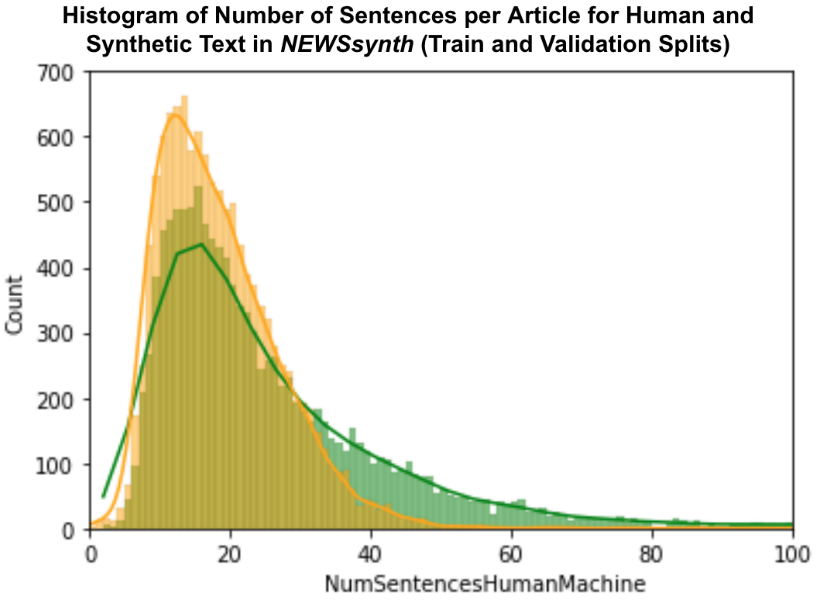


Figure 3.7: Histogram showing the number of sentences per article for human (green) and synthetic (orange) text in *NEWSsynth* (train and validation splits). Similar to Figure 3.6, human articles tend to be longer ($\overline{x}$ = 27) and have a much wider range of lengths (σ = 25) than synthetic articles ($\overline{x}$ = 18) and (σ = 9) respectively.

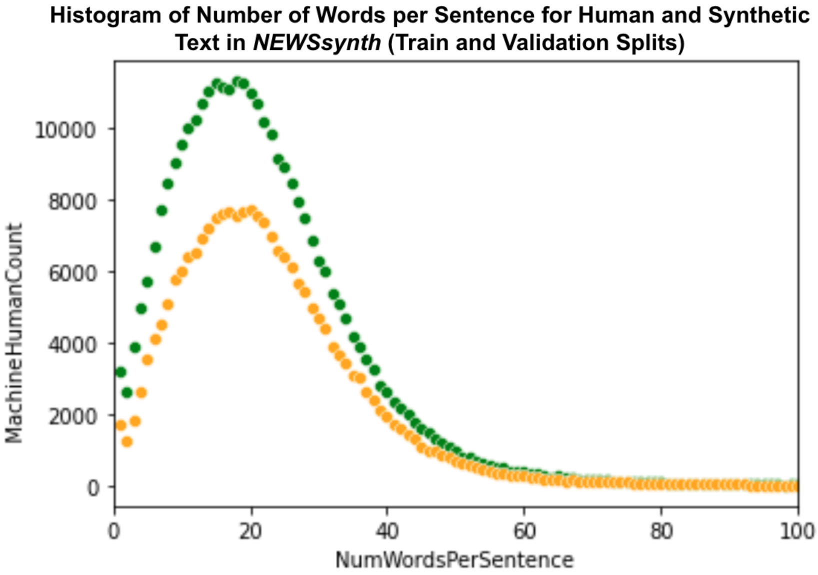


Figure 3.8: Histogram showing the number of words per sentence for human (green) and synthetic (orange) text in *NEWSsynth*. Human articles tend to have slightly fewer words per sentence ($\overline{x} = 22$) and a slightly narrower range of values ($\sigma = 16$) than synthetic articles ($\overline{x} = 23$) and ($\sigma = 17$) respectively. However, the human and synthetic distributions here are more similar than those in Figures 3.6 and 3.7.

suggest that recall is most sensitive to the size of the training set. This is perhaps because the PLM is already trained on human text during pretraining but not synthetic text (*exposure bias*), so more exposure to synthetic text increases the model's ability to detect synthetic text correctly with fewer false negatives.

3.4.3 Alternative Forms of `emoBERT`

We next examine the effect of using different emotion datasets to fine-tune our emotionally aware PLMs on the downstream task of synthetic text detection? We conduct experiments on `emoBERTsynth` by fine-tuning eight alternative `emoBERT` models:

- **GNE** involves fine-tuning using the GoodNewsEveryone dataset (§3.2.4) as in the main experiments;
- **GNE$_\mathrm{r}$** involves fine-tuning with a version of GNE with randomised labels. We do this to examine the extent to which the difference between `BERTsynth` and `emoBERTsynth` can be attributed to emotion or to the process of fine-

tuning on an arbitrary classification task with the GNE data;

- **AT** involves fine-tuning with the AffectiveText dataset comprising 1.5k news headlines in English annotated with respect to Ekman's 6 emotions (Strapparava and Mihalcea, 2008);
- **GA** is GNE and AT combined;
- **SST-2** involves fine-tuning on the task of sentiment polarity classification using the SST-2 dataset of 68,221 movie reviews in English (Socher et al., 2013);
- **GAS** is GNE, AT, and SST-2 combined; with SST-2 positive sentiment mapped to joy and negative sentiment mapped to sadness;
- **S-GA** involves first fine-tuning on sentiment using SST-2 and then fine-tuning on emotion using GA. This experiment is inspired by Kratzwald et al. (2018) who report that emotion classification can be improved by transfer learning from sentiment analysis;
- **GAS+-** is GAS but mapped to positive and negative sentiment.[6]

The results (Table 3.6) reveal that the best-performing `emoBERTsynth` models are those fine-tuned using GNE or using GNE and AffectiveText combined (GA). The latter achieves the highest accuracy and the former the highest F1. We attribute the relatively poor performance of AffectiveText on its own to its small size, comprising only 1.5k headlines (split 625 + 125 for training and dev splits respectively) compared to 5k for GNE and 68k for SST-2.

Fine-tuning on GNE outperforms fine-tuning with randomised labels ($\mathrm{GNE_r}$) by 1.1 percentage points. This drop in accuracy compared to GNE, albeit small, suggests that the emotion classification task does play a role in the improved performance of `emoBERTsynth` versus `BERTsynth`.

The results in Table 3.6 also suggest that fine-tuning on sentiment is not particularly helpful. The poor performance of the GAS model could be due

---

[6] Happiness was mapped to positive sentiment; sadness, fear, anger and disgust were mapped to negative sentiment; surprise was mapped to sentiment using a DistilBERT (base-uncased) (Sanh et al., 2020) sentiment classifier fine-tuned on the SST-2 dataset and available on HuggingFace. `https://huggingface.co/distilbert-base-uncased` 14.05% of 'surprise' mapped to positive, while the remaining 85.95% mapped to negative sentiment.

Dataset	Prec.	Rec.	F1	Acc.
GAS	81.95	85.58	83.72	83.36
S-GA	82.60	87.80	85.12	84.65
GAS+-	82.41	88.30	85.25	84.73
AT	**85.52**	83.88	84.69	84.84
SST-2	82.85	88.38	85.52	85.04
GNEr	82.44	89.93	86.02	85.39
GNE	83.84	**90.40**	**87.00**	86.49
GA	85.34	88.18	86.73	**86.51**

Table 3.6: Ablation experiments, using different emotion datasets for fine-tuning `emoBERT`, comparing `emoBERTsynth` (eBs) detectors on the task of synthetic text detection on the *NEWSsynth* dataset. GNE is the GoodNewsEveryone dataset which is used in the main experiments. GNE_r is GNE with randomised labels. AT is AffectiveText. GA is GNE and AT combined. SST-2 is the SST-2 sentiment dataset. GAS is the combined GNE, AT, and SST-2 datasets. S-GA is first fine-tuned on sentiment using the SST-2 dataset, and then fine-tuned on emotion using the GNE and AT datasets, and finally fine-tuned on synthetic text detection. GAS+- is GAS but mapped to positive and negative sentiment.

to the crude mapping of negative sentiment to sadness (because it could be any 1 of 5 Ekman emotions), which results in a large dataset imbalance across emotion labels. When we go in the opposite direction and map the emotion labels to sentiment labels (GAS+-), the results improve. Overall, however, the results suggest that mixing emotion and sentiment datasets is not a good idea (particularly if they are disproportionate in size and imbalanced), and that sentiment alone is not sufficient.

3.4.4 A Larger Detector Model

We next investigate what happens when we use a PLM larger than BERT to detect synthetic text. Using the same experimental setup described in §3.2, we substituted BLOOM (Scao et al., 2023) in place of BERT for the synthetic text detector. BLOOM is an open-science causal PLM alternative to GPT-3 (Brown et al., 2020). We use the BLOOM 560M size model. The results in Table 3.7 show that the emotionally-aware BLOOM PLM (`emoBLOOMsynth`) outperforms the standard BLOOM (`BLOOMsynth`) in all metrics.

	Prec.	Rec.	F1	Acc.
BLOOMsynth	81.90	85.95	83.79	83.40
Var.	(4.76)	(12.22)	(1.23)	(0.93)
emoBLOOMsynth	**85.98**	**88.02**	**86.90**	**86.75**
Var.	(5.72)	(9.96)	(0.27)	(0.15)
Δ	+4.08	+2.07	+3.11	+3.35

Table 3.7: Comparison of `BLOOMsynth` and `emoBLOOMsynth` against the *NEWSsynth* test set averaged over 5 runs (with variance in brackets). `emoBLOOMsynth` outperforms `BLOOMsynth` in Accuracy, F1, Recall, and Precision.

3.5 ChatGPT Experiments

All experiments so far have involved PLMs pre-trained with the self-supervised objective of predicting the next token or a masked token. We conduct a final experiment with ChatGPT, a more human-aligned Large Language Model (LLM) which has undergone a second training or "alignment" phase using Reinforcement Learning from Human Feedback on top of an underlying LLM (GPT 3.5 Turbo in our case) (OpenAI, 2022; Ouyang et al., 2022). We create a custom dataset comprising human articles and ChatGPT synthetic text from multiple non-news domains, and use it to compare our `BERTsynth` and `emoBERTsynth` models against ChatGPT (in a zero-shot setting) on the task of detecting ChatGPT's own synthetic text.

3.5.1 *ChatGPT100*

We create and release *ChatGPT100* - a dataset comprising human articles and synthetic articles generated by ChatGPT.[7] Following Clark et al. (2021) who collected 50 human articles and generated 50 articles using GPT2 and GPT3, we also collect 50 human articles, and we then use ChatGPT to generate 50 synthetic ones. The human written articles are from 5 different domains: Science, Entertainment, Sport, Business, and Philosophy. We used reputable websites for the human text which was gathered manually, see Table A.1 in Appendix A.

---

[7] We use ChatGPT (gpt-3.5-turbo-0314, Mar-14-2023 version) between dates 16-Mar-2023 and 24-Mar-2023.

The synthetic text was generated by providing ChatGPT with a prompt such as "*In less than 400 words, tell me about moral philosophy.*" where human text on the same topic, moral philosophy in this case, had already been found online. The data generated by ChatGPT is semantically correct and was checked manually. Subject areas in which we are knowledgeable were chosen so that the correctness of the synthetic text could be checked.

Table 3.8 contains two example articles from the *ChatGPT100* dataset - one human article retrieved from the internet, and one synthetic article generated by ChatGPT. The URL for the human article is shown, as is the prompt provided to ChatGPT to generate the synthetic article. Note that both articles are from the same domain, business, and both are on the subject of the stock market. Using human and synthetic article pairs on identical subjects and domains is intended to reduce any possible impact of subject knowledge (or lack of subject knowledge) on the detection task by ChatGPT.

3.5.2 Detection Task

To be comparable with the detectors presented in our earlier experiments, the articles were limited to a maximum of 384 words ($\approx$ 512 tokens) and truncated at a natural sentence boundary. The two articles were then made to be approximately the same length. Each article was appended to the following prompt to ChatGPT: "*Was the following written by a human or a computer, choose human or computer only, and briefly explain your choice?*" Having tested ChatGPT, we then tested our `BERTsynth` and `emoBERTsynth` models (the models fine-tuned on RealNews-Test from Table 3.3).

Table 3.9 shows the ChatGPT classification of the two examples provided in Table 3.8, including the prompt given to ChatGPT for the classification task. The ChatGPT predictions are comprehensive, plausible, and entirely incorrect. ChatGPT incorrectly predicts the human article is synthetic, and the synthetic article is human. We must caveat that ChatGPT, like other PLMs and LLMs,

Human Article	ChatGPT Article
URL: `https://www.investopedia.com/terms/s/stockmarket.asp`	Prompt: *In less than 400 words, tell me about the stock market.*
What Is the Stock Market? The term stock market refers to several exchanges in which shares of publicly held companies are bought and sold. Such financial activities are conducted through formal exchanges and via over-the-counter (OTC) marketplaces that operate under a defined set of regulations. Both “stock market” and “stock exchange” are often used interchangeably. Traders in the stock market buy or sell shares on one or more of the stock exchanges that are part of the overall stock market. The leading U.S. stock exchanges include the New York Stock Exchange (NYSE) and the Nasdaq. Understanding the Stock Market The stock market allows buyers and sellers of securities to meet, interact, and transact. The markets allow for price discovery for shares of corporations and serve as a barometer for the overall economy. Buyers and sellers are assured of a fair price, high degree of liquidity, and transparency as market participants compete in the open market. The first stock market was the London Stock Exchange which began in a coffeehouse, where traders met to exchange shares, in 1773. The first stock exchange in the United States began in Philadelphia in 1790. The Buttonwood Agreement, so named because it was signed under a buttonwood tree, marked the beginning of New York’s Wall Street in 1792. The agreement was signed by 24 traders and was the first American organization of its kind to trade in securities. The traders renamed their venture the New York Stock and Exchange Board in 1817. A stock market is a regulated and controlled environment. In the United States, the main regulators include the Securities and Exchange Commission (SEC) and the Financial Industry Regulatory Authority (FINRA). The earliest stock markets issued and dealt in paper-based physical share certificates. Today, stock markets operate electronically. How the Stock Market Works Stock markets provide a secure and regulated environment where market participants can transact in shares and other eligible financial instruments with confidence, with zero to low operational risk.	The stock market is a public marketplace where individuals and institutions can buy and sell shares in publicly traded companies. Shares represent ownership in a company, and stock prices fluctuate based on supply and demand, as well as factors such as the company’s financial performance, industry trends, and economic conditions. When investors purchase stock, they are essentially buying a piece of the company and are entitled to a portion of its profits. This can come in the form of dividends, which are payments made to shareholders based on the company’s earnings. Investors can also profit by selling their shares at a higher price than they paid for them. The stock market plays a critical role in the global economy. It provides companies with a way to raise capital by selling shares to the public, which they can then use to invest in their business operations or pay down debt. It also provides individuals and institutions with a way to invest in a diverse range of companies and industries, potentially earning significant returns on their investment. However, investing in the stock market comes with risks. Stock prices can be volatile, and investors can experience significant losses if they make poor investment decisions or the market experiences a downturn. It is important for investors to do their research and carefully consider the risks before investing in any stock or fund. There are a variety of ways to invest in the stock market, ranging from purchasing individual stocks to investing in mutual funds or exchange-traded funds (ETFs). Mutual funds and ETFs allow investors to diversify their holdings across a range of companies and industries, reducing the risk of a single stock or sector impacting their portfolio. Overall, the stock market plays a critical role in the global economy and provides individuals and institutions with a way to invest in the growth and success of publicly traded companies.

Table 3.8: Human and ChatGPT articles from the *ChatGPT100* dataset. The Human article was retrieved on 20-Mar-2023 from the URL shown. The ChatGPT article was generated (using gpt-3.5-turbo-0314) on the same day using the prompt shown.

Human Article	ChatGPT Article
Prompt: *Was the following written by a human or a computer, choose human or computer only, and briefly explain your choice?*	
What Is the Stock Market? The term stock market refers to several exchanges in which shares of publicly held companies are bought and sold. Such financial activities are conducted through formal exchanges and via over-the-counter (OTC) marketplaces that operate under a defined set of regulations. [Rest of article omitted for display purposes]	The stock market is a public marketplace where individuals and institutions can buy and sell shares in publicly traded companies. Shares represent ownership in a company, and stock prices fluctuate based on supply and demand, as well as factors such as the company's financial performance, industry trends, and economic conditions. [Rest of article omitted for display purposes]
ChatGPT Prediction	
The passage appears to be written in a factual, informative style with a focus on providing information about the stock market. The writing is clear and concise, without any obvious errors or omissions. Additionally, there is a lack of personal opinions or perspectives, which suggests that it was likely generated by an AI language model. The absence of any human-like idiosyncrasies or expressions further supports this assessment.	The passage is written in a clear, informative style, but it also has a certain level of nuance and expression that suggests it was written by a human rather than a computer. The writing includes a few examples and perspectives that are not strictly factual, such as the potential benefits and risks of investing in the stock market, which suggests a level of editorial judgment and experience that is more typical of a human author. Additionally, the language is slightly more varied and expressive than what might be expected from an AI language model, which tends to produce more formulaic and standardized writing. Overall, while the passage is focused on delivering information, it also has a human touch that makes it more engaging and relatable to readers.

Table 3.9: ChatGPT (gpt-3.5-turbo-0314) detecting Human and ChatGPT articles from *ChatGPT100*. The aricles were previously shown in Table 3.8 and are partially reproduced here. The classification was conducted on 21-Mar-2023. ChatGPT incorrectly predicts that the Human article is Synthetic, and that the Synthetic article is Human.

Model	Prec.	Rec.	F1	Acc.
ChatGPT	**75.00**	30.00	42.86	60.00
BERTsynth	60.24	**100.00**	75.19	67.00
emoBERTsynth	67.57	**100.00**	**80.65**	**76.00**

Table 3.10: Our emotionally aware PLM (`emoBERTsynth`) outperforms ChatGPT and `BERTsynth` at detecting synthetic text in the *ChatGPT100* dataset. Note that ChatGPT is performing the task zero-shot.

is prone to hallucinating during generation - this includes not just generation of articles, but also the generation of explanation for predictions.

3.5.3 Results

The results are shown in Table 3.10. The first thing to note is that no model performs particularly well. ChatGPT tends to misclassify its own synthetic text as human (hence the low recall score of 30%).[8] `BERTsynth` and `emoBERTsynth`, on the other hand, tend to classify text as machine-written and they both obtain 100% recall. We previously saw (§3.4.2) that recall is most sensitive to fine-tuning set size. The `emoBERTsynth` and `emoBERTsynth` models have been exposed to synthetic text during fine-tuning, whereas ChatGPT is performing the task zero-shot. This could explain some of the difference in recall between the ChatGPT and the two fine-tuned models.

Finally, as with our experiments with Grover-generated text, `emoBERTsynth` outperforms `BERTsynth` on all metrics. The dataset is small so we must be careful not to conclude too much from this result, but it does suggest that fine-tuning on emotion could be beneficial when detecting synthetic text from LLMs and more sophisticated generators, in non-news domains. This is in line with the results of our earlier experiments using variously size PLMs (such as Grover, BERT, BLOOM), used as generators and detectors in the news domain, and shows the potential for our approach with different generator models and in different domains.

---

[8] ChatGPTs responses suggest it may use fact-checking as a proxy during synthetic text detection.

3.6 Conclusion

We conducted experiments investigating the role that emotion recognition can play in the detection of synthetic text. An emotionally-aware PLM fine-tuned on emotion classification and subsequently trained on synthetic text detection (`emoPLMsynth`) outperformed a model with identical fine-tuning on synthetic text detection, but without emotion training, (`PLMsynth`). The results hold across different synthetic text generators, model sizes, datasets and domains. This work specifically demonstrates the benefits of considering emotion in the task of detecting synthetic text, it contributes two new datasets (*NEWSsynth* and *ChatGPT100*) and, more generally, it hints at the potential benefits of considering human factors in NLP and Machine Learning.

Is it possible that some other proxy for synthetic text is at play? We ruled out some potential proxies related to article length in §3.4.1. In ablation studies in §3.4.3, we showed that the emotion labels result in an improvement in performance compared to randomised labels for the same emotion dataset. Other potential proxies are nonsensical sentences, repetitive text, etc. However, we account for these by comparing our emotionally-aware PLMs (`emoPLMsynth`) against standard PLMs fine-tuned on synthetic text detection only (`PLMsynth`). Thus, any advantage or disadvantage of sentences without meaning (or any other factor) is also available to the non-emotionally-aware model against which we compare our emotionally-aware model.

Future work includes repeating the experiments using other PLMs for detection such as RoBERTa (Liu et al., 2019) which we use in RQ3 and RQ4, along with the latest version of the LLMs we used in these experiments i.e. BLOOM (Scao et al., 2023) and GPT 3.5 (OpenAI, 2022) and the latest open source LLMs like LLaMa (Touvron et al., 2023b), LLaMA 2 (Touvron et al., 2023a), Vicuna (Chiang et al., 2023), Falcon (Penedo et al., 2023), and Mistral (Jiang et al., 2023). Additional gains may be found by hyperparameter tuning, including prompt engineering (Liu et al., 2023), choosing a loss function that improves Recall over other metrics, and training for more epochs.

Later, in Chapter 6, we answer RQ4 where we investigate the *affective profile* (i.e. emotional content and characteristics) of human and synthetic text in an attempt to determine if there are measurable differences which may prove useful in the task of synthetic text detection. Before we get there, we must first turn our attention to RQ2 and RQ3, where we develop an emotion classification evaluation framework and a set of emotion classifiers respectively; both of which we use to address RQ4.

Chapter 4

ψ - a Novel Evaluation Framework (for Dimensional Emotion Classifiers)

In this chapter we answer Research Question 2: "*Is it possible to develop a framework for measuring and evaluating classifiers of Plutchik's 8 Emotions which takes account of their dimensional relationship?*". RQ2 was described in §1.2, with Related Work discussed in §2.4. It begins with an Overview (§4.1) and a definition of "A Novel Evaluation Framework for Emotion Classification" (§4.2) followed by a "Toy Example" (§4.3). We discuss "Extending $\phi(\delta)$ - a Flexible Loss Function" (§4.4) and finish with our Conclusions (§4.5).

We introduce ψ - a flexible evaluation framework which we use, in the first instance, to evaluate the performance of dimensional emotion classifiers in Chapter 5. In §6.4.3 we will see how ψ can be used in other scenarios demonstrating its flexibility and extensibility as an evaluation framework. ψ takes the important parts, such as exactitude, of error-based metrics like MAE and RMSE and combines them with the ease of interpretability of class based metrics like F1μ, Accuracy, Precision, Recall which are typically used in emotion classification. We saw in §2.4 that none of the popular metrics discussed are ideally suited to our needs of classifying emotions with a dimensional relationship. We also saw that we need a way to elevate Plutchik's Wheel of Emotions from an `Ordinal` scale to an `Interval` scale.

4.1 Overview

Elevating Plutchik's Wheel of Emotion to the `Interval` scale is achieved by defining a **distance between emotions**, which we refer to as δ and describe in detail in §4.2.1. The key point is that we assign a numeric value (0,1,2,3 etc.) as the difference between one emotion (e.g. gold label) and another emotion (e.g. predicted label). If we were to instead simply assign a numeric value to each emotion it would be arbitrary and without meaning.

We use δ, the distance between emotions, to define a numeric loss between a gold label and a predicted label. We refer to the loss function as $\phi(\delta)$ which we describe along with sample calculations in §4.2.2. We demonstrate the flexibility and extensibility of the ψ metric by using δ to define several custom loss functions $\phi(\delta)$ suited to specific tasks as defined by the researcher or user (§4.4).

Finally, we combine the loss for each example into a single percentage score, which we refer to as the ψ score. The ψ score is in the range 0..100 because it is easy to interpret and more familiar to the emotion classification research community. The ψ score is described in §4.2.3.

In the rest of this chapter we describe our metric, ψ, and demonstrate its implementation and calculation using a toy example.[1] We contrast ψ with F-score, MAE, and RMSE. For F-score we specifically use F1μ which is frequently used in emotion classification (Bostan et al., 2020; Bostan and Klinger, 2018), and note again that for single-label multi-class classification (as in our case) the Accuracy, Precision, Recall, and F1 have equal values.

---

[1] Note that we use ψ in 'real world' scenarios to evaluate the performance of emotion classifiers in Chapter 5 and to measure affective profiles in Chapter 6.

4.2 A Novel Evaluation Framework for Emotion Classification

4.2.1 Defining δ - a Measure of Emotion Similarity

We define δ as the distance between two emotions. Thus, δ is a measure of the similarity between emotions. The distance, δ, between Plutchik's 8 emotions is illustrated in Figure 4.1 and shown in Table 4.1. For example, Trust is adjacent to Joy - thus Trust and Joy have $\delta = 1$; Anticipation is also adjacent to Joy so they also have $\delta = 1$; while Trust and Anticipation have a $\delta = 2$ because they are two emotions apart. Opposing emotions e.g. Joy and Sadness have $\delta = 4$ which is the maximum distance for two emotions in Plutchik's 8 emotions. We define δ_{max} as the maximum distance possible between two emotions, and we will see δ_{max} is a constant for a given set of emotions ($\delta_{max} = 4$ for Plutchik's 8 emotions). Note that each emotion has $\delta = 0$ with itself, as shown by the diagonal line of 0's in Table 4.1.

	Joy	Sad	Fea	Ang	Dis	Sur	Tru	Ant
Joy	0	4	2	2	3	3	1	1
Sadness	4	0	2	2	1	1	3	3
Fear	2	2	0	4	3	1	1	3
Anger	2	2	4	0	1	3	3	1
Disgust	3	1	3	1	0	2	4	2
Surprise	3	1	1	3	2	0	2	4
Trust	1	3	1	3	4	2	0	2
Anticipation	1	3	3	1	2	4	2	0

Table 4.1: Distance, δ, between Plutchik's Emotions. The closer two emotions are (according to Plutchik's Wheel) the lower their δ value. Note that, for Plutchik's 8 emotions, $\delta_{max} = 4$.

According to Stevens (1946) theory of scales of measurement, by introducing the δ to Plutchik's Wheel we elevate it from an "*ordinal*" scale to an "*interval*" scale. This is important because it allows us to measure differences between emotions, and therefore to calculate relationships between the emotions including central tendencies and spread. We can use the δ between emotions as a more

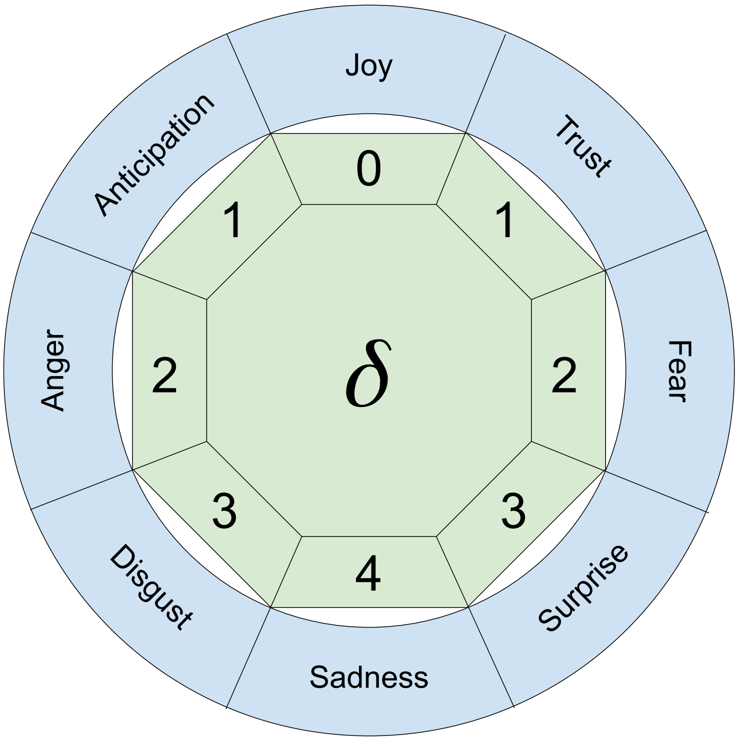


Figure 4.1: Plutchik Wheel of Emotion (in blue) showing, as an example, the δ (in the green octagon) between 'Joy' and the other emotions. Rotating Plutchik's wheel of emotions so that a given emotion aligns with 0 on the δ-octagon will give the δ between that emotion and the other Plutchik emotions.

fine-grained way to measure the performance of an emotion classifier. Given a gold label of 'Joy' and a predicted label of 'Trust': metrics likes Accuracy simply record an incorrect prediction. However, we can use the $\delta = 1$ to take account of the fact that, according to Plutchik's model, 'Trust' is incorrect but is a better prediction than say 'Sadness' with $\delta = 4$. Next we define a way to turn the δ value between two emotions into a Loss - using a Loss function.

4.2.2 Defining $\phi(\delta)$ - a Flexible Loss Function

We define $\phi(\delta)$ as a loss function[2] which takes a δ value and converts it into a numerical loss in the range 0..1. A loss of 0 indicates a correct prediction, a loss of 1 indicates a prediction that is maximally incorrect i.e. the opposite emotion according to Plutchik's Wheel. Exploiting the δ between emotions allows us to

---

[2] For now, we consider the specific loss function used in this work, later we describe the flexibility and extensibility of $\phi(\delta)$ in relation to other loss functions (§4.4).

δ	Score	$\phi(\delta) = \frac{\delta}{\delta_{max}}$ (Loss)
0	1	0/4
1	0.75	1/4
2	0.50	2/4
3	0.25	3/4
4	0	4/4

Table 4.2: The score awarded to a prediction is determined by δ i.e. how close the predicted emotion is to the correct emotion on Plutchik's Wheel. We use a linear loss which is calculated as the quotient of the distance between the two emotions and the maximum possible distance between two emotions i.e. $\phi(\delta) = \frac{\delta}{\delta_{max}}$

assign a partial score (effectively a Loss) which is proportional to the δ between two emotions, as shown in Table 4.2. For example, if the correct emotion is Joy and the predicted emotion is also Joy (δ=0) we assign this correct prediction a full score i.e. 1 with a Loss of 0. However, if the correct emotion is Joy but the predicted emotion is Trust or Anticipation (δ=1) we can assign this incorrect prediction a score close to a full score i.e. 0.75 with a Loss of 0.25; we reduce this score for $\delta = 2$ (0.50, a Loss of 0.50), reduce it further for $\delta = 3$ (0.25, a Loss of 0.75), and reduce it to 0 (a Loss of 1) if $\delta = 4$ (which is δ_{max} for 8 emotions) as it "couldn't be more wrong". Table 4.2 shows the δ, Score and Loss according to the Loss function i.e $\phi(\delta) = \frac{\delta}{\delta_{max}}$.

Equipped with an understanding of how we use δ to measure the distance between emotions, and use $\phi(\delta)$ as a loss function to convert a δ value into a loss, we now explain the process of taking the loss for each of these individual measurements - and combining them to give a ψ score in the range 0 - 100%.

4.2.3 Defining ψ - a Novel Metric for Emotion Classification Performance

The losses for all the individual predictions are combined to yield the overall loss. The overall loss is deducted from the total number of examples (for example, there are 10,000 examples in the NEWSsynth training split), the result of which is converted to a percentage value which is the ψ score.

For each δ value, we use the term n_δ to refer to the number of examples which have that specific δ. The product of the loss ($\frac{\delta}{\delta_{max}}$), and the number of examples with that loss (n_δ) gives us the total loss for that specific delta ($n_\delta \frac{\delta}{\delta_{max}}$). We sum over the losses for each specific δ in the range $\delta = 0$ to $\delta = \delta_{max}$ to give a total loss over all δ values for all the examples ($\sum_{\delta=0}^{\delta_{max}} n_\delta \frac{\delta}{\delta_{max}}$). The total loss is deducted from N, the total number of examples, where N essentially represents a perfect score i.e. a score of 1 per example ($N - \sum_{\delta=0}^{\delta_{max}} n_\delta \frac{\delta}{\delta_{max}}$). Finally, to obtain a percentage, we divide by the total number of examples (N) and multiply by 100 ($\frac{100}{N}$). This yields the final equation for our ψ metric, as shown in Equation 4.1.[3]

$$\psi = \frac{100}{N}\left(N - \sum_{\delta=0}^{\delta_{max}} n_\delta \frac{\delta}{\delta_{max}}\right) \tag{4.1}$$

Equation 4.1 defines the ψ metric, where:

N	Total number of examples being tested
δ	Distance between two emotions (0,1,2,3,4) per Table 4.1.
δ_{max}	Maximum distance between two emotions[4] i.e. $\delta_{max} = \lfloor \frac{n_{emo}}{2} \rfloor$
n_{emo}	Total number of emotions (8 for Plutchik)
n_δ	Number of occurrences of a particular δ value (0..N)

We use a percentage value for ψ because it is relatively easy to interpret (compared to MAE and RMSE for example), and readily comparable with F1, Accuracy, Precision and Recall (which are the most commonly used metrics in emotion classification). A ψ score of 100 is a perfect score where all predictions are correct ($\delta = 0$), while a ψ score of 0 is the worst possible score where all predictions are maximally incorrect ($\delta = \delta_{max}$). Later we will describe how Accuracy is essentially a specific instance of the ψ score.

Next we demonstrate the ψ metric using a toy example to show how it operates in practice, and compare results for the ψ score against Accuracy, MAE, and RMSE.

---

[3] Later we will define a more generalised version of this equation by replacing $\frac{\delta}{\delta_{max}}$ with a function which can be user defined for a given use-case.

[4] For Plutchik's 8 emotions: $\delta_{max} = \lfloor \frac{n_{emo}}{2} \rfloor = \lfloor \frac{8}{2} \rfloor = 4$

4.3 Toy Example

4.3.1 Demonstrating ψ

Table 4.3 is a toy example[5] demonstrating ψ in comparison to Accuracy (F1, Precision and Recall), MAE and RMSE for a classifier of Plutchik's 8 emotions. Two columns show the True (gold) Label and the Predicted Label (from a hypothetical classifier). The rightmost 4 columns show the Loss for each metric, along with the score for each metric calculated in the bottom three rows. The following paragraphs describe those calculations in detail.

Ex.	True Label	Predicted Label	Correct	Error (Loss) δ	δ^2	$\frac{\delta}{\delta_{max}}$
1	Joy	Joy	✓	0	0	0
2	Joy	Disgust	×	3	9	0.75
3	Sadness	Anger	×	2	4	0.50
4	Fear	Fear	✓	0	0	0
5	Fear	Anger	×	4	16	1
6	Anger	Anticipation	×	1	1	0.25
7	Disgust	Disgust	✓	0	0	0
8	Disgust	Joy	×	3	9	0.75
9	Surprise	Trust	×	2	4	0.50
10	Trust	Surprise	×	2	4	0.50
11	Trust	Trust	✓	0	0	0
12	Anticipation	Joy	×	1	1	0.25
		Total Error (Loss)	8	18	48	4.50
		Score	**33.33%**	**1.50**	**2.00**	**62.50%**
		Metric	**Acc.**	**MAE**	**RMSE**	$\boldsymbol{\psi}$

Table 4.3: Toy example illustrating the difference between our novel ψ metric and existing metrics including Accuracy, MAE, and RMSE. Note that MAE is the mean δ value, RMSE is the square root of the mean δ^2 value. Accuracy is the quotient of the number of correct predictions and the total number of predictions, N. ψ is calculated per Equation 4.1. Most notable is that ψ is higher than Accuracy (62.50% vs 33.33%) - showing that ψ rewards predictions which may be "incorrect" but close to the correct emotion. This is evident in the $\frac{\delta}{\delta_{max}}$ column - which is effectively the loss of the prediction - because the losses are not all 1 when incorrect, they also include losses of only 0.75, 0.5, and 0.25.

---

[5] Chapter 5 provides real-world results from our experiments with dimensional emotion classifiers using ψ, F1_{μ}, MAE and RMSE.

Accuracy is defined in Equation 4.2,[6] and calculated using the sample values in Table 4.3. Given the true value y_i and the predicted value $\hat{y}_i$ of the i-th example, the percentage of correct predictions over all examples (N) is defined as:

$$\begin{aligned}\text{Accuracy}(y, \hat{y}) &= \frac{100}{N} \sum_{i=0}^{N-1} 1(\hat{y}_i = y_i) \\ &= \frac{100}{12}\, 4 \\ &= 33.33\%\end{aligned} \tag{4.2}$$

MAE is defined in Equation 4.3,[7] and calculated using the sample values in Table 4.3. Given the true value y_i and the predicted value $\hat{y}_i$ of the i-th example, the MAE over all examples (N) is defined as:

$$\begin{aligned}\text{MAE}(y, \hat{y}) &= \frac{1}{N} \sum_{i=0}^{N-1} |y_i - \hat{y}_i| \\ &= \frac{1}{N} \sum_{i=0}^{N-1} \delta_i \qquad \text{where } \delta_i = |y_i - \hat{y}_i| \\ &= \frac{1}{12}\,(0+3+2+0+4+1+0+3+2+2+0+1) \\ &= \frac{1}{12}\, 18 \\ &= 1.50\end{aligned} \tag{4.3}$$

We substitute δ for $|y_i - \hat{y}_i|$ using the δ value specified in Table 4.1 for each pair of actual emotion (y_i) and predicted emotion ($\hat{y}_i$).[8]

RMSE is defined in Equation 4.4,[9] and calculated using the sample values in Table 4.3. Given the true value y_i and the predicted value $\hat{y}_i$ of the i-th

---

[6] `https://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score`

[7] `https://scikit-learn.org/stable/modules/model_evaluation.html#mean-absolute-error`

[8] δ values are positive so there is no need to calculate the absolute part of $|y_i - \hat{y}_i|$.

[9] `https://scikit-learn.org/stable/modules/model_evaluation.html#mean-squared-error`

example, the RMSE over all examples (N) is defined as:

$$\begin{aligned}
\mathrm{RMSE}(y,\hat{y}) &= \sqrt{\frac{1}{N}\sum_{i=0}^{N-1}(y_i-\hat{y}_i)^2} \\
&= \sqrt{\frac{1}{N}\sum_{i=0}^{N-1}\delta_i^2} \qquad \text{where } \delta_i^2=(y_i-\hat{y}_i)^2 \\
&= \sqrt{\frac{1}{12}\left(0+9+4+0+16+1+0+9+4+4+0+1\right)} \\
&= \sqrt{\frac{1}{12}\,48} \\
&= \sqrt{4} \\
&= 2.00
\end{aligned} \tag{4.4}$$

ψ is defined in Equation 4.5,[10] and calculated using the sample values in Table 4.3. Given the true value y_i and the predicted value $\hat{y}_i$ of the i-th example

---

[10] To be consistent with the format of the equations for Accuracy, MAE, and RMSE - we rewrite Equation 4.1 (which sums over δ) to sum over i when calculating ψ in this example.

where there are n_{emo} total emotions, the ψ over all examples (N) is defined as:

$$\begin{aligned}
\psi_{(y,\hat{y})} &= \frac{100}{N}\left(N - \sum_{i=0}^{N-1} \frac{|y_i - \hat{y}_i|}{\lfloor \frac{n_{emo}}{2} \rfloor}\right) \\
&= \frac{100}{N}\left(N - \sum_{i=0}^{N-1} \frac{\delta_i}{\delta_{max}}\right) \qquad \text{where } \delta_i = |y_i - \hat{y}_i|, \text{ and } \delta_{max} = \lfloor \frac{n_{emo}}{2} \rfloor \\
&= \frac{100}{12}(12 - \sum_{i=0}^{N-1} \frac{\delta_i}{\delta_{max}}) \\
&= \frac{100}{12}(12 - (0 + 0.75 + 0.50 + 0 + 1 + 0.25 + 0 + 0.75 + 0.50 + 0.50 + 0 + 0.25)) \\
&= \frac{100}{12}(12 - 4.50) \\
&= \frac{100}{12}(7.50) \\
&= \frac{750}{12} \\
&= 62.50\%
\end{aligned} \tag{4.5}$$

4.3.2 Comparing Metrics

We compare our ψ metric to popular existing metrics for emotion classification, covered by Accuracy, and also MAE and RMSE. Accuracy is 33.33% because there are four correct predictions from twelve examples (Ex.1 to Ex.12). F1, Precision and Recall will also have scores of 33.33% since this is single-label multi-class classification. Note the binary nature of Accuracy - either a prediction is correct or incorrect. Accuracy does not take account of how close a prediction is to the correct answer. To illustrate this deficit, Table 4.3 Example 5 shows an incorrect prediction of Anger where the correct label is Fear; and Example 6 shows an incorrect prediction of Anticipation where the correct label is Anger. While Accuracy is right to consider these predictions incorrect, it fails to consider the dimensional relationship between the emotions as defined by Plutchik. By considering the dimensional relationship between emotions we see a different picture emerge. Table 4.1 shows the incorrect prediction in Example 5 is incorrect by the maximum amount ($\delta = 4 = \delta_{max}$), whereas Example 6 is incorrect by the minimum amount ($\delta = 1$). Accuracy fails to consider the dimensional relationship between the emotions and treats both inaccuracies the same even though they are not the same - as we have demonstrated.

MAE is 1.50 for our hypothetical classifier.[11] But how should we interpret 1.50? Is it indicative of a good, bad, or indifferent classifier? We can say that, as we have defined MAE in Equation 4.3, a MAE of 1.50 tells us that the mean δ for a prediction is 1.50 i.e. on average a predicted emotion is 1.5 emotions 'away' from the correct emotion. We can visualise this distance (δ) by referring to Figure 4.1. While this is useful for interpretation it is, nonetheless, somewhat meaningless without knowing how many emotions are being predicted and the resulting maximum error (δ_{max}) the system *could* potentially have. If there are 6 emotions like Ekman[12] then $\delta_{max} = 3$ so a MAE of 1.50 suggests a mediocre to poor classifier. If there are 8 emotions like Plutchik then $\delta_{max} = 4$ so an MAE of 1.50 might suggest a mediocre to good classifier. However, if there are 15 emotions (Bostan et al., 2020) then $\delta_{max} = 7$ so a MAE of 1.50 indicates a very strong classifier. Thus, it is difficult to interpret 1.50 in a way which allows us to compare it to other classifiers in a meaningful way - unless we know their δ_{max} and do some arithmetic like $\frac{\delta}{\delta_{max}}$. In general, it is difficult to interpret MAE because it can range from 0 to ∞ depending on the maximum size of errors (or Loss). This highlights another issue with MAE - it is difficult to compare across different numbers of classes - emotions in our case. Furthermore, the emotion classification research community is accustomed to using metrics with a range from 0 to 100 (Accuracy, F1, Precision and Recall) so using a metric with varying ranges, determined by the number of classes, introduces an additional

---

[11] We use δ as the absolute error when calculating MAE.

[12] Ekman treats emotions as distinct classes with no dimensional relationship between them. However, we could use the dimensions specified by Plutchik (which includes Ekman's 6 emotions) and the δ values shown in Table 4.1 to treat Ekman's 6 emotions dimensionally.

burden and possibility for misinterpretation.

RMSE is 2.00 for our hypothetical classifier.[13] All of the issues with MAE apply to RMSE so we will not restate them. As with MAE we take the error to be δ, thus for RMSE the squared-error is δ^2. A note of caution when using RMSE is that, by design, its effect is different depending on the magnitude of a number. Specifically, in our case for $2 \leq \delta \leq 4$ the effect is greater as δ increases. Whereas for $\delta = [0, 1]$ squaring has no effect. This can be seen by comparing the δ and δ^2 columns in Table 4.3. Squaring is a perfectly valid operation where, for example, you want to increase the loss or more severely punish increasingly incorrect predictions. RMSE, in contrast to Mean Squared Error (MSE), returns the square-root. This has the effect that, although skewed, the units returned by RMSE are the same as that of the original error.

ψ is 62.50% for our hypothetical classifier.[14] The calculation is shown in Equation 4.5. A ψ score of 62.50% (which uses a linear loss function) shows that our classifier is almost two-thirds correct in its predictions i.e. overall the predictions are approximately within one-third of the correct emotion and two-thirds away from the worst possible prediction. While there is a bit of learning involved with every new metric, the ψ score is more meaningful and easier to interpret than the error-based metrics, MAE and RMSE. The ψ score gives a more representative evaluation than simpler binary class-based methods like Accuracy, because ψ considers the similarity between the actual and predicted emotion and rewards closer predictions. Furthermore, ψ score is comparable across classifiers with different numbers of emotions - unlike MAE and RMSE.

For any given ψ score there are multiple ways to achieve that score, this is also the case with MAE and RMSE. From our toy example, if we consider a ψ score of 50% this could be achieved with half (6) of the examples having $\delta = 0$ and the other half (6) of the examples having $\delta = 4$. But a ψ score of 50% could also be achieved if all of the examples (12) had $\delta = 2$ and so on. The ψ

---

[13] We use δ^2 as the squared error when calculating RMSE.

[14] We use $\frac{\delta}{\delta_{max}}$ as the error when calculating ψ.

n_δ	Count
n_0	4
n_1	2
n_2	3
n_3	2
n_4	1

Table 4.4: Distribution of n_δ counts from the toy example in Table 4.3 as defined by ψ in Equation 4.1. The table shows that 4 examples had $\delta = 0$ (n_0), 2 examples had $\delta = 1$ (n_1) etc. This distribution of δ counts provides additional insight into the performance of the classifier.

metric conveniently allows us to know how the ψ score was achieved by viewing the distribution of n_δ counts. This is shown in Table 4.4 using the counts from our toy example. The distribution can be used by users so they can make more informed decisions affecting their classification model performance as defined by their specific requirements. The distribution could also be used as features in classical machine learning models etc. In §6.4.3 we will use the distribution of n_δ to measure the affective coherence of human and synthetic news articles using ψ.

One feature that is not evident from our toy example is the stronger correlation between ψ and MAE and RMSE, compared to Accuracy (and therefore Precision, Recall, and F-score). We consider this a benefit because it indicates that ψ reflects the fine-grained measurement of MAE and RMSE compared to the coarser binary 'correct - incorrect' measurement of Accuracy. We will identify and measure the correlations in detail in later experiments (§5.4.2).

4.4 Extending $\phi(\delta)$ - a Flexible Loss Function

We now consider several loss functions $\phi(\delta)$ which can be used with our ψ metric, and tailored to user requirements. The loss functions map from a given δ value to a resulting loss, using the specified loss function $\phi(\delta)$. For the following examples we will adhere to our current use case of δ values in the range 0..4 (for Plutchik's 8 emotions), and a loss function $\phi(\delta)$ in the range 0..1.

We can now rewrite Equation 4.1 to include $\phi(\delta)$, rather than our specific use-case where $\phi(\delta) = \frac{\delta}{\delta_{max}}$, as shown in Equation 4.6:

$$\psi = \frac{100}{N}\left(N - \sum_{\delta=0}^{\delta_{max}} n_\delta \quad \phi(\delta)\right) \tag{4.6}$$

Linear Loss Function In our toy example we used $\frac{\delta}{\delta_{max}}$ as our loss function and noted that it was linear. It is linear because it conforms to the general equation of a line as shown in Equation 4.7.

Linear Loss function:

$$\begin{aligned} f(x) &= mx + c \\ \phi(\delta) &= m.\delta + c \qquad \text{where } m = \frac{1}{\delta_{max}} \text{ and } c = 0 \\ \phi(\delta) &= \frac{\delta}{\delta_{max}} \qquad \text{where } \delta_{max} = \lfloor \frac{n_{emo}}{2} \rfloor = \frac{8}{2} = 4 \\ \phi(\delta) &= \frac{\delta}{4} \end{aligned} \tag{4.7}$$

We plot this linear equation in Figure 4.2 and see that the points for $\phi(\delta)$ vs δ match those in Table 4.2 i.e. (0, 0), (1, 0.25), (2, 0.5), (3, 0.75), (4, 1).

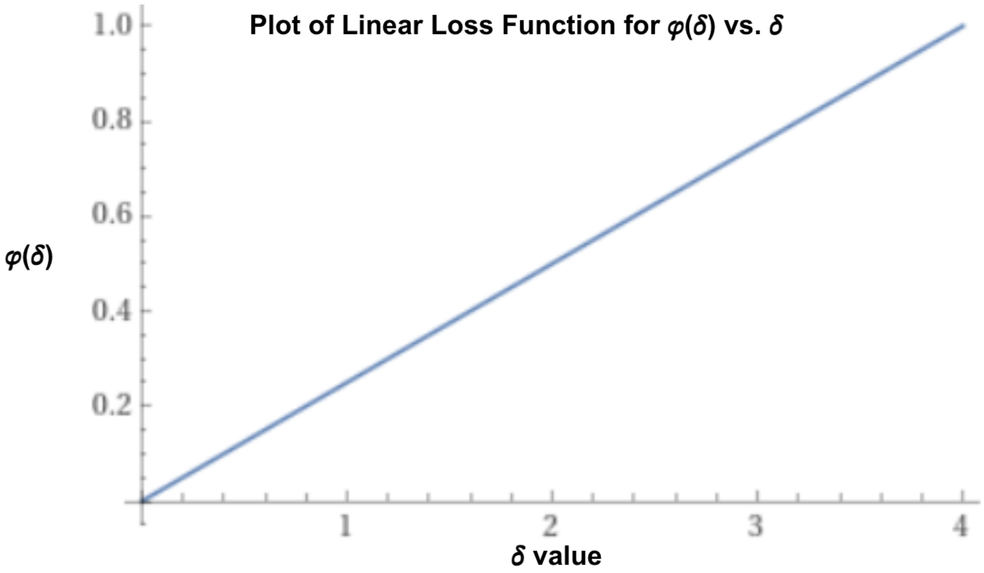


Figure 4.2: Plot of a linear loss function $\phi(\delta)$ vs δ for our toy example from Equation 4.7. This is the same loss function we use in later experiments and is suitable for Plutchik's 8 emotions.

Logistic Loss Function We can specify a logistic function as the loss function. The general form of a logistic function is shown in Equation 4.8 and illustrated in Figure 4.2 using suitable values for our use case. The logistic function gives low loss for δ values of 0 and 1, and high loss to δ values of 3 and 4. This can be particularly useful if the use-case requires the predicted emotion to be either correct or "out-by-one" but penalises more for being outside that range and particularly for being the opposite emotion or "out-by-one" of the opposite emotion. We choose values for $L, k,$ and δ_0 to suit our use case but they can be specified by users to change the function and match their requirements.

$$\begin{aligned}
&\text{Logistic Loss function:}\\
&\qquad f(x) = \frac{L}{1+e^{-k(x-x_0)}}\\
&\qquad \phi(\delta) = \frac{L}{1+e^{-k(\delta-\delta_0)}}\\
&\qquad \text{where } L = \text{max value, we choose } L = 1\\
&\qquad k = \text{growth rate, we choose } k = 3\\
&\qquad \delta_0 = \text{midpoint, we choose } \delta_0 = 2\\
&\qquad \phi(\delta) = \frac{1}{1+e^{-3(\delta-2)}}
\end{aligned} \tag{4.8}$$

Quadratic Loss Function We can specify a quadratic function as the loss function. The general form of a quadratic function is shown in Equation 4.9 and illustrated in 4.4. The quadratic function might seem an unusual choice for a loss function because it has minimal loss for both $\delta = 0$ and $\delta = \delta_{max}$. However, one scenario where this might be desirable is predicting both synonyms and antonyms.

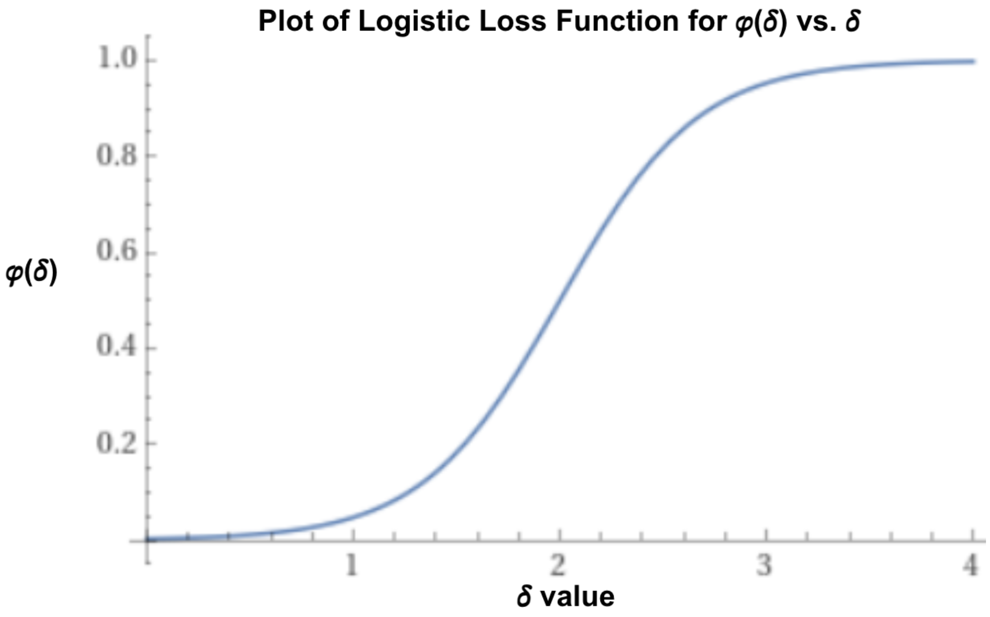


Figure 4.3: Plot of a logistic loss function $\phi(\delta)$ vs δ using Equation 4.8.

Quadratic Loss function:

$$\begin{aligned} f(x) &= ax^2 + bx + c \\ \phi(\delta) &= a\delta^2 + b\delta + c \qquad (4.9) \\ \text{where } a &= -\frac{1}{4}, b = 1, c = 0 \\ \phi(\delta) &= -\frac{\delta^2}{4} + \delta \end{aligned}$$

Step Loss Function We can specify a step function as the loss function. The linear loss function we used in our toy example could instead be described by a step function, as shown in Equation 4.10. Step functions are not continuous, unlike the other functions described, and step functions can be defined as required by specific use-cases which may not conform well to known functions or are simply not continuous. For example, we can specify a step loss function which will make ψ behave like the Accuracy metric as shown in Equation 4.11. Thus, Accuracy is a specific use case of one type of loss function $\phi(\delta)$ which can be used with the ψ metric.

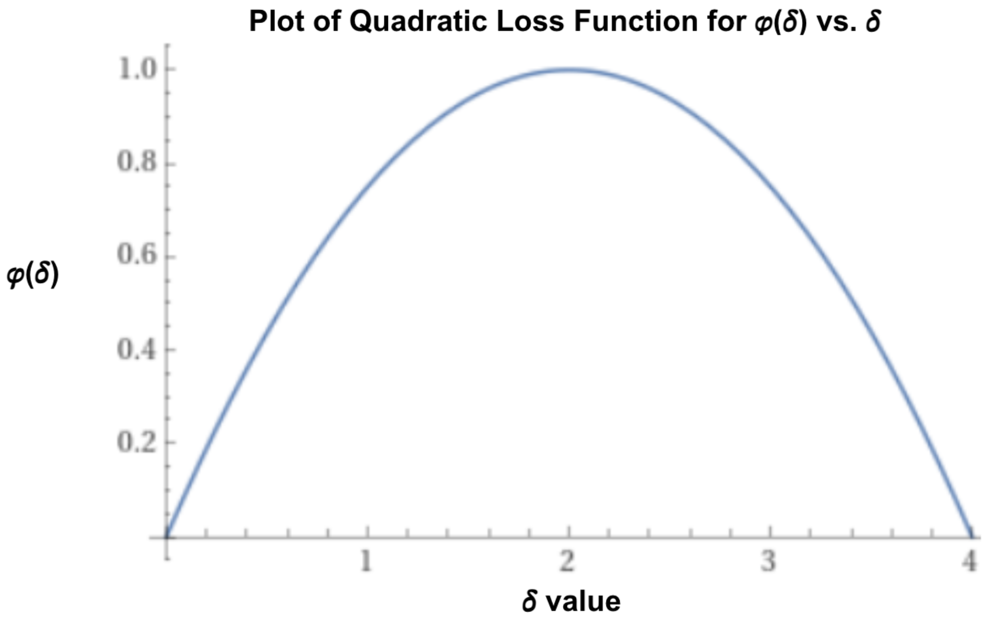


Figure 4.4: Plot of a quadratic loss function $\phi(\delta)$ vs δ using Equation 4.9.

Step Loss function:

$$\phi(\delta) := \begin{cases} 0 & \delta = 0 \\ 0.25 & \delta = 1 \\ 0.50 & \delta = 2 \\ 0.75 & \delta = 3 \\ 1 & \delta = 4 \end{cases} \tag{4.10}$$

Step Loss function - behave like Accuracy:

$$\phi(\delta) := \begin{cases} 0 & \delta = 0 \\ 1 & \delta \geq 1 \end{cases} \tag{4.11}$$

Other Loss Functions There are many other loss functions possible which we will not cover, including polynomial functions, exponential functions,[15] and

---

[15] Exponential functions can be tricky with 0 values, so a geometric transformation could be used for example from (0,0) to (1,1) to facilitate their use.

other functions which domain experts can apply as required. We do not prescribe limitations on the $\phi(\delta)$ loss function - it can be defined by users to match their specific use case. Some caveats do apply - such as the ψ function should return a value in the range 0..100 so that it is consistent and comparable with other ψ scores. As such one would expect the $\phi(\delta)$ loss function to return a value in the range 0..1.

4.5 Conclusion

In this chapter we answer Research Question 2 in the affirmative: "*Is it possible to develop a framework for measuring and evaluating classifiers of Plutchik's 8 Emotions which takes account of their dimensional relationship?*"

We introduced a novel flexible framework ψ for measuring and evaluating classifiers of Plutchik's 8 emotions. We elevated Plutchik's Wheel from the `Ordinal` scale to the `Interval` scale by introducing δ - the distance between two emotions. We defined a loss function $\phi(\delta)$ which converted δ value to a loss. All of this was demonstrated with a toy example and calculations. Real world usage will be shown in §6.4.3 for measuring affective coherence, and we will see in §5.4.2 the correlation between ψ, $F1_\mu$, MAE and RMSE, as compared to Accuracy (Precision, Recall, and F-score).

We compared ψ to popular metrics like Accuracy, MAE and RMSE and highlighted the many advantages of ψ including its ability to give fine-grained results in a familiar easy to interpret 0..100% scale. The distribution of counts for n_δ in Table 4.4 demonstrated that ψ provides detail on how a classifier is performing which can be critical information for classifier developers and users. The flexibility and extensibility of ψ as a framework was demonstrated with multiple loss functions $\phi(\delta)$ defined for various potential use-cases. Continuous and step functions were shown along with calculations and plots, to illustrate their usefulness.

We will now use the ψ metric to answer RQ3 by evaluating and selecting an emotion classifier. The selected emotion classifier will be used in RQ4 to generate affective profiles.

Chapter 5

Developing and Selecting a Dimensional Emotion Classification Model

In this chapter we answer Research Question 3: "*Can we exploit the dimensional relationship between Plutchik's 8 emotions to enable optimal selection of an automatic emotion classifier?*". RQ3 was described in §1.2, with Related Work discussed in §2.3.4. We describe how we create, evaluate and select our own automatic emotion classifier, having discussed previously (§2.3.4) the absence of an automatic emotion classifier suitable for our specific needs (Plutchik's 8 emotions in the news domain). We will then use this emotion classifier in Chapter 6 to generate affective profiles for the NEWSsynth dataset, and compare the affective profile for human versus synthetic news articles.

In Chapter 3 we created several emotion classifiers (*emoBERT*) which, while not directly suitable for our needs here, inform our methodology (§5.1) for RQ3. Specifically, the lessons from Chapter 3 inform the choices we now make regarding the fine-tuning process, optimal dataset selection, label balancing, PLM selection, and emotion-model selection. We will see that the results (§5.4.1) for our new emotion classification models are slightly better than those in Chapter

3 (§3.2.4) - but on the harder task of classifying Plutchik's 8 emotions rather than Ekman's 6 emotions.

We begin with a description of our experimental Method (§5.1) and Experiments (§5.2) followed by Results (§5.3) and Analysis (§5.4). Our Analysis includes additional experiments correlating ψ with other metrics (§5.4.2), and zero-shot NLI for emotion classification (§5.4.3), and finishes with our Conclusions (§5.5).

5.1 Method

We develop an automatic emotion classifier. Similar to our approach in RQ1 (§3.2.4), we fine-tune a PLM on the task of emotion classification using an appropriate dataset and suitable evaluation metrics. We take advantage of our results from RQ1 and apply them in RQ3; specifically in relation to dataset selection, dataset balancing, and PLM selection. In addition, we fine-tune a suite of ensembled models on the same task. Unlike RQ1, we use Plutchik's 8 emotions (elevated to the `Interval` scale) rather than Ekman's 6 basic emotions (`Nominal` scale). The best performing individual model, or ensemble, is selected based on the ψ score and will be used to generate affective profiles to address RQ4 in Chapter 6.

5.2 Experimental Setup

In this section, we give a brief reminder of our choice of emotion model, before describing the dataset preparation, and choice of PLMs that were fine-tuned, and detailing the fine-tuning process itself.

5.2.1 Emotion Model

In RQ2 we elevated Plutchik's 8 emotions from the `Ordinal` scale to the `Interval` scale by introducing δ - the distance between emotions. Now we build upon this work by using Plutchik's 8 emotions in concert with ψ to evaluate the performance of the PLMs we fine-tune on the task of emotion classification.

Furthermore, by using Plutchik's 8 emotions in our emotion classifiers we get a more granular breakdown of the emotions (and more granular affective profile) in the NEWSsynth dataset in RQ4; compared to Ekman's 6 emotions. We previously alluded to some of the other advantages of using Plutchik in concert with ψ which will be discussed in Chapter 6.

5.2.2 Dataset for Fine-Tuning Emotion Classifiers

There is no suitable dataset available for Plutchik's 8 emotions in the news domain. Consequently, we construct a new dataset from existing datasets in the news domain labelled using emotions. Our choices in creating this dataset are informed by our results from RQ1 - in relation to choice of datasets, emotion mapping, and label balancing.

Datasets We saw, in Table 3.6, the best performing emoBERTsynth models were those fine-tuned, for the emoBERT phase, using GoodNewsEveryone (GNE), and GoodNewsEveryone combined with Affective Text (GNE+AT). We also know that, generally speaking, more (suitable) data tends to result in better performing models. As a result, we combine GNE+AT into a single dataset.

Emotion Mapping We map GNE to Plutchik's 8 emotions using the mapping schema shown in Table 5.1. This mapping is similar to RQ1, Table 3.1, where we mapped the 15 emotions of the GNE dataset to Ekman's 6 emotions.

The mapping schema is consistent with Bostan and Klinger (2018) with two exceptions, i.e. 'pride' and 'love/like', which do not map directly to Plutchik's 8 emotions. Bostan and Klinger (2018) do not list 'pride' in their mapping. We map 'pride' to 'anticipation' because, according to TenHouten (2017) 'pride' is a combination of 'joy' and 'anger', and according to Plutchik's Wheel (Figure

GoodNewsEveryone		Plutchik	%
trust, love/like	→	trust	(4%)
joy	→	joy	(5%)
disgust	→	disgust	(8%)
fear	→	fear	(8%)
pride, positive anticipation / optimism, negative anticipation / pessimism	→	anticipation	(14%)
sadness, guilt, shame	→	sadness	(14%)
anger, annoyance	→	anger	(18%)
negative surprise, positive surprise	→	surprise	(30%)

Table 5.1: Emotion Mapping Schema: GoodNewsEveryone (15 emotions) to Plutchik 8 emotions. % shows the emotion label distribution in the dataset.

2.1) 'anticipation' is the emotion which lies between 'joy' and 'anger'.

Bostan and Klinger (2018) map 'love/like' individually to 'love' and 'trust' respectively. According to Plutchik's Wheel, 'love' is a combination of 'joy' and 'trust' - so we could reasonably choose either. We choose to map 'love/like' to 'trust' because 'like' is mapped to 'trust' by Bostan and Klinger (2018); 'love' is partially comprised of 'trust' according to Plutchik's Wheel. In addition 'trust' occurs less in the datasets and doesn't occur in AT at all since 'trust' is not one of Ekman's 6 basic emotions; thus it can contribute towards balancing the dataset labels.

It is notable that the dimensional relationship between emotions on Plutchik's Wheel can help resolve emotion mapping issues. This highlights a very real and practical advantage of an `Ordinal` scale like Plutchik's Wheel, over a `Nominal` scale like Ekman's basic emotions.

There is no mapping required for the AT dataset because it is labelled with Ekman's 6 emotions - which are a subset of Plutchik's 8 emotions.

Label Balancing The emotion distribution for the GNE, AT, and GA (GNE + AT) datasets are shown in Table 5.2. The emotions (labels) are very imbalanced. We noted in RQ1 (§3.2.4) the varying performance of emotion classification correlating with emotion distribution in the dataset, as shown by the Confusion Matrix in Figure 3.2. Consequently, we upsample and downsample the emotions so they are balanced across the combined GA dataset - which we refer to as GAB (GNE + AT Balanced).

Emotion	GNE (Plutchik)	AT (Ekman)	GA (GNE+AT)	GAB (GA Balanced)
trust	187	0	187	1000
joy	264	441	705	1000
disgust	382	42	424	1000
fear	419	194	613	1000
anticipation	707	0	707	1000
sadness	680	265	945	1000
anger	885	91	976	1000
surprise	1476	217	1693	1000
Total	5000	1250	6250	8000

Table 5.2: Table showing the emotion counts for the GNE and AT datasets, combing those datasets (GA), and upsampling and downsampling to balance across all 8 Plutchik Emotions (GAB).

GNE+AT were combined (GA) and shuffled randomly using a seed (313) for reproducibility. The shuffled GA was separated into 8 labels (8 Plutchik emotions). Each label was split into train, validation and test splits.[1] Each of these splits was upsampled or downsampled, according to the required split ratio, to give a total of 1,000 examples per label.[2] This yields a GAB dataset of 8,000 examples distributed evenly at 1,000 examples per Plutchik emotion, with a train-validation-test split ratio of 7-1-2.

We avoid using synthetic text to create additional examples because we use

---

[1] It is crucial to split the dataset (train, validation, test) *before* balancing - otherwise duplicate examples could appear in multiple splits - thereby contaminating the validation and test splits.

[2] Downsampling was necessary only for 'surprise' and was done by removing the extraneous examples. Upsampling was required for the other 7 emotions. Upsampling was done by copying all of the examples for a label until the total amount of examples met or exceeded the desired amount of examples - the extraneous examples were then removed. This approach was taken, rather than random sampling, because some of the labels e.g. 'disgust' must be upsampled from 424 examples to 1,000 examples. Thus it makes more sense to copy the entire set of examples rather than randomly select individual examples to copy one at a time. For this reason, after balancing the GA dataset, we randomly shuffled the splits again using a different seed for each split (train, val test : 17, 11, 23 respectively).

the GAB dataset to fine-tune emotion classifiers which then generate affective profiles in order to contrast human versus synthetic text. At this point we do not want to mix human and synthetic text. Future work could investigate the effect of exposure bias and balancing the dataset with human and synthetic news annotated with emotions.

In RQ1 we used a train:validation:test split ratio of 5-1-4 (specifically 10k-2k-8k for NEWSsynth and RealNews-Test) to be consistent with Zellers et al. (2019) and make comparison of results e.g. Table 3.3 more meaningful. We also used this ratio for emoPLM models throughout, for example in the ablation experiments in Table 3.6. However, in RQ3 we use the more common train-validation-test split ratio of 7-1-2.

5.2.3 Fine-Tuning Emotion Classifiers

We use the GAB dataset of 8,000 examples balanced across Plutchik's 8 emotions to fine-tune 15 PLMs using a train-validation-split ratio of 7-1-2 i.e. 5,600-800-1,600 examples. The models are fine-tuned for 10 epochs, with the best performing epoch on validation (based on $F1_\mu$) run against the test split. We evaluate 6 different ensembles of models on the validation split, and the best performing ensemble (based on $F1_\mu$) is run against the test split. The other hyper-parameters are the same as those used in RQ1. All models were downloaded from the HuggingFace model hub.[3] Fine-tuning, validation and testing were run on an NVIDIA GeForce RTX 3090 with 24GB RAM.

The 15 individual PLMs include models used in RQ1, such as BERT$_{\text{base}}$, BERT$_{\text{large}}$, and BLOOM 560M - all used for emoPLMsynth; and DistilBERT$_{\text{base}}$. We also use BLOOM 1B1; RoBERTa$_{\text{large}}$; ALBERT - base, large, and xxlarge; XLM; XLM-RoBERTa -base and large; XLNet - base and large; and DeBERTa$_{\text{xlarge}}$. This gives a broad range of model architectures including CLMs and MLMs, pretrained on different datasets, with different

---

[3] `https://huggingface.co/models`

learning objectives, including monolingual and multilingual models, with different model sizes.

We also use 6 ensembles of the above 15 models. We rank the models in descending order based on their performance on the validation set according to $F1_{\mu}$ score. The 6 ensembles are the top-5, middle-5, bottom-5, top-10, bottom-10 performing individual models, and all-15 models. We use a majority vote to determine the prediction for each ensemble. In the event of a tie for an ensemble, we select the highest ranking model which voted for the tied predictions - and choose its individual prediction as the ensemble prediction.

5.3 Results

The results for the 15 emotion classification models, and best performing ensemble, against the GAB test set are shown in Table 5.3. Ensemble-top5 is the best performing model based on $F1_{\mu}$ score, with RoBERTa-large a close second. ψ and MAE rank the Ensemble-top5 as as the second best performing model, while ψ, MAE and RMSE rank RoBERTa-large as the best performing model.

We can summarise our observations on the 15 models and Ensemble-top5 (against GAB test) as follows:

- $F1_{\mu}$ ranks Ensemble-top5 first, whereas ψ and MAE rank it second and RMSE ranks it third.
- ψ and the continuous metrics (MAE and RMSE) both rank RoBERTa-large first; whereas $F1_{\mu}$ ranks it second.
- All 4 metrics rank the same classifiers in the top-4 i.e. Ensemble-top5, RoBERTa-large, XLM-RoBERTa_large, and DeBERTa_xlarge_v2.
- The results appear to show a good correlation amongst the metrics for scores and for ranking. We see this again in Table 5.4, which shows the results of the 5 ensembles on the GAB validation set.

PLM	F1$_\mu$ ↑	ψ	MAE	RMSE
XLM	17.44	48.67	2.05	2.45
BLOOM_1B1	29.63	61.22	1.55	2.05
BLOOM560M	31.13	60.39	1.58	2.09
ALBERT_xxlarge_v2	31.38	61.20	1.55	2.06
ALBERT_base_v2	32.81	63.20	1.47	1.99
XLNet_large	33.88	63.13	1.48	1.99
XLM-RoBERTa_base	35.25	64.75	1.41	1.94
BERTlarge	35.25	64.84	1.41	1.94
XLNet_base	36.31	63.97	1.44	1.99
BERTbase	36.38	65.02	1.40	1.94
ALBERT_large_v2	36.38	65.83	1.37	1.90
DistilBERT-base	36.44	65.25	1.39	1.93
DeBERTa_xlarge_v2	37.25	67.02	1.32	1.85
XLM-RoBERTa_large	38.69	66.55	1.34	1.88
RoBERTa-large	40.50	**68.16**	**1.27**	**1.82**
Ensemble-top5	**40.75**	67.69	1.29	1.86
Mean	34.34	63.55	1.46	1.98
σ	5.33	4.44	0.18	0.14

Table 5.3: Table showing results for 15 emotion-classification PLMs and ensemble (best of 6 ensembles on Validation, see Table 5.4 for ensemble results) against the GAB test set, which were fine-tuned with X-Entropy loss, and validated on $F1_\mu$. Note that for MAE and RMSE lower scores are better. The table is sorted by $F1_\mu$ ascending. Best score in bold, second best score underlined. The mean and standard deviation (σ) are shown for each metric.

- There appears to be a strong negative correlation between ψ and RMSE, and a very strong negative correlation between ψ and MAE.

Table 5.4 shows the results of the 6 ensembles on the GAB validation set. The best performing ensemble on validation is Ensemble-top5 - according to all 4 metrics. Ensemble-top5 was run against the GAB test set and is included in Table 5.3 where we saw it is ranked first by $F1_\mu$ score.

We can summarise our observations on the 6 ensembles (against GAB validation) as follows:

- All 4 metrics have the same ranking of the six ensembles. RMSE has a tie for the first and second ensembles. $F1_\mu$ has a tie for the fourth and fifth ensembles.

Ensemble Name	F1$_{\mu}$ ↑	ψ	MAE	RMSE
Ensemble-bottom5	38.88	65.84	1.37	1.93
Ensemble-middle5	43.75	69.16	1.23	1.82
Ensemble-bottom10	43.75	70.06	1.20	1.78
Ensemble-all15	46.50	71.75	1.13	1.73
Ensemble-top10	<u>47.25</u>	<u>72.75</u>	<u>1.09</u>	**1.68**
Ensemble-top5	**47.38**	**72.97**	**1.08**	**1.68**

Table 5.4: Table showing ensembles of models against the validation set. The best performing ensemble, Ensemble-top5, was subsequently run against the test set as shown in Table 5.3.

- As one might expect, the ensembles of better performing models tend to out-perform the ensembles of less well performing models.
- Ensemble-b10 and Ensemble-m5 have the same $F1_{\mu}$ score, in a tie on $F1_{\mu}$ the ensemble with the higher ψ score is considered the winner.

5.4 Analysis

In this section we analyse the results of the main experiment i.e. fine-tuning PLMs on the task of emotion classification, and choose the best performing model. We then analyse the correlation between the results of the experiment using four metrics ψ, $F1_{\mu}$, MAE and RMSE. Finally, we conduct an additional experiment, using Natural Language Inference (NLI) in the task of zero-shot emotion classification, to determine if NLI yields a model that performs better than the fine-tuned models.

5.4.1 Choosing an Emotion Detection Model

The $F1_{\mu}$ results are in the range 17.44% to 40.75%, although XLM is an outlier at 17.44%. The random chance for 8 emotions is 12.5% and all of the models, except XLM, perform well above this. The mean $F1_{\mu}$ is 34.34% across all 15 models and Ensemble-top5. The top ranked Ensemble-top5 has an $F1_{\mu}$ of 40.75%. This outperforms our earlier emoBERT model (§3.2.4) which had an

$F1_\mu$ of 39.40%. However, emoBERT classifies only 6 emotions with a random chance of 16.7%, so it is more difficult to classify 8 emotions with a random chance of 12.5%. The results are within the range 31% to 98% reported by Bostan and Klinger (2018) for within-corpus emotion classification. Nonetheless, an $F1_\mu$ score of 40.75% is not high and shows the difficulty of the task of emotion classification.

We use the ψ score as the metric to choose the best performing model or ensemble. Therefore, the best performing model is RoBERTa-large. We will use RoBERTa-large as the emotion classifier to generate affective profiles in RQ4 which we answer in Chapter 6. Our choice of RoBERTa-large is consistent with Labat et al. (2022) who reported RoBERTa performed best for emotion classification, although their domain was tweets. However, in contrast to our results, Labat et al. (2022) reported that XLM performed well. We believe this difference is due to our XLM model having a poor random initialisation. Our analysis showed that after the second epoch of fine-tuning the XLM model performance dropped from an $F1_\mu$ of 20.25% to 12.50% where it was constant for the remaining 8 epochs of fine-tuning. The model only predicted three emotions (Joy, Trust, Surprise) from the possible 8 emotions. Poor random initialisation is typically resolved by using a different random seed, however for consistency we used the same seed for all 15 individual models.

One advantage of using a single model, as opposed to an ensemble, at inference time is it uses less resources (e.g. carbon, compute, time, memory) compared to running and combining 5 models as required by Ensemble-top5. This is important in use cases where large amounts of text must be classified on an ongoing basis e.g. social media platform, large e-tailers, etc. The carbon cost is a factor for the environment during development and deployment for the lifetime of model. Making models more efficient (Jooste et al., 2022) is another way to reduce the carbon cost and environmental impact.

5.4.2 Correlating ψ with $F1_{\mu}$, MAE, and RMSE

The ψ scores are numerically quite different to the $F1_{\mu}$ scores. The ψ scores range from 46.87% to 68.16%, with a mean of 63.55%. In contrast, the $F1_{\mu}$ scores range from 17.44% to 40.75%, with a mean of 34.34%. The higher numerical values for ψ, compared to $F1_{\mu}$, support the contention that ψ gives a graduated loss to incorrect predictions. The amount of loss is proportional to the δ value between the correct emotion and the predicted emotion. In our case we use a linear loss function, $\phi(\delta) = \frac{\delta}{\delta_{max}}$, as described in §4.2.2 and §4.4.

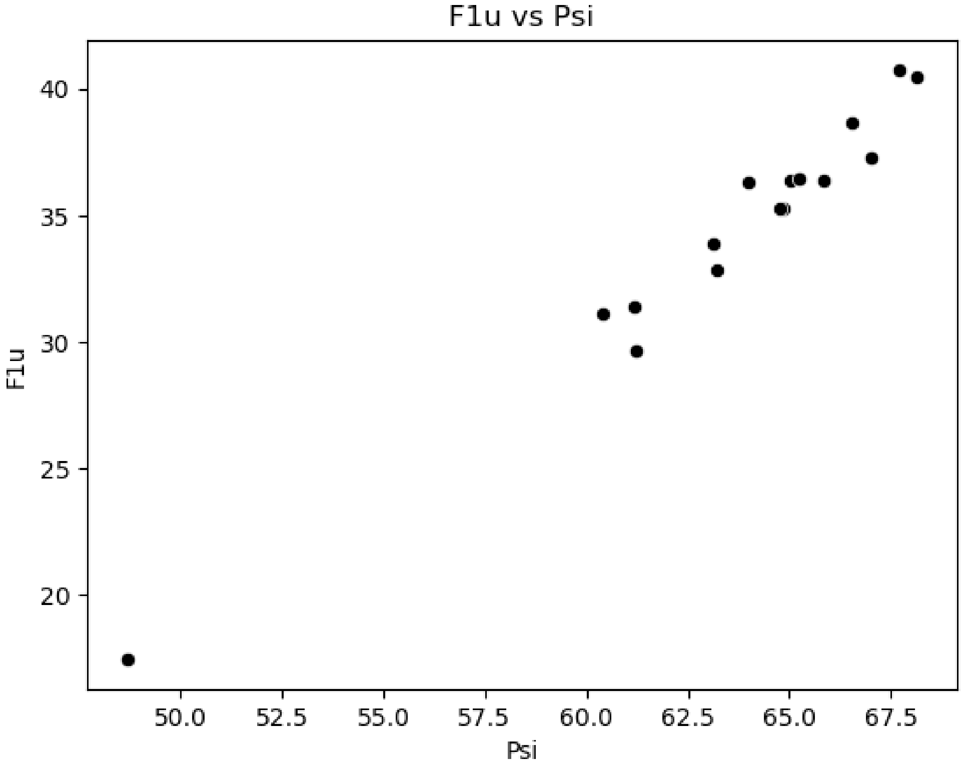


Figure 5.1: Scatter Plot of $F1_{\mu}$ vs ψ for 15 models and 1 ensemble on the GAB test set. XLM is the outlier. There is strong positive correlation between the metrics.

While the numerical scores for ψ and $F1_{\mu}$ are quite different, we want to know how well ψ scores correlate with $F1_{\mu}$, MAE, and RMSE. We also want to know the correlation between the ranking of the models for each of the metrics i.e. how similar is the order of the models when we rank them according to the different metrics. To address these questions we create three scatter plots of

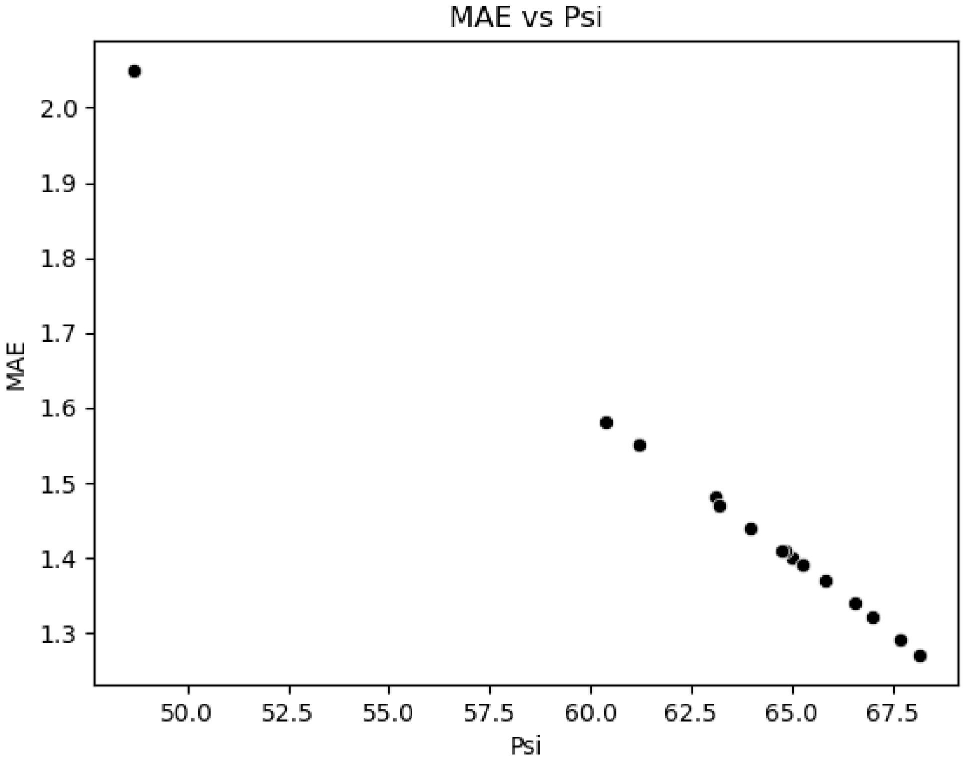


Figure 5.2: Scatter Plot of MAE vs ψ for 15 models and 1 ensemble on the GAB test set. XLM is the outlier. There is very strong negative correlation between the metrics.

$F1_{\mu}$, MAE, and RMSE vs ψ; and calculate the Pearson ρ correlation, and the Spearman ρ correlation respectively.

Figure 5.1 is a scatter plot of $F1_{\mu}$ vs ψ and shows a strong positive correlation. Figures 5.2 and 5.3 are scatter plots of MAE vs ψ and RMSE vs ψ respectively. They both show a strong negative correlation between ψ and MAE, and ψ and RMSE. The Pearson correlation, ρ, between each of the metrics is shown in Table 5.5. We recall from Table 2.4 that $\rho > 0.7$ indicates a strong relationship. The ρ values are all above 0.97 which shows strong correlation between all the metrics.

	ψ	$F1_{\mu}$	MAE	RMSE
ψ	1.000000	0.984394	-0.999899	-0.997641
$F1_{\mu}$	0.984394	1.000000	-0.985005	-0.975421
MAE	-0.999899	-0.985005	1.000000	0.997329
RMSE	-0.997641	-0.975421	0.997329	1.000000

Table 5.5: Pearson correlation ρ between four metrics ψ, $F1_{\mu}$, MAE, and RMSE.

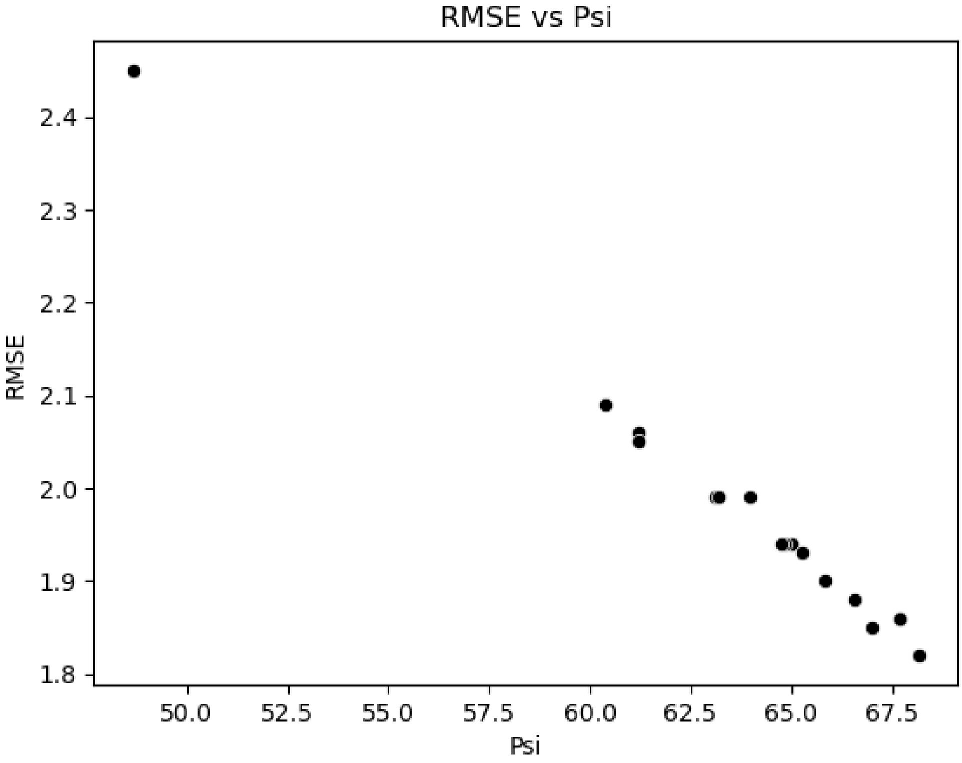


Figure 5.3: Scatter Plot of RMSE vs ψ for 15 models and 1 ensemble on the GAB test set. XLM is the outlier. There is very strong negative correlation between the metrics.

We can summarise our observations on the scatter plots and Pearson ρ correlation of metrics as follows:

- The scatter plots show a strong (positive) correlation between ψ and $F1_{\mu}$, with a Pearson correlation of 0.984.
- There is a strong (negative) correlation between ψ and MAE, and ψ and RMSE; with a Pearson correlation of -0.999 and -0.997 respectively. We expect a negative correlation since a higher score is preferred for ψ, while a lower score is preferred for MAE and RMSE.
- The Pearson correlations between all metrics are strong.
- It is interesting that the absolute correlation between ψ and MAE and RMSE is higher than the absolute correlation between ψ and $F1_{\mu}$. This is perhaps because ψ, MAE, and RMSE are all derived from δ, and therefore are more granular than the coarse binary correct/incorrect nature of $F1_{\mu}$ (and Accuracy, Precision, and Recall) as discussed in Chapter 4.

Thus far we have seen that, numerically, ψ is quite different to $F1_\mu$, but both are strongly correlated with a Pearson ρ of 0.984. We still have one question to answer - how do the different metrics affect the ranking of the models? The ranking is crucial because it typically determines which model is selected for use in a downstream task. We calculate the Spearman ρ correlation and show the results in Table 5.6. The results show a strong correlation between all metrics, indicating that the ranking of the emotion classifiers is quite similar across all four metrics.

	ψ	$F1_\mu$	MAE	RMSE
ψ	1.000000	0.967600	-0.998528	-0.991142
$F1_\mu$	0.967600	1.000000	-0.971976	-0.946676
MAE	-0.998528	-0.971976	1.000000	0.991121
RMSE	-0.991142	-0.946676	0.991121	1.000000

Table 5.6: Spearman correlation ρ between four metrics ψ, $F1_\mu$, MAE, and RMSE.

We summarise our observations on the Spearman ρ correlation of metrics:

- The Spearman ranking correlations show strong correlation between the rankings of all 4 metrics.
- It is notable that, for Spearman, ψ has a stronger correlation with MAE and RMSE than it does with $F1_\mu$. This is consistent with the Pearson ρ correlation, and could be because all three are derived from δ as mentioned earlier.
- The strong Spearman correlation between ψ and MAE, and ψ and RMSE supports the hypothesis that the ψ metric takes advantage of the fine-grained exactitude of the continuous metrics.

Given the strong Pearson and Spearman correlations between ψ and the other metrics, one might ask if there is any benefit from the introduction of ψ. We believe the answer is a definitive "yes". There is a benefit from the ψ score (and the $\phi(\delta)$ loss function, and δ - which we used to calculate the MAE and RMSE scores). The ψ score is easy to interpret - like $F1_\mu$, Accuracy, Precision, and Recall. But, unlike those metrics, ψ benefits from the same granularity as MAE and RMSE - as indicated by higher scores than $F1_\mu$. Notwithstanding the strong Spearman correlation between ψ and $F1_\mu$ ranking the emotion classifiers, ψ ranked RoBERTa-large as the top performing model whereas $F1_\mu$ ranked Ensemble-top5 as the top performing model/ensemble. This is crucial because our ultimate choice of emotion classifier for building affective profiles in RQ4 is determined by the choice of metric. This demonstrates that ψ can be impactful. Note that MAE and RMSE also rank RoBERTa-large as the top performing model. The strong correlation between the 4 metrics also acts as a validation for the ψ score. It demonstrates that ψ is consistent with existing metrics and is not a radical departure from existing norms.

5.4.3 Zero-Shot NLI for Emotion Classification

We conducted experiments using zero-shot NLI for emotion classification. NLI was discussed in (§2.3.4) including advantages of this approach over fine-tuning, and the role of prompting.

We use fifteen models for zero-shot NLI classification of Plutchik's 8 emotions using the GAB dataset. We measure the performance of these models using four metrics: ψ, F1μ, MAE and RMSE as shown in Table 5.8. Similar to the work of Basile et al. (2021) and Plaza-del Arco et al. (2022) we use RoBERTa-large, BART-large, XLM-RoBERTa-large, and DeBERTa-xlarge. We include several models used by Bareiß et al. (2024) and made available by Laurer et al. (2024) - such as Distilled XLM-RoBERTa-large, RoBERTa, XLM-V-base, and mDeBERTa all fine-tuned on XNLI and MNLI. We include variants of some of these models in addition to the two top performing models (at the time of running the experiment) published by Laurer et al. (2024) i.e. deberta-v3-large-zeroshot-v2.0 and deberta-v3-base-zeroshot-v2.0. All of the models are available on HuggingFace.[4] The 15 models include several different architec-

---

[4] `https://huggingface.co/models`

tures, training datasets and domains, and multiple languages.

The HuggingFace zero-shot-pipeline was used whereby a *premise* (news headline) and a list of *hypotheses* (Plutchik's 8 emotions) were sent to the model via the zero-shot pipeline. The pipeline returns an *entailment* (prediction) which is a soft-maxed probability for each *hypothesis* (emotion). The *hypothesis* (emotion) with the highest probability is *entailed* (predicted). As discussed previously (§2.3.4) we use a prompt type "EMO-Name" as defined by Plaza-del Arco et al. (2022). Table 5.7 shows an example of using zero-shot NLI with the "EMO-Name" prompt.

Premise	`‘Ukraine election:Comedian Volodymyr Zelensky wins’`	
Hypotheses	`[‘joy’, ‘sadness’, ‘fear’, ‘anger’, ‘disgust’, ‘surprise’, ‘trust’, ‘anticipation’]`	
Entailment	`disgust`	`0.0039077484980225560`
	`anger`	`0.0044322255998849870`
	`fear`	`0.0047090123407542706`
	`sadness`	`0.0051015573553740980`
	`trust`	`0.0416014716029167200`
	`anticipation`	`0.0712974146008491500`
	`surprise`	`0.1438306123018264800`
	`joy`	`0.7251200079917908000`

Table 5.7: An example of zero-shot NLI from DeBERTa-base on the GAB dataset (training split), using the "EMO-Name" prompt type from Plaza-del Arco et al. (2022). A *premise* (sentence) and *hypotheses* (emotions) are used as prompt to the model, which returns an *entailment* (prediction) i.e. a soft-maxed probability for each *hypothesis*. The *hypothesis* (emotion) with the highest probability is *entailed* (predicted) from the *premise* (sentence). In this example, `‘joy’` is the prediction of the zero-shot NLI model, which agrees with the gold label from GAB.

As shown in Table 5.8, DeBERTa-base (deberta-v3-base-zeroshot-v2.0) from Laurer et al. (2024) is the best performing model on the task of emotion classification using zero-shot NLI on the GAB test set. This is consistent with Laurer et al. (2024) who state that DeBERTa-base is their best performing base-size model[5] (mean F1-macro on 28 different text classification tasks). In contrast to our results, Laurer et al. (2024) report DeBERTa-large outperforms

---

[5] At the time of the experiments (05-April-2024) and time of writing (24-May-2024).

NLI Model	$\mathrm{F1}_\mu$	$\psi \uparrow$
multilingual-MiniLMv2-L6-mnli-xnli*	18.31	51.55
vicgalle/xlm-roberta-large-xnli-anli	18.56	53.03
microsoft/deberta-v2-xxlarge-mnli	22.19	56.38
mDeBERTa-v3-base-mnli-xnli*	24.69	56.78
mDeBERTa-v3-base-xnli-multilingual-nli-2mil7*	22.19	57.48
facebook/bart-large-mnli	24.00	57.55
xlm-v-base-mnli-xnli*	25.00	57.88
ernie-m-base-mnli-xnli*	24.44	59.31
ernie-m-large-mnli-xnli*	23.38	59.83
roberta-large-mnli	27.44	59.91
microsoft/deberta-xlarge-mnli	28.63	60.89
microsoft/deberta-large-mnli	29.94	61.42
microsoft/deberta-v2-xlarge-mnli	30.69	61.53
deberta-v3-large-zeroshot-v2.0*	<u>31.13</u>	<u>61.80</u>
deberta-v3-base-zeroshot-v2.0*	**32.63**	**63.94**
Mean	25.55	58.62
σ	4.27	3.23

Table 5.8: Table (sorted by ψ ascending) comparing 15 emotion-classification zero-shot NLI models against the GAB Test set. Best score in bold, second best score underlined. Models with an asterisk (*) are provided by Laurer et al. (2024). The mean and standard deviation (σ) are shown for ψ and $F1_\mu$.

DeBERTa-base averaged over 28 tasks (including 3 tasks related to emotion) by 5.70 percentage points on F1.[6] However, our results show DeBERTa-base outperforms DeBERTa-large by 1.50 percentage points on F1. It may be possible to get more performance from the models using few-shot NLI, and prompt engineering; we leave this for future work. Nonetheless, DeBERTa-large is ranked second by both ψ and $F1_\mu$. We compare our best performing zero-shot NLI model against our best performing fine-tuned model. The zero-shot NLI DeBERTa-base ψ score of 63.94%, is less than the ψ score of 68.16% achieved by our fine-tuned RoBERTa-large (§5.4.1). Therefore, we will use RoBERTa-large fine-tuned on GAB to create the affective profile for human and synthetic articles in the NEWSsynth dataset for RQ4 in Chapter 6.

We make the following observations, relating to ψ and $F1_\mu$, from Table 5.8:

---

[6] Laurer et al. (2024) report F1-macro scores while we report $F1_\mu$ scores. However, these are equivalent because GAB is balanced across labels and therefore $F1_\mu$ = F1-macro.

- Both ψ and $F1_\mu$ have the same order for the top 7 models.
- The standard deviation, σ, for ψ is 3.23, whereas for $F1_\mu$ it is 4.27. This indicates ψ is less volatile than $F1_\mu$. Remember, both metrics are scoring the exact same predictions - the only difference is how they measure and calculate the loss.
- The difference between the mean ψ and mean $F1_\mu$ scores differ across the two experiments i.e. fine-tuning versus zero-shot NLI for classification:
    - Mean $F1_\mu$ is 34.34% for fine-tuning, and 25.55% for zero-shot NLI.
    - Mean ψ is 63.55% for fine-tuning, and 58.62% for zero-shot NLI.
    - The difference in the mean $F1_\mu$ values is 8.79 = (34.34 - 25.55)
    - The difference in the mean ψ values is 4.93 = (63.55 - 58.62)
    - This also supports the notion that ψ is less volatile than $F1_\mu$.
- The lower difference in mean ψ (compared to the larger difference in mean $F1_\mu$) indicates that ψ is rewarding the zero-shot NLI models for their predictions being close to the correct label.
- The results using both metrics indicate that fine-tuning gives better performance than zero-shot NLI in the task of emotion classification.
- Inspecting the Confusion Matrices shows that many of the NLI models over-predict on certain emotions which negatively impacts performance.

5.5 Conclusion

In this chapter we answer Research Question 3 in the affirmative: "*Can we exploit the dimensional relationship between Plutchik's 8 emotions to enable optimal selection of an automatic emotion classifier?*"

We chose Plutchik's 8 emotions and created a custom dataset (GAB) which we used to fine-tune 15 individual PLMs and 6 ensembles. RoBERTa-large was the best performing model based on the ψ metric with a score of 68.16%. We used these results to correlate the ψ metric with $F1_\mu$, MAE, and RMSE and showed strong correlation for both Pearson and Spearman coefficients. We also used zero-shot NLI on 15 PLMs. DeBERTa-base was the best performing model with a ψ score of 63.94%. The results indicate that zero-shot NLI is not as performant as fine-tuning PLMs in the task of emotion classification.

Equipped with our custom dataset of human and synthetic articles (*NEWSsynth*) and a RoBERTa-large PLM fine-tuned on our custom dataset (*GAB*) of Plutchik's emotions, evaluated with our custom metric (ψ), we are now ready to answer RQ4 and generate affective profiles.

Chapter 6

Affective Profiling of Human and Synthetic Text

In this chapter we answer Research Question 4: "*Can we generate and compare the affective profile of human text and synthetic text?*". RQ4 was described in §1.2, with Related Work discussed in §2.3.4. We investigate human and synthetic text to determine if there is a difference in their *affective profile*, i.e. their emotional content and characteristics, and whether any such difference can be beneficial in the task of differentiating human from synthetic text. We previously saw two examples of an affective profile (§3.3). Now we generate 10k affective profiles, for human and synthetic articles from our NEWSsynth dataset, using our emotion classifier from Chapter 5.

The chapter begins with a description of our experimental Method (§6.1) and Experiments (§6.2) followed by Results (§6.3) and Analysis (§6.4). Our Analysis begins with affective coherence in emotion n-grams (§6.4.1), and includes additional experiments on affective coherence in affective profiles (emotion) of news articles (§6.4.2), and measuring affective coherence using δ-profiles and ψ (§6.4.3). Our analysis continues with experiments generating affective profiles using sentiment (§6.5), affective coherence in sentiment n-grams, and affective coherence in affective profiles (sentiment) of news articles. We finish with our Conclusions (§6.6).

6.1 Method

We generate an affective profile for each article in a dataset of human and synthetic articles. The affective profile is generated by splitting each article into a sequence of sentences, and assigning an emotion to each sentence. The affective profile for each article is the resulting sequence of emotions, for example `[joy, sadness, joy, joy]`. Each affective profile is broken into a sequence of n adjacent emotions i.e. n-grams of emotion, for example the 3-grams `[joy, sadness, joy]`, `[sadness, joy, joy]`. For each n-gram of emotion, we count the number of times the emotion changes between successive emotions in the n-gram. For the two example affective profiles above, that is `2, 1` respectively. The number of emotion changes is used because it is an indication of affective coherence, and we hypothesise affective coherence may help differentiate human and synthetic text. The reason we do not use n-grams directly is because they can become unwieldy since there are N^n permutations; where $N =$ number of emotions, and $n =$ size of n-gram. For example, the number of 5-grams for Plutchik's emotions is $8^5 = 32,768$. We then compare the distribution of the number of emotion changes within the n-grams; for human and synthetic text, and contrast these with random chance.

6.2 Experiments

We used 10k human and synthetic news articles from the NEWSsynth training split. The training split is used to avoid contaminating the test (and validation) splits, which can be used in future work. Each article was divided into a sequence of individual sentences using spaCy.[1] The emotion for each sentence was predicted by the best performing emotion classifier from Chapter 5 (i.e.

---

[1] spaCy for English (en_core_web_sm) from `https://spacy.io/models/en`

RoBERTa-large fine-tuned on GAB and evaluated using ψ). The resulting sequence of emotions, corresponding to the sequence of sentences in the article, is the affective profile for that article.

Preserving the order of sentences (and their emotions) is crucial because the affective profile is the *sequence* of emotions in an article. Furthermore, by measuring the change between successive emotions in an article, we can calculate a value for affective coherence - which we demonstrate in §6.4.3. If the order of sentences (and their emotions) is lost, then we effectively have a *bag-of-emotions* with no relationship between them.

We take the affective profile of each article (i.e. the predicted emotion for each sentence in the article) in NEWSsynth (train 10k) and calculate n-grams of emotion over the article. We will consider 3-grams for our explanation of the experimental process and reuse the example above. Given the affective profile for a news article `[joy, sadness, joy, joy]`, we extract the 3-grams: `[joy, sadness, joy]`, `[sadness, joy, joy]`. In each 3-gram, we count the number of times the emotion changes within the 3-gram. There are 2 emotion changes in `[joy, sadness, joy]` and 1 emotion change in `[sadness, joy, joy]`. For a 3-gram there can be either 0, 1, or 2 emotion changes. We count the number of occurrences of emotion changes (0,1,2) for all 3-grams across the whole dataset. We compare the distribution of emotion changes for human and synthetic news articles, and contrast this with random chance. We repeat the above process separately for n-grams of size 2, 3, 4, and 5.

It is important to include chance because it gives a baseline against which to compare the human and synthetic results. Note that, for example in 2-grams there is either 0 change or 1 change only, and for 8 emotions and 2-grams there are 8^2 permutations = 64. However, only 8 of these 64 correspond to 0 changes (i.e. `[joy,joy]`, `[sadness,sadness]`, `[fear,fear]`, `[anger,anger]`, `[disgust,disgust]`, `[surprise,surprise]`, `[trust,trust]`, `[anticipation,anticipation]`).

Affective Profile	3-grams		Emotion Changes
`[fear, fear, fear, surprise, surprise, fear, fear, fear, anticipation, fear]`	`[fear, fear, fear]`	→	0
	`[fear, fear, surprise]`	→	1
	`[fear, surprise, surprise]`	→	1
	`[surprise, surprise, fear]`	→	1
	`[surprise, fear, fear]`	→	1
	`[fear, fear, fear]`	→	0
	`[fear, fear, anticipation]`	→	1
	`[fear, anticipation, fear]`	→	2

Table 6.1: An example *affective profile*, 3-grams, and number of emotion changes; for a human article from the *NEWSsynth* dataset. The affective profile is parsed into 3-grams of emotion. The number of changes in each 3-gram is counted. We analyse the distribution of emotion changes in 3-grams and compare human versus synthetic, and random chance. We repeat the above process for 2, 3, 4, and 5-grams of emotion on the *NEWSsynth* (train, 10k) dataset.

The remaining 56 permutations correspond to 1 change (e.g. `[joy,sadness]`, `[joy,fear]`, `[joy,anger]`, `[joy,disgust]`, `[joy,surprise]`, `[joy,trust]`, `[joy,anticipation] ... [anticipation, trust]`). Therefore, for emotion 2-grams of Plutchik's 8 emotions, there is a 12.5% ($= 100\frac{8}{64}$) chance of 0 emotion changes, but an 87.5% ($= 100\frac{56}{64}$) chance of 1 emotion change.

To illustrate the experimental method, a sample affective profile is shown in Table 6.1 using a human article from the *NEWSsynth* (train, 10k) dataset. The article contains 10 sentences, and the Plutchik emotion for each sentence represents the affective profile of the article - as shown in the leftmost column of Table 6.1. We generate the emotion 3-grams from the affective profile, as shown in the centre column. We count the number of emotion changes in each 3-gram, as shown in the right hand column. We combine the 3-grams for all human articles and group them by the number of emotion changes. We then sum the count of 3-grams in each group, which yields a distribution of emotion changes (i.e. 0, 1, or 2, for 3-grams) for human articles. We repeat the last two steps - but this time for synthetic articles. We then compare the distribution of emotion changes in human articles versus synthetic articles, and random chance. We repeat the above process for 2, 3, 4, and 5-grams of emotion.

6.3 Results

The distribution of emotion changes within emotion n-grams are shown in Figures 6.1 to 6.4. We examine the distribution of emotion changes in turn by comparing: human text to random chance, synthetic text to random chance, and human text to synthetic text. We will see a consistent pattern of results emerge from all three comparisons, for 2, 3, 4, and 5-grams.

Comparing human text to random chance Figure 6.1 shows the distribution of emotion changes for 2-grams. If we first compare the human (green) to random chance (grey) we see that human articles exhibit affective coherence (0 emotion changes) in 40.07% of emotion 2-grams. This compares to 12.50% for random chance. Thus, human articles exhibit affective coherence (0 emotion changes) in emotion 2-grams well above random chance. On the other hand, human articles exhibit affective incoherence (1 emotion change) in 59.93% of emotion 2-grams, compared to 87.50% for random chance. Thus, human article emotion 2-grams show affective incoherence (1 emotion change) 27.57 percentage points below random chance.

Figures 6.2 to 6.4 illustrate the same consistent pattern, for emotion 3, 4, and 5-grams, to the pattern we see in Figure 6.1 for emotion 2-grams. That is, human text exhibits affective coherence (0 emotion changes) well above chance, and human text exhibits affective incoherence ($n - 1$ emotion changes) well below chance. For 3-grams, in Figure 6.2: human text shows affective coherence 20.90% compared to 1.56% for random chance; and affective incoherence 40.79% compared to 76.56% for random chance. For 4-grams, in Figure 6.3: human text shows affective coherence 12.41% compared to 0.20% for random chance; and affective incoherence 27.17% compared to 66.99% for random chance. For 5-grams, in Figure 6.4: human text shows affective coherence 7.98% compared to 0.02% for random chance; and affective incoherence 18.73% compared to 58.62% for random chance.

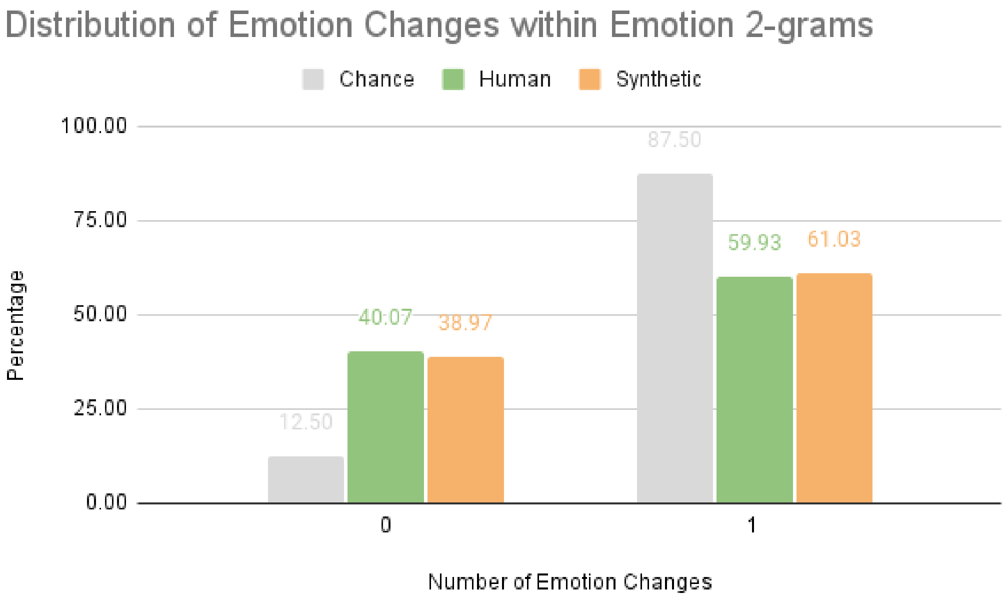


Figure 6.1: The distribution of emotion changes for emotion 2-grams. Human text exhibits affective coherence (emo changes = 0) well above chance (40.07% vs 12.50%). Human text exhibits affective incoherence (emo changes = 1) well below chance (59.93% vs 87.50%). The results indicate human text is affectively coherent. However, synthetic text exhibits similar affective coherence to human text. Nonetheless, human text is more affectively coherent than synthetic text, and human text is less affectively incoherent than synthetic text, and this difference is statistically significant ($p < .001$).

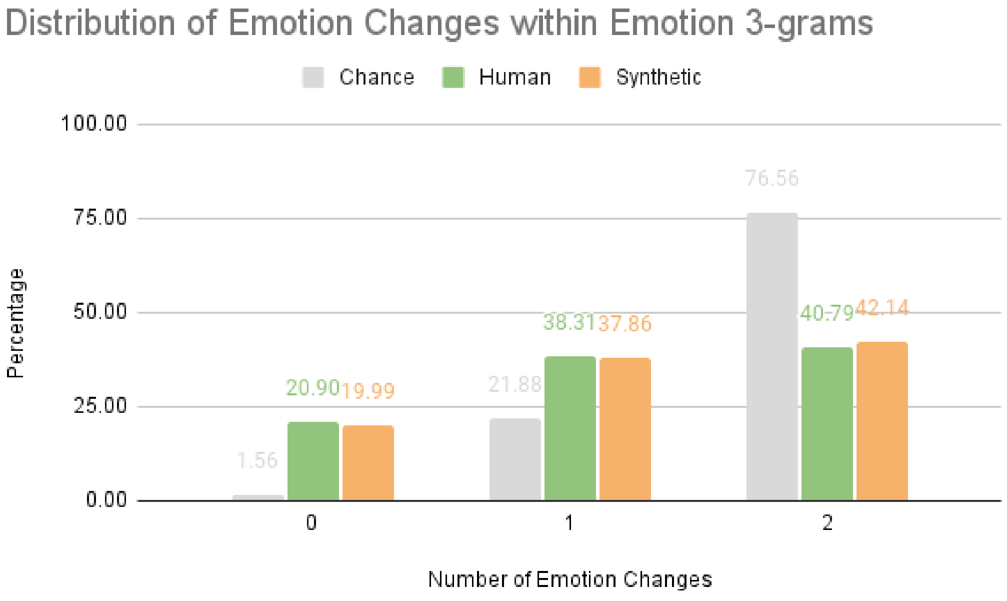


Figure 6.2: The distribution of emotion changes for emotion 3-grams. Human text exhibits affective coherence (emo changes = 0) well above chance (20.90% vs 1.56%). Human text exhibits affective incoherence (emo changes = 2) well below chance (40.79% vs 76.56%). The human and synthetic distribution difference is significant ($p < .001$). The results are consistent with 2-grams in Figure 6.1.

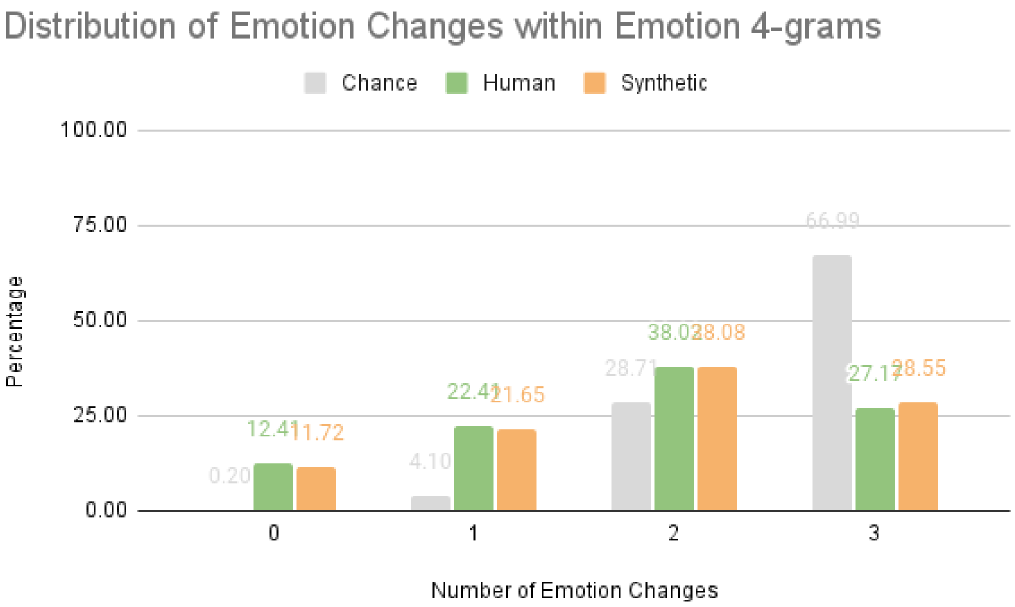


Figure 6.3: The distribution of emotion changes for emotion 4-grams. Human text exhibits affective coherence (emo changes = 0) well above chance (12.41% vs 0.20%). Human text exhibits affective incoherence (emo changes = 3) well below chance (22.17% vs 66.99%). The human and synthetic distribution difference is significant ($p < .001$). The results are consistent with 2-grams and 3-grams in Figures 6.1 and 6.2.

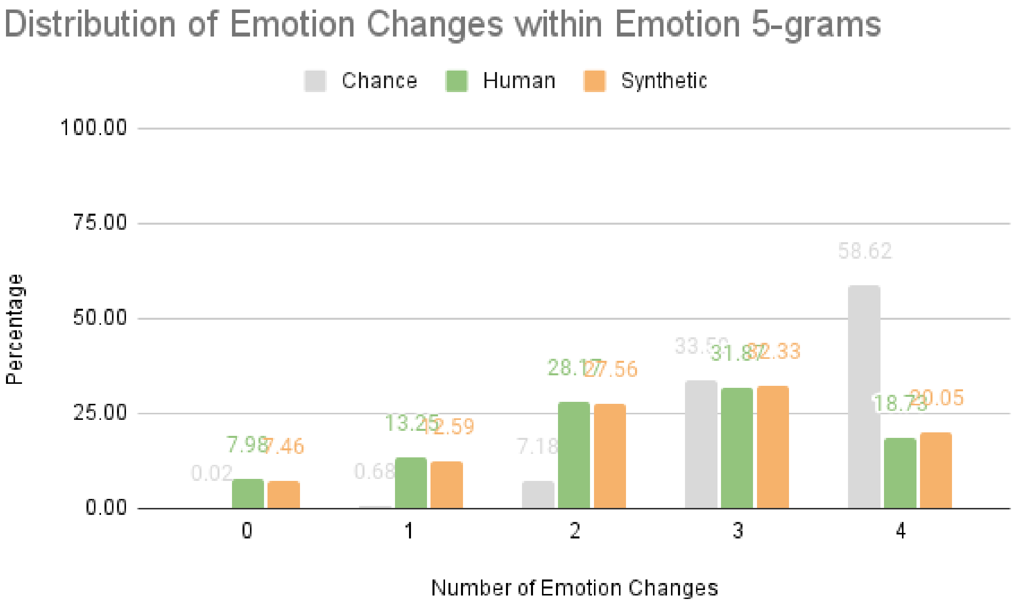


Figure 6.4: The distribution of emotion changes for emotion 5-grams. Human text exhibits affective coherence (emo changes = 0) well above chance (7.98% vs 0.02%). Human text exhibits affective incoherence (emo changes = 4) well below chance (18.73% vs 58.62%). The human and synthetic distribution difference is significant ($p < .001$). The results are consistent with 2-grams, 3-grams, and 4-grams in Figures 6.1 to 6.3.

Comparing synthetic text to random chance Figures 6.2 to 6.4 illustrate another consistent pattern for emotion 2, 3, 4, and 5-grams. Comparing synthetic text (orange) to random chance (grey) we see that, similar to human text, synthetic text exhibits affective coherence well above random chance, and synthetic text exhibits affective incoherence well below random chance.

For 2-grams in Figure 6.1: synthetic text shows affective coherence 38.97% compared to 12.50% for random chance; and affective incoherence 61.03% compared to 87.50% for random chance. For 3-grams in Figure 6.2: synthetic text shows affective coherence 19.99% compared to 1.56% for random chance; and affective incoherence 42.14% compared to 76.56% for random chance. For 4-grams in Figure 6.3: synthetic text shows affective coherence 11.72% compared to 0.20% for random chance; and affective incoherence 28.55% compared to 66.99% for random chance. For 5-grams in Figure 6.4: synthetic text shows affective coherence 7.46% compared to 0.02% for random chance; and affective incoherence 20.05% compared to 58.62% for random chance.

Comparing human text to synthetic text A third consistent pattern emerges when we compare human (green) to synthetic (orange) text in Figures 6.2 to 6.4. Human text and synthetic text exhibit similar levels of affective coherence and affective incoherence. However, human text exhibits slightly higher affective coherence and slightly lower affective incoherence than synthetic text across 2, 3, 4, and 5-grams of emotion. It is important to note that the differences between human and synthetic distributions of emotion changes in emotion n-grams ($2 \leq n \leq 5$) are statistically significant ($p < .001$). Data and calculations are provided in Appendix F.

For 2-grams in Figure 6.1: human text shows affective coherence 40.07% compared to 38.97% for synthetic text; and affective incoherence 59.93% compared to 61.03% for synthetic text. For 3-grams in Figure 6.2: human text shows affective coherence 20.90% compared to 19.99% for synthetic text; and affective incoherence 40.79% compared to 42.14% for synthetic text. For 4-

grams in Figure 6.3: human text shows affective coherence 12.41% compared to 11.72% for synthetic text; and affective incoherence 27.17% compared to 28.55% for synthetic text. For 5-grams in Figure 6.4: human text shows affective coherence 7.98% compared to 7.46% for synthetic text; and affective incoherence 18.73% compared to 20.05% for synthetic text. We reiterate that, although they may appear slight, the differences between human and synthetic distributions regarding the number of emotion changes in emotion n-grams is statistically significant in all cases ($p < .001$).

Revisiting the Chapter 3 (RQ1) Experiments We remind the reader that in Chapter 3 (RQ1) our emotionally-aware models outperformed standard models in the task of synthetic text detection. In many cases the improvements were in the order of a few percentage points. When concluding Chapter 3 we signposted that in Chapter 6 we would investigate the affective profile of human and synthetic text to determine if there are "*measurable differences which may prove useful in the task of synthetic text detection*".

The results of the comparison between the affective profile of human and synthetic text are presented above and in the remainder of this chapter. The results indicate "*measurable differences*" which are statistically significant. While the differences between human and synthetic text may appear small to the human reader, the fact they are measurable and statistically significant suggests they could "*prove useful in the task of synthetic text detection*". For example, a Machine Learning algorithm or model may benefit from the features associated with the affective profile and we refer to this possibility in the concluding chapter (§7.4).

We note two similarities between the results in Chapter 3 (RQ1) regarding synthetic text detection, and the results in Chapter 6 (RQ4) regarding affective profiles of human and synthetic text: Firstly, in RQ1 there is (for the most part) a small difference in the performance of emotionally-aware and standard models in the task of synthetic text detection (e.g. 1.4% and 3.4% for models based on BERT and BLOOM respectively); and in RQ4 there is a small difference in the affective profile of human and synthetic text (e.g 0.5% for mean affective coherence using ψ, which we will see in §6.4.3). Secondly, in RQ1 there is a large difference in the performance of both emotionally-aware and standard models versus chance (e.g. 50% chance versus 85% for BERT and BLOOM based detectors); and in RQ4 there is a large difference in the affective profile of both human and synthetic text versus chance (e.g. 50% chance versus 65% for sentiment n-grams, which we will see in §6.5.2). We cannot conclude that the affective profile (encompassing affective coherence and affective incoherence) is a factor, let alone a decisive factor, in the improved performance of emotionally-aware models over standard models. However, we suggest that the affective profile is a candidate for future work using explainable AI techniques to determine the extent to which it can impact detector performance, and we refer to this in Future Work (§7.4).

6.4 Analysis

Our analysis begins with the results (§6.3) from the emotion 2, 3, 4, and 5-grams (§6.4.1). We then describe further experiments comparing the affective profile of human and synthetic articles (§6.4.2), and measuring the affective coherence of human and synthetic articles using δ-profile, and ψ (§6.4.3). Additionally, we describe experiments comparing sentiment 2, 3, 4, and 5-grams (§6.5).

In our description of Figures 6.2 to 6.4 we treat affective coherence as 'number of emotion changes = 0', and affective incoherence as 'number of emotion changes = n-1' where n is the size of the emotion n-gram i.e. n = 2, 3, 4, and 5. However, for emotion n-grams where $n > 2$, there are more possible values for 'number of emotion changes' i.e. 1, 2, 3 which we have not referenced. Examining these additional values we see that the same three consistent patterns are present. In short, both human and synthetic text exhibit affective coherence (generally the centre to left of Figures 6.2 to 6.4) above random chance; similarly both human and synthetic text exhibit affective incoherence (generally the centre to right of Figures 6.2 to 6.4) below random chance. The same pattern emerges for human text being slightly more affectively coherent than synthetic text, and human text being slightly less affectively incoherent than synthetic text; these differences are statistically significant ($p < .001$).

Calculating random chance for these experiments is complicated, given the complexity of permutations for emotion changes in emotion n-grams, particularly as the 'number of emotion changes' increases in size. To resolve this complexity, a python program was written which generated all of the possible permutations of Plutchik's 8 emotions for 2, 3, 4 and 5 grams. The program then calculated the counts for all the number of emotion changes within each n-gram group, to yield the random chance. The count for each 'number of emotion changes' within each n-gram was checked to ensure it matched the total number of permutations, and that the total random chance for each n-gram equalled 100%.

6.4.1 Affective Coherence (in Emotion N-grams) for Human and Synthetic News Articles

Comparing human text to random chance Human text exhibits affective coherence well above random chance, and affective incoherence well below random chance, for emotion changes in emotion 2, 3, 4, and 5-grams. Human text is orders of magnitude (e.g. more than 300% for emotion 2-grams) above random chance in relation to affective coherence (0 emotion changes), and well below random chance for affective incoherence (n-1 emotion changes). These results support the hypothesis that human text exhibits affective coherence. In addition, the proportion of affective coherence (0 emotion changes) above random chance increases as the length of the n-gram increases for human text; while the proportion of random chance above affective incoherence (n-1 emotion changes) also increases as the length of the n-gram increases for human text.

Comparing synthetic text to random chance Synthetic text exhibits affective coherence well above random chance. Synthetic text exhibits affective incoherence well below random chance. As with human text, these findings are consistent across 2, 3, 4, and 5-grams of emotion.

This result does not support the hypothesis that synthetic text is affectively incoherent; at least not in relation to emotion changes for emotion 2, 3, 4 and 5-grams. This may be because the emotion distribution in the human text used to pretrain Grover (which generated the synthetic text the *NEWSsynth* dataset) is being "parroted" in the generated synthetic text.

Comparing human text to synthetic text Human text exhibits slightly more affective coherence than synthetic text. Human text exhibits slightly less affective incoherence than synthetic text. These results are statistically significant ($p < .001$) and consistent across 2, 3, 4, and 5-grams of emotion. This consistency across n-grams is not guaranteed. For example, we could have three 2-grams `[joy, joy]`, `[sadness, sadness]` and `[fear, fear]` which all have zero emotion changes. However, when combined to make a 5-gram `[joy, joy, sadness, sadness, fear]` there are 2 emotion changes. While the difference between human and synthetic text is slight, in terms of percentage across the whole training dataset of 10k articles, it is statistically significant and may be useful (for example using machine learning techniques) for discriminating human versus synthetic text.

Finally, we note that the disparity between affective coherence and random chance, and affective incoherence and random chance, increases with n-gram length; for human and synthetic text.

6.4.2 Affective Coherence (in Emotion Affective Profiles) for Human and Synthetic News Articles

Having considered the emotion changes within n-grams of emotion from NEWSsynth (train 10k), we now consider the emotion changes within each individual article. We accomplish this by taking the affective profile of an article (i.e. the sequence of emotions within the article) and counting the number of times the emotion changes within the article. However, the mean number of sentences in a human article is 29.33, while it is 19.51 for synthetic articles in the NEWSsynth (train 10k). Therefore, we use percentage values, rather than absolute values, to account for the differing length of articles. The following is an example of calculating the number of emotion changes in an article: An affective profile with 11 emotions can have a maximum of 10 emotion changes; if there are 4 emotion changes then we say the percentage of emotion changes is $(4/10) * 100 = 40\%$. We calculate the percentage of emotion changes in each individual article in NEWSsynth (train, 10k) and compare human and synthetic articles.

	Human	Synthetic
Mean	59.12	60.35
σ	17.65	18.64
Median	61.54	62.50

Table 6.2: Statistics for percent emotion changes for human and synthetic articles in NEWSsynth (train 10k). The human and synthetic distributions are similar for mean, median, and σ (standard deviation), as illustrated in Figure 6.5. Nonetheless, the differences are statistically significant ($p < .001$).

As shown in Table 6.2 and Figure 6.5, the distribution of emotion changes in human and synthetic articles from NEWSsynth (train 10k) are quite similar, nonetheless, the differences are statistically significant ($p < .001$). Data and calculations are provided in Appendix F. The mean 'percent emotion change' in human text is 59.12% with a σ of 17.65%, compared to 60.35% and 18.64% respectively for synthetic text. The median for human text is 61.54% and for

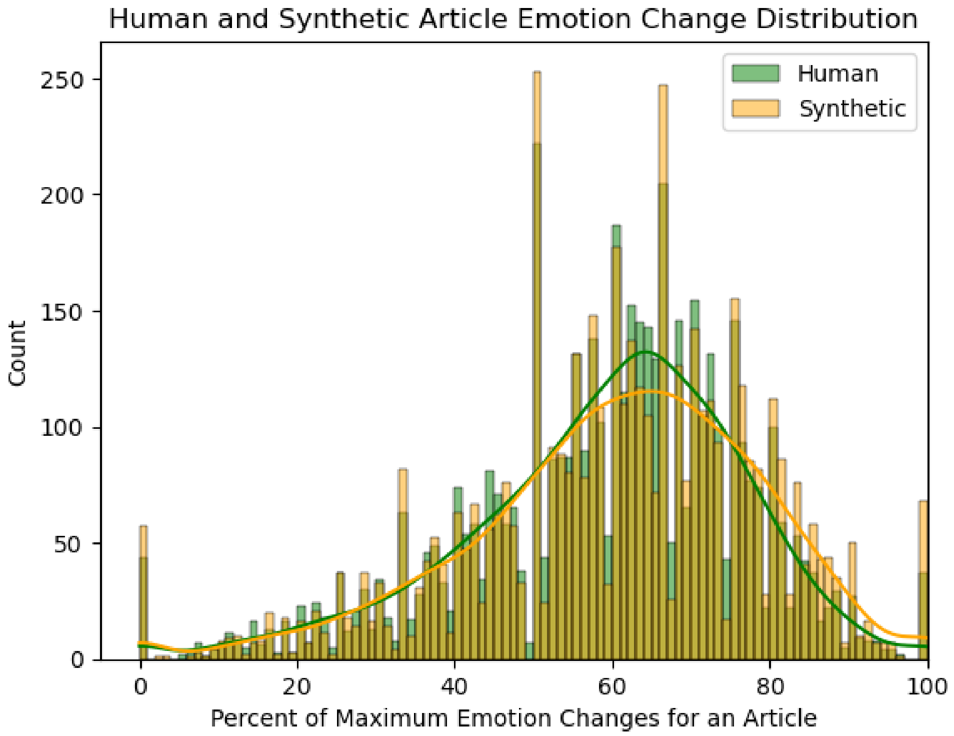


Figure 6.5: Distribution of emotion changes per article for human and synthetic text in NEWSsynth (train 10k). The distributions are similar for human and synthetic text, peaking at just over 60%, as shown in Table 6.2. Nonetheless, the differences are statistically significant ($p < .001$)

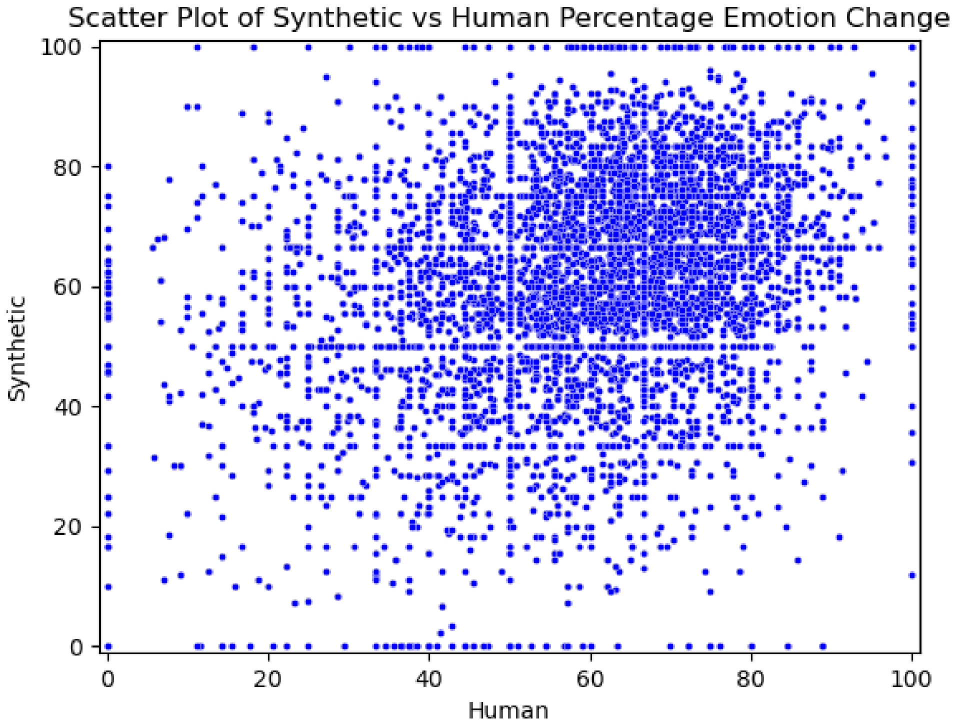


Figure 6.6: Scatter plot of synthetic vs human percentage emotion changes per article - for a human article (used as prompt) and the corresponding article. There is weak correlation, with Pearson $\rho = 25.92$ and Spearman $\rho = 24.84$.

synthetic text it is 62.50%. This similarity is consistent with our previous results (§6.4.1) showing human and synthetic text have similar distributions of emotion change over 2, 3, 4, and 5-grams of emotion, in the NEWSsynth dataset.

Figure 6.6 is a scatter plot comparing the emotion change for a human article, used as prompt to Grover, and the resulting synthetic article from NEWSsynth (train 10k). The Pearson ρ is 25.92 and the Spearman ρ is 24.84; indicating a very weak correlation between the emotion changes in a human article and the emotion changes in the corresponding synthetic article. Therefore, you cannot predict the affective coherence of a synthetic article, even if you know the affective coherence of the human prompt article, when using Grover. The implication is that you cannot control the emotion changes in a synthetic article by selecting emotion changes in the human prompt article. This very weak correlation is indicative of the need for methods to control the affective content of synthetic text as mentioned in §2.2.4. We will see similar results for affective coherence (§6.4.3) and sentiment (§6.5).

6.4.3 Measuring Affective Coherence with ψ and δ-Profile

Thus far, in this chapter, we measured emotion changes in a binary fashion - either the emotion changed or it stayed the same. We now leverage δ, the difference between emotions which we introduced in Chapter 4, to give a more fine-grained description of emotion changes in NEWSsynth (train 10k).

Measuring affective coherence with δ-profile We take the affective profile for an article and measure the δ between successive emotions, thus producing a **δ-profile** for each article. Table 6.3 shows an example affective profile, and the resulting δ-profile. We calculate the δ-profile for all articles in NEWSsynth (train 10k), and compare the distribution of δ values for human and synthetic articles as shown in Figure 6.7. The greater values for affective coherence ($\delta = 0$) and lower values for affective incoherence ($\delta = 4$), in addition to the similarity between human and synthetic distributions, are consistent with our earlier findings on affective coherence - using n-grams (§6.4.1), and affective profile (§6.4.2).

Affective Profile	δ-profile
`[fear, fear, fear, surprise, surprise, fear, fear, fear, anticipation, fear]`	`[0, 0, 1, 0, 1, 0, 0, 3, 3]`

Table 6.3: An example *affective profile* and resulting δ-profile for a human article from the *NEWSsynth* dataset. The δ-profile corresponds to the δ between successive emotions. For example, 'fear'-'fear' in the affective profile results in '0' in the δ-profile; while 'fear'-'anticipation' results in '3' in the δ-profile. See Table 4.1 for the δ between Plutchik's emotions.

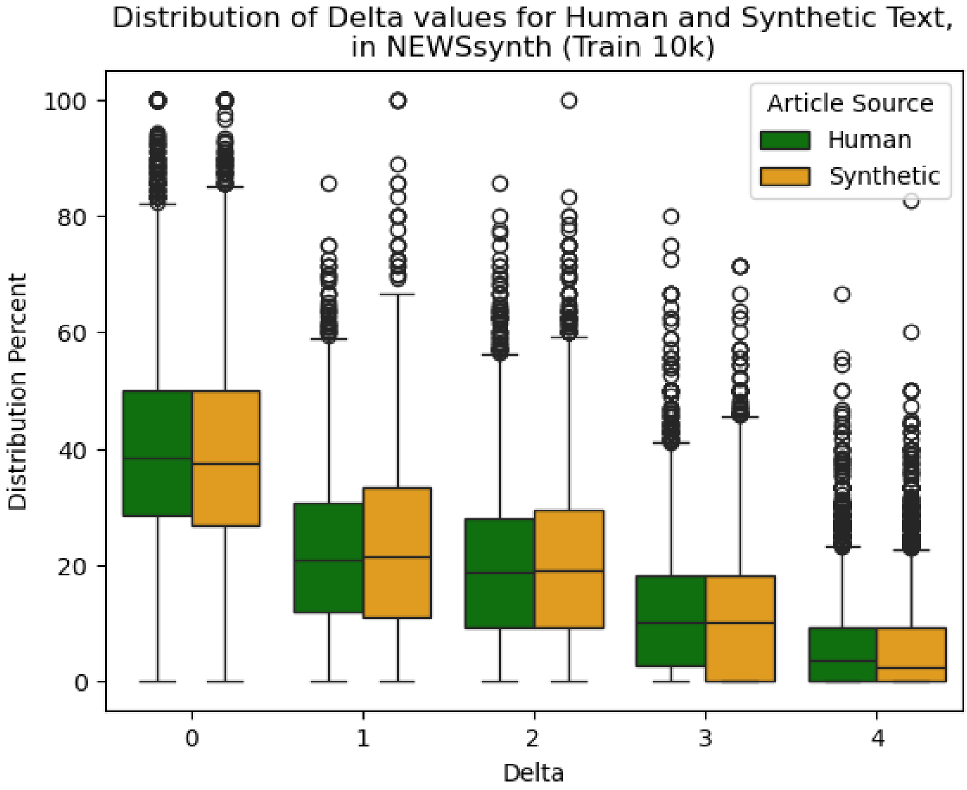


Figure 6.7: Distribution of δ values for human and synthetic articles in NEWSsynth (train 10k). The greater values for affective coherence ($\delta = 0$) and lower values for affective incoherence ($\delta = 4$), in addition to the similarity between human and synthetic distributions, are consistent with our earlier findings on affective profiles.

Measuring affective coherence with ψ Utilising the δ-profile, which we calculated for every article in NEWSsynth (train 10k), in concert with the ψ metric (Equation 6.1) - allows us to calculate a ψ score for each article. But what is the meaning of a ψ score for an article based on its δ-profile? The ψ score for an article is a measure of the affective coherence of the article. For example, if an article has an affective profile with 11 emotions which are all the same, then it will have a δ-profile consisting of ten 0's. If we plug the δ-profile into Equation 6.1 we will get a ψ score of 100%. On the other hand, if the emotions in the article alternate between opposing emotions (say 'joy' and 'sadness') then the affective profile will consist of ten 4's. If we plug the δ-profile into Equation 6.1 we will get a ψ score of 0%. Therefore, a higher ψ score indicates affective coherence, while a lower ψ score indicates affective incoherence.

$$\psi = \frac{100}{N}\left(N - \sum_{\delta=0}^{\delta_{max}} n_\delta \frac{\delta}{\delta_{max}}\right) \tag{6.1}$$

Equation 6.1 shows the ψ metric, where:

N	Total number of examples being tested
δ	Distance between two emotions (0,1,2,3,4) per Table 4.1.
δ_{max}	Maximum distance between two emotions[2] i.e. $\delta_{max} = \lfloor \frac{n_{emo}}{2} \rfloor$
n_{emo}	Total number of emotions (8 for Plutchik)
n_δ	Number of occurrences of a particular δ value (0..N)

ψ	Strength of Affective Coherence
$\psi < 30$	None or very weak
$30 < \psi < 50$	Weak
$50 < \psi < 70$	Moderate
$\psi > 70$	Strong

Table 6.4: Table showing ψ and the strength of affective coherence.

Therefore, we can use ψ to measure and compare the affective coherence of human and synthetic articles in NEWSsynth (train 10k). Furthermore, we

---

[2] For Plutchik's 8 emotions: $\delta_{max} = \lfloor \frac{n_{emo}}{2} \rfloor = \lfloor \frac{8}{2} \rfloor = 4$

can ‘bucket’ ψ scores, similar to Table 2.4 for correlations, to label the affective coherence as ‘very weak’, ‘weak’, ‘moderate’, or ‘strong’; as shown in Table 6.4.

We introduced ψ as a metric for measuring and evaluating emotion classifiers, but we also introduced ψ as a flexible and extensible framework. This ‘overloading’ of ψ, to measure affective coherence, is an example of the flexibility and extensibility of ψ.

Table 6.5 and Figure 6.8 show the distribution of ψ scores for human and synthetic text in NEWSsynth (train 10k). The distributions are very similar, with human text showing slightly higher affective coherence than synthetic text. The mean and σ (standard deviation) are almost identical for human and synthetic text: a mean of 70.05% & 69.66% and σ of 12.00 & 12.23, for human and synthetic text respectively. Similarly, the median values are 69.64% and 69.35% for human and synthetic text are almost identical. This is consistent with our previous results (§6.4.2) whereby human and synthetic text have similar distributions.

	Human	Synthetic
Mean	70.05	69.66
σ	12.00	12.23
Median	69.64	69.35

Table 6.5: Statistics for ψ (affective coherence) for human and synthetic articles in NEWSsynth (train 10k). The human and synthetic distributions are similar for mean, median, and σ (standard deviation), as illustrated in Figure 6.8.

Figure 6.9 is a scatter plot comparing the ψ score (affective coherence) for a human article, used as prompt to Grover, and the resulting synthetic article from NEWSsynth (train 10k). The Pearson ρ is 31.19 and the Spearman ρ is 31.65; indicating a weak (borderline very weak) correlation between the ψ score in a human article and the ψ score in the corresponding synthetic article. Therefore, as we saw earlier with the distribution of emotion changes (§6.4.2), you cannot predict (or control) the affective coherence (ψ score or δ-profile) of a synthetic article - even if you know the affective coherence (ψ score or

δ-profile) of the human prompt article, when using Grover. This borderline weak correlation is consistent with our earlier result for distribution of emotion changes (§6.4.2), and we will see similar for sentiment changes (§6.5.3) in the next section.

6.5 Sentiment Experiments

Thus far, in this chapter, we generated the affective profile for an article from Plutchik's 8 emotions. However, by using sentiment we can generate a more coarse view of the affective profile of an article, and this coarse view may elucidate some features of the affective profile which are potentially diluted by the larger number of emotions used by Plutchik. We use binary sentiment (positive, negative), rather than ternary sentiment (positive, negative, neutral), because it is the topmost division of positive and negative valence as described by Ortony et al. (1988), and is, therefore, the most coarse affective profile possible.

We evaluate two candidate sentiment classifiers (§6.5.1) and choose the best performing model to generate binary sentiment affective profiles (similar to Chapter 5 for emotion) for NEWSsynth (train 10k). We compare the affective profiles of human and synthetic articles for sentiment changes in §6.5.3 (similar to §6.4.2 for emotion); and sentiment n-grams in §6.5.2 (similar to §6.4.1 for emotion).

6.5.1 Selecting a Sentiment Classifier

Creating an affective profile for sentiment requires a suitable sentiment classifier. We previously used DistilBERT$_{\text{base}}$ uncased (Sanh et al., 2020) finetuned on the SST-2 dataset (Socher et al., 2013) for sentiment analysis (§3.4.3). Subsequently, Hartmann et al. (2023) published and released SiEBERT - a RoBERTa-large model fine-tuned and evaluated on 15 diverse sentiment datasets. Hartmann et al. (2023) reported superior performance from SiEBERT, compared

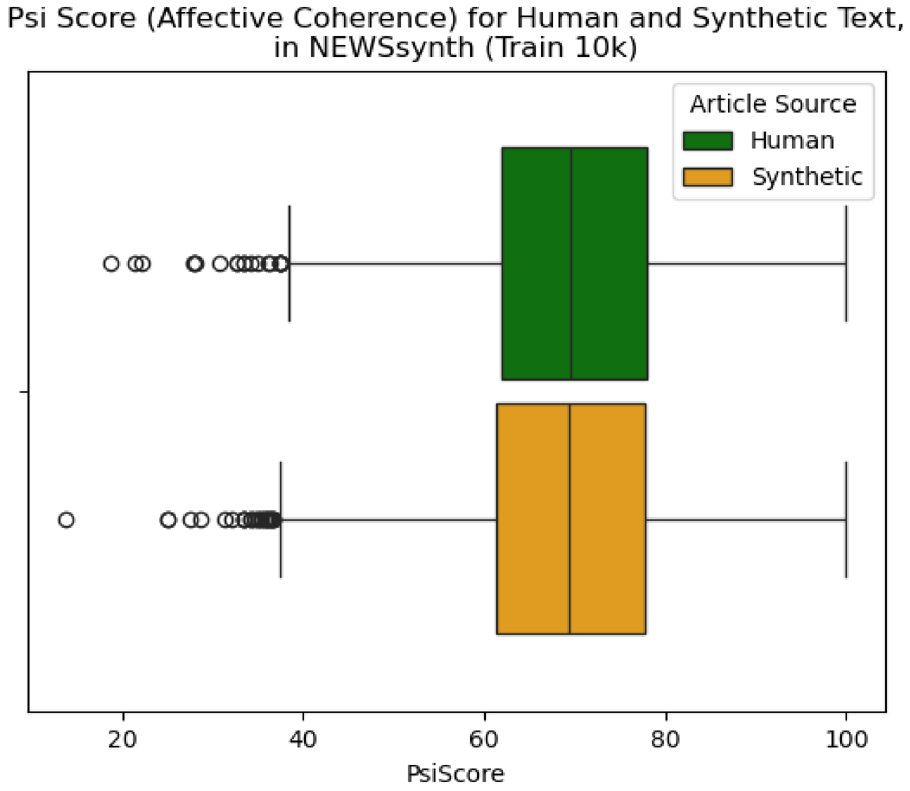


Figure 6.8: Distribution of ψ score per article for human and synthetic text in NEWSsynth (train 10k). The distributions are similar for human and synthetic text, with mean and median around 70%, and σ around 12, as in Table 6.5.

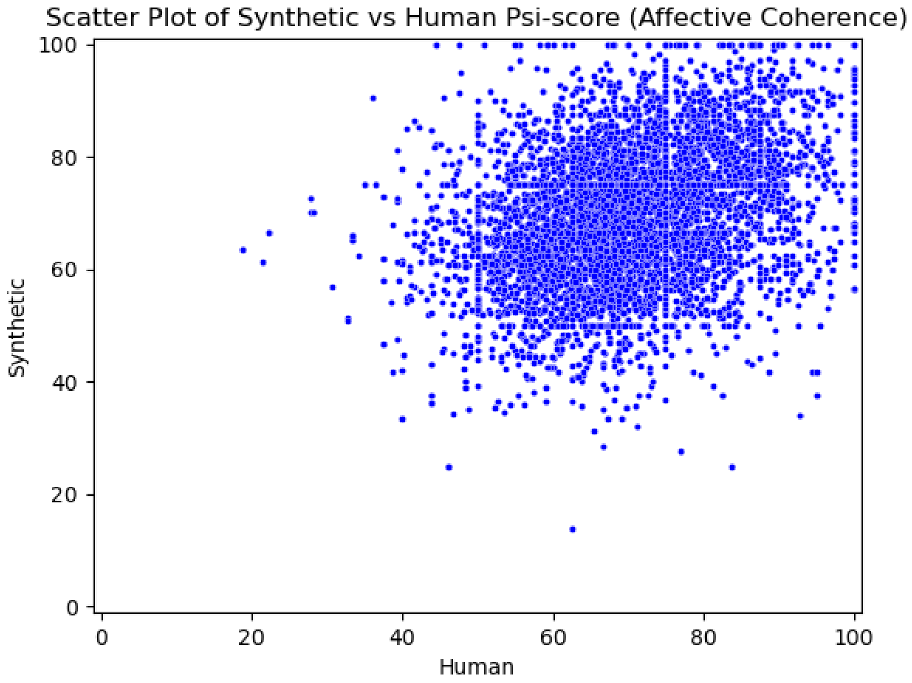


Figure 6.9: Scatter plot of synthetic vs human ψ score per article - for a human article (used as prompt) and the corresponding article. There is weak correlation, with Pearson $\rho = 31.19$ and Spearman $\rho = 31.65$.

to DistilBERT$_{\text{base}}$, on all 15 datasets with an average accuracy of 78.1% for DistilBERT$_{\text{base}}$, and an average accuracy of 93.2% for SiEBERT.

We choose two familiar datasets in the news domain on which to evaluate the two sentiment classifiers. We evaluate SiEBERT and DistilBERT$_{\text{base}}$ models on the GNE and AT datasets which were mapped to sentiment, as shown in Table 6.6. After mapping GNE to sentiment there was a label imbalance - 1418 positive and 3582 negative examples. The labels were balanced, by randomly removing excess negative examples, to give 1418 positive and 1418 negative examples. Similarly, AT was imbalanced after mapping to sentiment, with 441 positive and 592 negative examples. Random negative examples were removed to give a balanced dataset with 441 positive and 441 negative examples.

As shown in Table 6.6 'surprise' was removed from the AT dataset before mapping. In order to determine if 'surprise' is positive or negative sentiment, we previously used DistilBERT$_{\text{base}}$ to classify examples from AT labelled 'surprise'. If we used DistilBERT$_{\text{base}}$ to predict sentiment for 'surprise', and then test DistilBERT$_{\text{base}}$ against these predictions - we would introduce bias in favour of DistilBERT$_{\text{base}}$. Therefore, we remove the 'surprise' examples from AT before mapping to sentiment. We believe this removal is acceptable for several reasons (in addition to the removal of possible bias). First, removing 'surprise' impacts both positive and negative sentiment, and actually *reduces* the imbalance in AT. We saw above that AT without 'surprise' is imbalanced when mapped to sentiment (441-positive, 592 negative), and we saw in §3.4.3 that 85.95% of 'surprise' mapped to negative sentiment. Therefore, including 'surprise' would result in a greater imbalance, and necessitate the removal of additional negative examples. Second, we are evaluating on sentiment, which is coarser than individual emotions, so the removal of a specific emotion (particularly one that maps to an over-represented label) is not as critical. Thirdly, 'surprise' is still represented in GNE and the sentiment classifiers are evaluated on both datasets. Finally, we see that both datasets give similar results for the sentiment classifiers.

Dataset	Emotion		Sentiment	Count
GNE	trust, love/like, joy, pride, positive anticipation / optimism, positive surprise	→	positive	1418
	disgust, fear, negative anticipation / pessimism, sadness, guilt, shame, anger, annoyance, negative surprise	→	negative	1418
AT-s	joy	→	positive	441
	sadness, fear, anger, disgust	→	negative	441
	(surprise)*			0

Table 6.6: Sentiment Mapping Schema: GoodNewsEveryone (15 emotions), and Affective text (6 emotions) mapped to binary sentiment. Count shows the sentiment label distribution in the dataset after balancing. * Surprise was removed from AT because we previously used DistilBERT-base to map surprise to sentiment, so testing DistilBERT-base on 'surprise' from AT would bias the results in favour of DistilBERT-base.

Table 6.7 shows the results of evaluating $\text{DistilBERT}_{\text{base}}$ and SiEBERT on two datasets mapped to sentiment i.e GNE and AT-s (Affective Text minus 'surprise'). Accuracy, True Positive Rate (TPR) and True Negative Rate (TNR) are reported. SiEBERT has a lower TNR but outperforms $\text{DistilBERT}_{\text{base}}$ in Accuracy and TPR. Due to its superior performance in our evaluation, we choose SiEBERT as the sentiment classifier for generating affective profiles on the NEWSsynth (train 10k) dataset.

Dataset	Model	Accuracy	TPR	TNR
GNE	$\text{DistilBERT}_{\text{base}}$	70.52	59.45	**81.59**
	SiEBERT	**76.97**	**72.99**	80.96
AT-s	$\text{DistilBERT}_{\text{base}}$	79.59	72.79	**86.39**
	SiEBERT	**87.19**	**91.38**	82.99

Table 6.7: Results for sentiment classifiers tested against the (mapped to sentiment) GNE and AT-s (Affective Text minus 'surprise') datasets. TPR is True Positive Rate (aka Sensitivity, Recall), and TNR is True Negative Rate (aka Specificity). TPR, TNR, and Accuracy are reported. SiEBERT has a lower TNR, but outperforms $\text{DistilBERT}_{\text{base}}$ for Accuracy and TPR.

6.5.2 Affective Coherence (in Sentiment N-grams) for Human and Synthetic News Articles

We reuse our methodology from §6.1, but modify it to use n-grams of sentiment rather than n-grams of emotion. As before, we generate 2, 3, 4, and 5-grams. The distribution of sentiment changes within sentiment n-grams are shown in Figures 6.10 to 6.13. The pattern of results we saw for emotion n-grams (§6.4.1) are essentially identical to the results we see for sentiment n-grams. The consistent pattern of results for sentiment 2, 3, 4, and 5-grams are:

1. Human text displays affective coherence well above chance, and affective incoherence well below chance.
2. Synthetic text displays affective coherence and affective incoherence similar to human text; however, the difference is statistically significant ($p < .001$). Data and calculations are provided in Appendix F.
3. Human text displays slightly more affective coherence, and slightly less affective incoherence, than synthetic text, and the difference is statistically significant ($p < .001$).
4. The disparity between affective coherence and random chance, and affective incoherence and random chance, increases with n-gram length; for human and synthetic text.

6.5.3 Affective Coherence (in Sentiment Affective Profiles) for Human and Synthetic News Articles

We reuse our methodology from §6.4.2, but modify it to use sentiment rather than emotion. SiEBERT predicted the binary sentiment (positive, negative) for each sentence in each news article of the NEWSsynth (train 10k). The 10k articles are distributed equally between human and synthetic i.e. 5k articles each. The sequence of sentiments in each article represents that articles *affective profile*. A sample affective profile for a human article comprising 10 sentences,

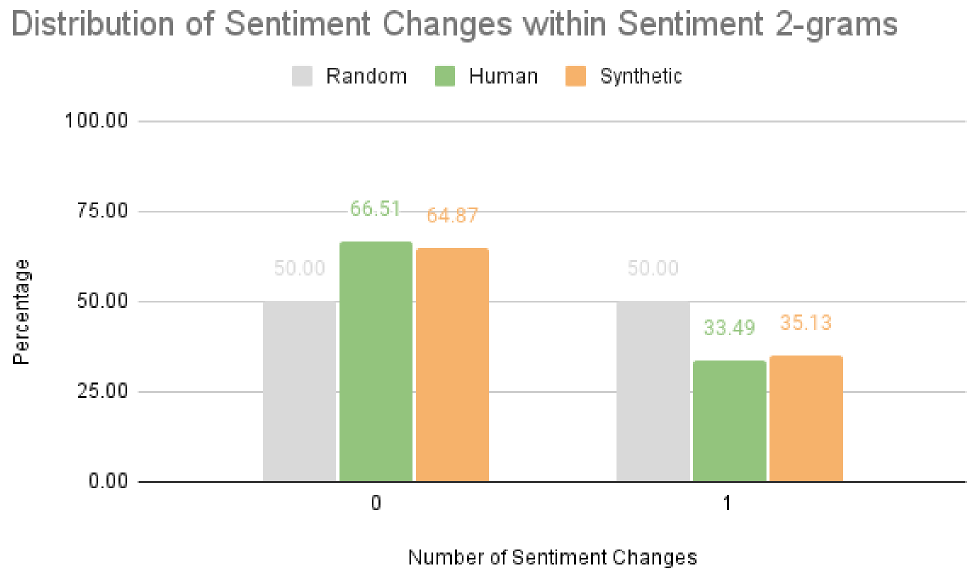


Figure 6.10: The distribution of sentiment changes for sentiment 2-grams. Human text exhibits affective coherence (sentiment changes = 0) well above chance (66.51% vs 50.00%). Human text exhibits affective incoherence (sentiment changes = 1) well below chance (33.49% vs 50.00%). The results indicate human text is affectively coherent. However, synthetic text exhibits similar affective coherence to human text. Nonetheless, human text is more affectively coherent than synthetic text, and human text is less affectively incoherent than synthetic text, and this difference is statistically significant ($p < .001$).

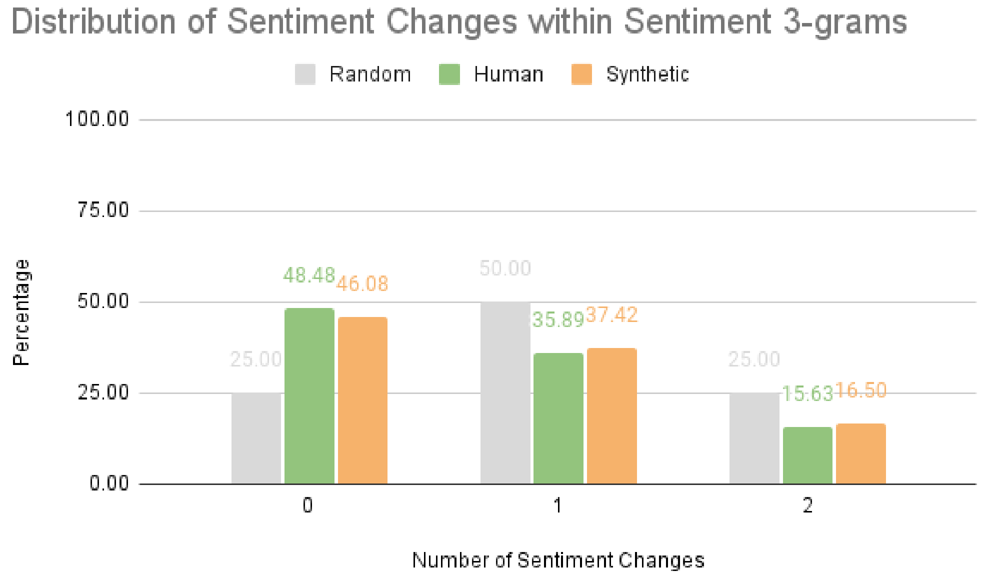


Figure 6.11: The distribution of sentiment changes for sentiment 3-grams. Human text exhibits affective coherence (sentiment changes = 0) well above chance (48.48% vs 25.00%). Human text exhibits affective incoherence (sentiment changes = 2) well below chance (15.63% vs 25.00%). The human and synthetic distribution difference is significant ($p < .001$). The results are consistent with 2-grams in Figure 6.10.

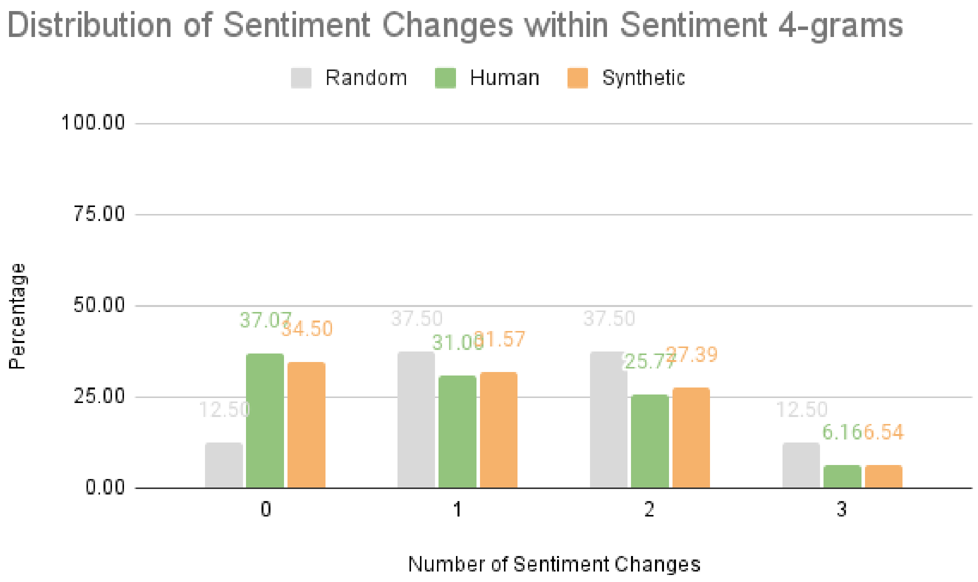


Figure 6.12: The distribution of sentiment changes for sentiment 4-grams. Human text exhibits affective coherence (sentiment changes = 0) well above chance (37.07% vs 12.50%). Human text exhibits affective incoherence (sentiment changes = 3) well below chance (6.16% vs 12.50%). The human and synthetic distribution difference is significant ($p < .001$). The results are consistent with 2-grams and 3-grams in Figures 6.10 and 6.11.

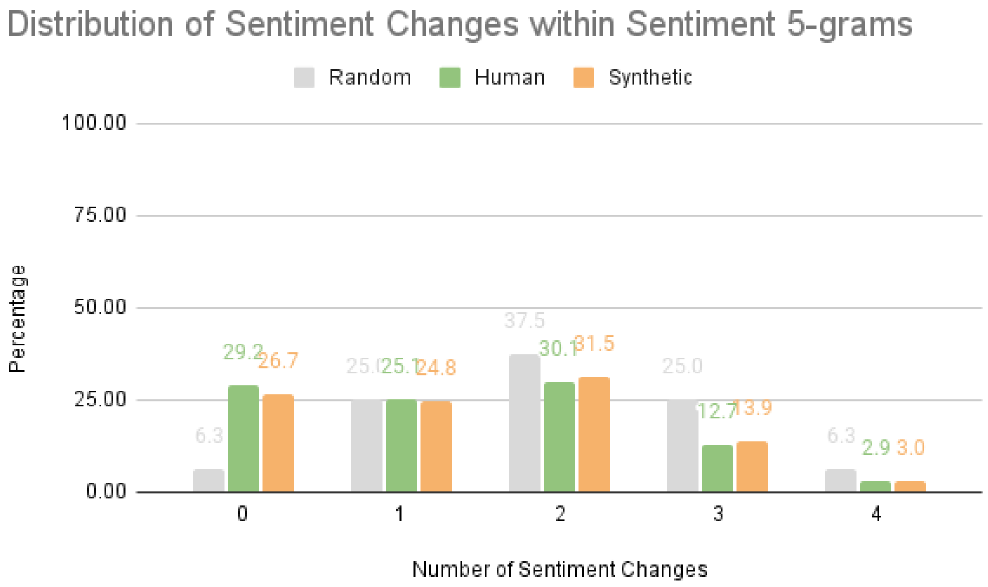


Figure 6.13: The distribution of sentiment changes for sentiment 5-grams. Human text exhibits affective coherence (sentiment changes = 0) well above chance (29.19% vs 6.25%). Human text exhibits affective incoherence (sentiment changes = 4) well below chance (2.90% vs 6.25%). The human and synthetic distribution difference is significant ($p < .001$). The results are consistent with 2-grams, 3-grams, and 4-grams in Figures 6.10 to 6.12.

where `'P'` is positive sentiment and `'N'` is negative sentiment is shown in Table 6.8, along with the affective profile of the corresponding synthetic article generated by Grover comprising 11 sentences. Both articles are from *NEWSsynth* (train 10k) dataset described in §3.2.2. As before, we use the affective profile to calculate the 'percent sentiment changes', where previously we calculated 'percent emotion changes' (§6.4.2).

Source	Affective Profile (binary sentiment)
Human	`[ N, N, N, P, N, P, P, P, P, N ]`
Synthetic	`[ N, N, P, N, P, N, P, N, P, N, N ]`

Table 6.8: An example *affective profile* for binary sentiment of a human article and the corresponding synthetic article. The synthetic article was generated by Grover (using the human article as prompt). Both articles are taken from the *NEWSsynth* dataset. 'N' is negative sentiment, 'P' is positive sentiment.

The results for 'percent sentiment change' are consistent with our previous results for 'percent emotion change' (§6.4.2). As shown in Table 6.9 and Figure 6.14, the distribution of sentiment changes in NEWSsynth (train 10k) are quite similar for human and synthetic articles, nonetheless, the differences are statistically significant ($p < .001$). Data and calculations are provided in Appendix F. The mean 'percent sentiment change' in human text is 32.12% with a σ of 16.40%, compared to 34.08% and 17.44% respectively for synthetic text. The median for human text is 33.33% and for synthetic text it is 35.29%.

	Human	Synthetic
Mean	32.12	34.08
σ	16.40	17.44
Median	33.33	35.29

Table 6.9: Statistics for percent sentiment changes for human and synthetic articles in NEWSsynth (train 10k). The human and synthetic distributions are similar for mean, median, and σ (standard deviation), as illustrated in Figure 6.14. Nonetheless, the differences are statistically significant ($p < .001$).

Figure 6.15 is a scatter plot comparing the 'percent sentiment change' for a human article, used as prompt to Grover, and the resulting synthetic article from NEWSsynth (train 10k). The Pearson ρ is 28.10; indicating a very weak correlation between the percent sentiment changes in a human article and the percent sentiment changes in the corresponding synthetic article. Therefore, as we saw earlier with both the distribution of 'percent emotion changes' (§6.4.2) and ψ-score (§6.4.3), you cannot predict (or control) the affective coherence of a synthetic article - even if you know the affective coherence of the human prompt article, when using Grover.

The very weak correlation does not impact the use-case in this thesis because we classify a single article in isolation, not a pair of human and synthetic articles. However, it may impact Turing Test (Turing, 1950) style classification tasks where a pair of articles (human prompt article and corresponding synthetic article) are to be classified - one human and one synthetic. The very weak correlation suggests that knowing the length of the human article, or the 'percent emotion changes', or the ψ-score (for affective coherence), or the 'percent sentiment changes', does not give an indication of the equivalent value in the synthetic article. If, however, it had been the case that there was a 'strong' correlation, then it may be possible to infer which article is more likely to be human and which is more likely to be synthetic. For example, if synthetic articles tended to be twice the length of the human prompt article then, given two articles, one could infer that the longer article is more likely to be synthetic. However, the very weak correlation across all experiments suggests no such inference can be made. Future work could investigate the correlation between combinations of such features, for example combining article length and ψ-score, in human prompt articles and the corresponding synthetic article. The very weak correlation is indicative of the need for methods to control the affective content of synthetic text, as previously mentioned in §2.2.4, §6.4.2, and §6.4.3, and surveyed in relation to sentiment by Lorandi and Belz (2023).

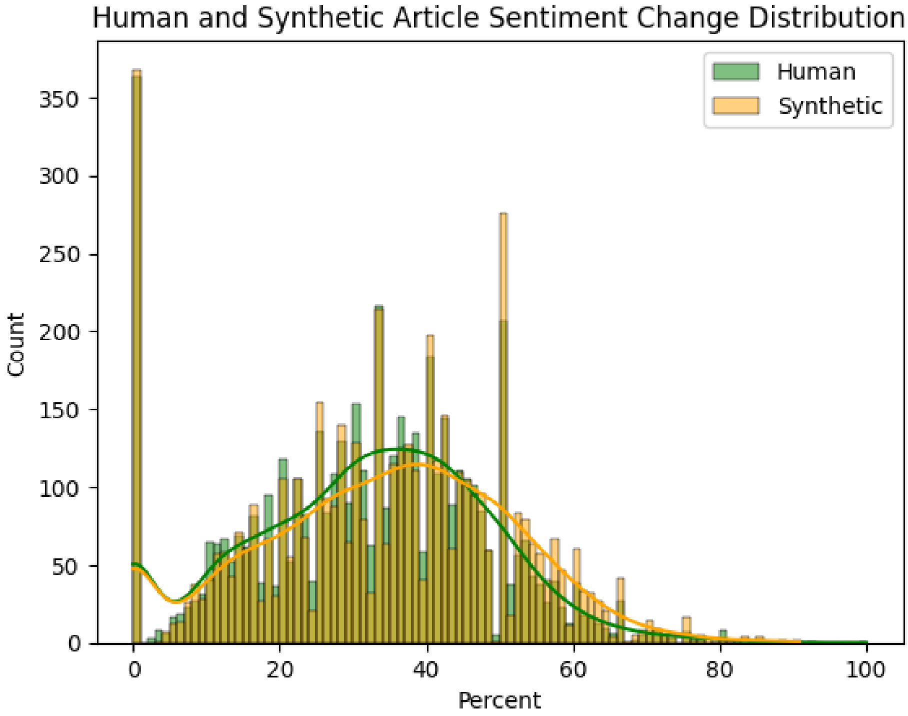


Figure 6.14: Distribution of 'percent sentiment change' per article for human and synthetic text in NEWSsynth (train 10k). The distributions are similar for human and synthetic text, with mean and median around 34%, and σ around 17%, as in Table 6.9. Nonetheless, the differences are statistically significant ($p < .001$).

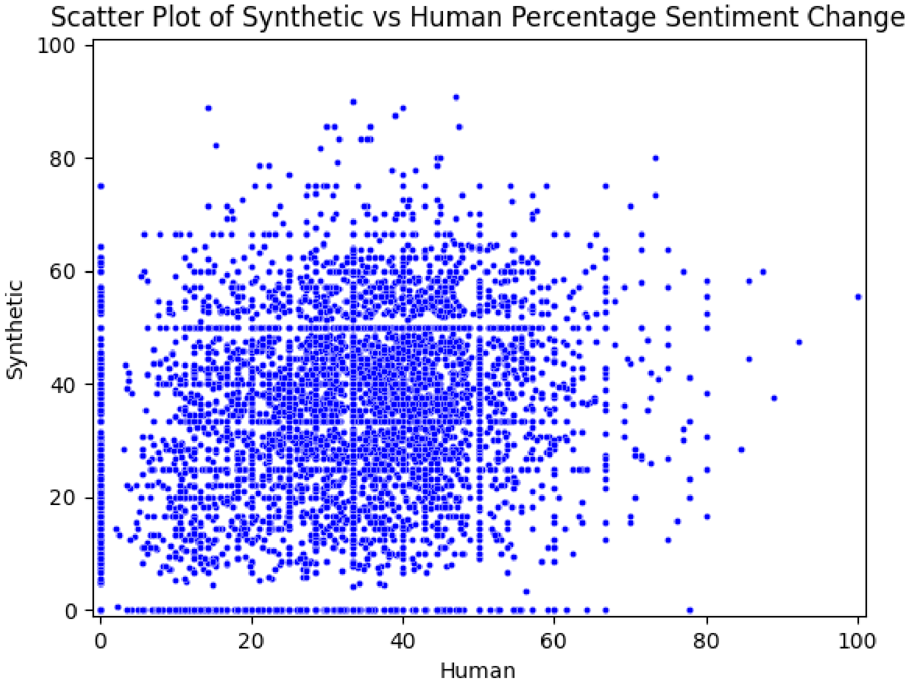


Figure 6.15: Scatter plot of synthetic vs human 'percent sentiment change' per article - for a human article (used as prompt) and the corresponding synthetic article. There is very weak correlation, with Pearson $\rho = 28.10$.

6.6 Conclusion

In this chapter we answer Research Question 4 in the affirmative: "*Can we generate and compare the affective profile of human text and synthetic text?*". We generated affective profiles for human and synthetic articles, in our NEWSsynth (train 10k) dataset from Chapter 3, using our fine-tuned PLM (RoBERTa-large) for emotion classification from Chapter 5. We compared the affective profiles of human and synthetic articles by analysing their affective coherence using several methods.

We first used the affective profiles to generate emotion n-grams for NEWSsynth (train 10k), and compared human versus synthetic text, which yielded a pattern of results that we refer to as PR1 - as shown in Table 6.10. We then used the affective profile to calculate 'percent emotion change' for each article in NEWSsynth (train 10k), and compared human versus synthetic text, which yielded a pattern of results that we refer to as PR2 - as shown in Table 6.10. We generated a δ-profile to calculate the δ-distribution for human and synthetic articles, yielding PR1 results. We then used the δ-profile to calculate the ψ score for human and synthetic articles, yielding PR2 results.

Then we turned to sentiment, and evaluated two sentiment classifiers (DistilBERT$_{\text{base}}$ and SiEBERT). We selected SiEBERT based on its superior performance in our evaluation. We generated affective profiles (sentiment) using SiEBERT for NEWSsynth (train 10k). Taking a similar approach to sentiment, as we did with emotion, we used the affective profiles to generate sentiment n-grams over the NEWSsynth (train 10k) dataset. We compared human versus synthetic text, which yielded PR1 results. Finally, we used the affective profile to calculate 'percent sentiment change' for each article in NEWSsynth (train 10k), and compared human versus synthetic text, which yielded PR2 results. Statisitcal significance testing was conducted using chi-squared (χ^2) for PR1 experiments, and using Kolmogorov-Smirnov for PR2 experiments. Data and calculations are provided in Appendix F.

Table 6.11 shows each experiment in RQ4 and its corresponding pattern of results (PR1 or PR2). Each pattern of results (PR1 and PR2) is produced in three separate experiments, covering different classifiers (RoBERTa-large, SiEBERT), emotion models (Plutchik, sentiment), and experimental methodologies (n-grams, affective profile, δ-profile, ψ metric).

With Research Questions 1-4 answered, next we present our conclusions.

Name	Pattern of Results
PR1	1. Human text displays affective coherence well above chance, and affective incoherence well below chance. 2. Synthetic text displays affective coherence and affective incoherence similar to human text. For both emotion and sentiment n-grams the difference is statistically significant ($p < .001$). 3. Human text displays slightly more affective coherence, and slightly less affective incoherence, than synthetic text. For both emotion and sentiment n-grams the difference is statistically significant ($p < .001$). 4. The disparity between affective coherence and random chance, and affective incoherence and random chance, increases with n-gram length; for human and synthetic text.
PR2	1. Central tendency and spread are very similar for the distribution of changes of emotion or sentiment in human and synthetic articles from NEWSsynth (train 10k). For both emotion and sentiment n-grams the difference is statistically significant ($p < .001$). 2. Specifically, the mean, median, and σ are very similar for human and synthetic text. For both emotion and sentiment n-grams the difference is statistically significant ($p < .001$). 3. There is very weak (or borderline very weak) correlation between the emotion or sentiment in a human article and the corresponding synthetic article.

Table 6.10: Two patterns of results arise consistently across all RQ4 experiments. We refer to these patterns of results as PR1 and PR2.

§	Experiment	Results
6.4.1	Affective Coherence (in emotion n-grams) for Human and Synthetic News Articles	PR1
6.4.2	Affective Coherence (in emotion affective profiles) for Human and Synthetic News Articles	PR2
6.4.3-δ	Measuring Affective Coherence with ψ, and δ-profile	PR1
6.4.3-ψ	Measuring Affective Coherence with ψ, and δ-profile	PR2
6.5.2	Affective Coherence (in sentiment n-grams) for Human and Synthetic News Articles	PR1
6.5.3	Affective Coherence (in sentiment affective profiles) for Human and Synthetic News Articles	PR2

Table 6.11: Table showing our RQ4 experiments and their corresponding pattern or results. Each pattern of results (PR1 and PR2) is produced in three separate experiments, covering different classifiers, emotion models (Plutchik, sentiment), and experimental methodologies (n-grams, affective profile, δ-profile, ψ metric).

Chapter 7

Conclusion

In this concluding chapter, we summarise the contributions of the research (§7.1), and review the key findings and insights (§7.2), while also addressing limitations of the research (§7.3) and highlight potential directions for future work (§7.4).

7.1 Contributions

The contributions of this research are as follows:

1. We demonstrated that **emotion can help in the task of synthetic text detection**. We did so using multiple PLMs of different sizes and architectures - for both generation of text and detection of text. We conducted experiments primarily in the news domain, but also other domains. We used multiple datasets for testing (NEWSsynth, ChatGPT100, RealNews-Test), and compared our results to other authors reported results.

2. We created and released ***NEWSsynth*** - a dataset of 20k human and synthetic (Grover) news articles. The hyper-parameters used for generating synthetic text with Grover are provided in Appendix A.

3. We created and released ***ChatGPT100*** - a dataset of 100 human and synthetic articles (ChatGPT) in various domains, similar to Clark et al. (2021),

using an LLM for synthetic text. Using ChatGPT100 we demonstrated that our "emotionally-aware" model outperformed the standard model in synthetic text detection (§3.5).

4. We released **multiple PLMs fine-tuned on emotion classification** for Plutchik's 8 emotions, and Ekman's 6 emotions. The absence of suitable emotion classifiers necessitated the creation of our own classifiers (and custom datasets in many cases). We release the emotion classifiers spanning multiple PLMs, and fine-tuned on multiple datasets.

5. We released **multiple datasets customised with emotion mappings** to Plutchik's 8 emotions, Ekman's 6 emotions, and binary sentiment. Once again the absence of suitable resources necessitated the creation of our own resources. We combined existing emotion datasets and created mapping schemata as required, in some cases we also balanced the dataset across emotions.

6. We introduced **$\boldsymbol{\psi}$ - a novel flexible framework for evaluating classifiers** (of dimensional emotions, and more). Inspired by the dimensional relationship between emotions in Plutchik's Wheel of Emotions we introduced **$\boldsymbol{\delta}$ - a measure of emotion similarity**. By introducing $\boldsymbol{\delta}$ to Plutchik's Wheel of Emotions we elevated it from an `Ordinal` scale to an `Interval` scale which, crucially, allowed us to measure difference between emotions. We introduced **$\boldsymbol{\phi(\delta)}$ - a loss function** which is flexible and extensible and can be specified to suit user requirements. We added some terms to the $\boldsymbol{\phi(\delta)}$ loss function to turn it into a percentage score, yielding the final equation for **$\boldsymbol{\psi}$ - a novel flexible framework for evaluating classifiers**. We used $\boldsymbol{\psi}$ to evaluate and select emotion classifiers (RQ3). We demonstrated the flexibility and extensibility of $\boldsymbol{\psi}$ when we used it to measure and numerically quantify *affective coherence* in human and synthetic text (RQ4).

7. We presented an **analysis of *affective profiles* in human and synthetic text news articles**. We analysed the affective profile of the NEWSsynth training set of 10k human and synthetic articles. We did so using n-grams of emotion, and whole articles using different analytical techniques (i.e. 'percentage change of emotion', and $\boldsymbol{\delta}$-profiles). We conducted the same experiments using binary sentiment. The results across all experiments showed two consistent patterns of results emerging - for human and synthetic text - which differ greatly from random chance. The results showed that human and synthetic text were similar, but slightly different, across all experiments. Nonetheless, for both emotion and sentiment the difference between human and synthetic text is statistically significant ($\boldsymbol{p < .001}$).

We believe our contributions represent a valuable resource for researchers, and those in industry, who wish to work with human and synthetic text, emotion classification, and those who will find use beyond these applications.

7.2 Research Questions Revisited

We revisit each of our four research questions in turn, answering each research question in the affirmative.

7.2.1 RQ1 - Can the performance of automatic synthetic text detectors be improved by making them 'emotionally-aware' compared to standard detectors?

We demonstrated that the performance of automatic synthetic text detectors can be improved by making them 'emotionally-aware' compared to standard detectors. We did so by creating an 'emotionally-aware' synthetic text detector (by first fine-tuning it on emotion classification) and subsequently comparing it against a standard synthetic text detector. We confirmed these results using different size models (for synthetic text generation, and detection), different architectures, and different datasets. We conducted an experiment using randomised emotion labels to demonstrate the benefit of the emotion classification task (as opposed to the benefit of additional fine-tuning) in the subsequent task of synthetic text detection. However, the experiment also showed that the additional fine-tuning with randomised emotion labels showed some improvement over the baseline. We conducted an ablation study using multiple emotion datasets in isolation and combination to determine which datasets gave the best results in the subsequent task of synthetic text detection; we use these results in this experiment, and subsequently, to inform our choice of datasets in RQ3 when fine-tuning emotion classifiers. We conducted additional experiments on the length of human and synthetic articles (words per article, words per sentence, sentence per article); this enabled us to exclude length (in its various forms) as a potential proxy for synthetic text detection. We demonstrated weak correlation between the length of a human article and the length of the corresponding synthetic article; weak correlation between a human prompt article and the resulting synthetic article is also observed in RQ4 in relation to emotion.

7.2.2 RQ2 - Is it possible to develop a framework for measuring and evaluating classifiers of Plutchik's 8 emotions which takes account of their dimensional relationship?

We developed a framework for measuring and evaluating classifiers of Plutchik's 8 emotions which takes account of their dimensional relationship. We refer to this framework as the ψ metric, or simply ψ. Leveraging the dimensional relationship between Plutchik's 8 emotions, we introduced δ - a measure of emotion similarity. We defined a loss function $\phi(\delta)$ which, with some additional terms, we used to define an equation for ψ. We described the flexibility and extensibility of ψ through examples of different loss functions, $\phi(\delta)$, such as linear, quadratic, logistic, step etc. We compared ψ against Accuracy, MAE, and RMSE highlighting its advantages over each of them - such as its fine-grained measurement (rather than using binary correct or incorrect), and ψ scores are always (regardless of, for example, the number of emotions) in the familiar range 0..100%.

We saw a real-world application of ψ, in RQ3, when we used it to evaluate and select an emotion classifier. Then, in RQ4, we demonstrated the flexibility and extensibility of ψ when measuring affective coherence. We describe the aspects of those experiments relating to ψ here, rather than in relation to RQ3 and RQ4 respectively.

Using ψ for Emotion Classification Having used ψ in the evaluation and selection of an emotion classifier in RQ3, we can make the following conclusions:

- The results in the RQ3 experiments demonstrate that ψ is a useful metric.
- ψ gives a more nuanced result (than binary in/correct metrics) and gives a more fine-grained measure of classifier performance.
- ψ tends to give a different score than $F1_\mu$ and can rank models differently than $F1_\mu$. Therefore, ψ can impact model selection and potentially benefit downstream tasks. We saw an example of this with the fine-tuning test results in §5.3.
- Notwithstanding the last point, ψ correlates (Pearson and Spearman) strongly with existing metrics (Accuracy, MAE, RMSE) - demonstrating it is not an outlier.
- ψ appears to be less volatile than $F1_\mu$, having lower standard deviation, σ.

Using ψ (and δ-profile) for measuring Affective Coherence Having used both ψ and δ-profile for measuring affective coherence in RQ4, we can make the following conclusions:

- The results in the RQ4 experiments demonstrate that ψ is a useful metric, as is the δ-profile, because they provided us with novel ways to accomplish our task i.e. measure *affective coherence.*
- The application of both ψ and δ-profile to measuring *affective coherence* was a novel application which was not anticipated when ψ was developed; thereby supporting the contention that ψ is flexible and extensible - beyond our initial use-cases.
- The fact that the pattern of results for both ψ and δ-profile were consistent with all the other experiments in RQ4, reinforces our conclusion above from RQ3 i.e. it demonstrates that ψ, while novel, is not an outlier.

7.2.3 RQ3 - Can we exploit the dimensional relationship between Plutchik's 8 emotions to enable optimal selection of an automatic emotion classifier?

We exploited the dimensional relationship between Plutchik's 8 emotions to enable optimal selection of an automatic emotion classifier. We created a custom dataset by combining the two best performing emotion datasets from our ablation study in RQ1, mapping them to Plutchik's 8 emotions, and balancing the resulting dataset across emotions. We used the custom dataset to fine-tune 15 individual PLMs and 6 ensembles, in addition to testing 15 PLMs using zero-shot NLI. We used our metric ψ, from RQ2, to evaluate and choose the best performing classifier from the resulting 36 candidate individual models and ensembles.

In RQ3, we took advantage of several aspects of our work from RQ1 and RQ2. We used a similar method for fine-tuning the emotion models as we did in RQ1. We also benefited from the ablation study which identified the best performing emotion datasets. We evaluated and selected the best performing emotion classifier based on our evaluation using ψ from RQ2. However, we also evolved and refined our approach from our earlier work. We balanced our custom emotion dataset across labels as this can improve performance. In addition to individual models, we also used ensembles of individual models. We experimented with zero-shot NLI for emotion classification, which is a different approach to fine-tuning which we used heretofore. Crucially, we used Plutchik's 8 emotions instead of Ekman's 6 emotions. We made this choice because of the dimensional relationship between Plutchik's emotions and our ability to exploit this by introducing δ (a measure of emotion similarity) which elevated Plutchik's Wheel of Emotion from the `Ordinal` scale to the `Nominal` scale. This opened up numerous opportunities including: δ, δ-profile, $\phi(\delta)$, and ψ itself; which we subsequently used for evaluating and selecting an emotion classifier in addition to calculating a ψ score for *affective coherence* which was not originally anticipated. We hope that ψ, and its components (δ, δ-profile, $\phi(\delta)$) will find application in other areas, and we allude to some in the Future Work section.

7.2.4 RQ4 - Can we generate and compare the affective profile of human text and synthetic text?

We generated and compared the affective profile of human text and synthetic text. We used our emotion classifier (RQ3) to generate affective profiles for 10k human and synthetic articles from the training split of our NEWSsynth dataset (RQ1). We compared the affective coherence of the human and synthetic text for n-grams of emotion, and whole articles. We analysed them in relation to 'percent emotion change', and their ψ score (RQ2) which we calculated based on the δ-profile of each article. We repeated the process using binary sentiment, after evaluating two candidate sentiment classifiers. Two consistent patterns of results (PR1 and PR2) emerged across all of the experiments. PR1 showed that human and synthetic text both displayed affective coherence well above chance, and affective incoherence well below chance. It also showed that human text and synthetic text are quite similar, but human text is slightly more affectively coherent, and slightly less affectively incoherent, than synthetic text. Chi-squared analysis showed that, for both emotion and sentiment n-grams, the difference between human and synthetic text is statistically significant ($p < .001$). Finally, PR1 showed the disparity between affective coherence and random chance, and the disparity between affective incoherence and random chance, increases with n-gram length. PR2 showed that the central tendency (mean and median) and spread (standard deviation (σ)) are very similar for the distribution of changes of emotion or sentiment in human and synthetic text. Kolmogorov-Smirnov analysis showed that, for both emotion and sentiment distributions, the difference between human and synthetic text is statistically significant ($p < .001$). PR2 also showed there is a very weak, or borderline very weak, correlation between the emotion or sentiment in a human article and the corresponding synthetic article.

The consistency of the pattern of results (PR1 and PR2) across various length n-grams and full news articles, and for both emotions and sentiment, is noteworthy. It supports the validity of the mapping schemata for emotion (Table 5.1 mapping GNE from 15 emotions to Plutchik's 8 emotions), and sentiment (Table 6.6 mapping GNE and AT from 15 and 6 emotions respectively to binary sentiment). It supports our RQ3 selection of RoBERTa-large, using the ψ metric, as the emotion classification model because it yielded a similar pattern of results for the affective profiles as SiEBERT yielded. SiEBERT is a sentiment classifier with an average Accuracy of 93.2% across 15 datasets (Hartmann et al., 2023). It supports the validity of our method i.e. fine-tuning RoBERTa-large on our custom dataset (GAB) using our custom metric (ψ) for evaluation on the task of classifying Plutchik's 8 emotions; while SiEBERT is also a RoBERTa-large model but fine-tuned on 15 different datasets on the task of binary sentiment analysis by Hartmann et al. (2023). Both models were run against our custom dataset, NEWSsynth (train 10k), and both models yielded a consistent pattern of results. In summary, the consistency of results between methods which start with similar language models (no doubt initialised differently), which are then fine-tuned with different datasets (GAB vs 15 different datasets), and different emotion models (Plutchik vs binary sentiment), from different researchers (our work vs Hartmann et al. (2023)), and using different analyses (n-grams, full articles, 'percent emotion change', δ-profile, ψ score) gives some support to the results.

Another consistent result in the RQ4 experiments is the very weak, or weak (borderline very weak), correlation (Pearson, and Spearman) between the emotion or sentiment in a human article and the corresponding synthetic article. This is consistent with RQ1 where we saw very weak correlation between the length of a human prompt article and the length of the resulting synthetic article. This very weak correlation, between a human prompt article and the resulting synthetic article, is evidence of the difficulty of controllable text generation. It also suggests that Turing Test scenarios, classifying a pair of articles, may not benefit from the emotion and sentiment techniques used in our experiments.

RQ4 brought together aspects of our work from RQ1 (NEWSsynth dataset), RQ2 (ψ, δ-profile), and RQ3 (emotion classifier). RQ1 results support the hypothesis that PLMs may have an *affective deficit*. The results from RQ4 show some similarity, between human and synthetic text when compared against random chance, in their *affective profile* as it relates to measures of *affective coherence*. There is a consistent pattern where human text displays slightly more affective coherence, and slightly less affective incoherence, than synthetic text. The difference between human and synthetic text is statistically significant ($p < .001$). This difference may also prove to be exploitable in the task of synthetic text detection.

7.2.5 Overarching Hypothesis Revisited

In the Introduction (Chapter 1) we introduced four Research Questions (§1.2) and stated that the overarching hypothesis of this research is that **the task of synthetic text detection can be improved by exploiting emotion**. The results from RQ1 support this hypothesis by demonstrating several 'emotionally-aware' PLMs outperforming otherwise equivalent standard PLMs in the task of synthetic text detection. The results from RQ4 show that, while human and synthetic text show similar distributions, there are differences in their *affective profile* which are statistically significant ($p < .001$) and could be exploited in the task of synthetic text detection. We discuss this possibility in Future Work (§7.4). But first, we will address some limitations of our work.

7.3 Limitations

In this section, we discuss limitations which the reader should be cognisant of when interpreting the findings we have presented.

The datasets used in this work (synthetic text datasets, emotion datasets, and sentiment dataset) are English language and, where those datasets were used to fine-tune PLMs, the resulting PLM performance may vary for other languages. We primarily focus on the news domain and, while performance in other domains may vary (Merchant et al., 2020), we include experiments in several non-news domains (§3.5). Our rationale for choosing the news domain, and a description of the emotional content of our NEWSsynth dataset, was mentioned in the Introduction (Chapter 1).

Emotion classification is a difficult task (§2.3.4), notwithstanding that, the performance of the emotion classifiers in RQ1 and RQ3 leaves room for improvement - which, in turn, could lead to improvements in the downstream tasks i.e. synthetic text detection in RQ1, and improvements in the 'quality' of the affective profiles in RQ4 (which are generated by the top performing emotion classifier from RQ3). In RQ1 the emotion datasets are imbalanced across emotion labels, this imbalance can negatively impact performance of models fine-tuned on those datasets. However, we conducted ablation experiments to find the best combination of emotion and sentiment datasets (§3.4.3). The 15 emotions in GoodNewsEveryone were mapped to Plutchik's 8 emotions and Ekman's 6 emotions, factoring in Plutchik's wheel of emotion (Plutchik, 1980, 2001), but there is no firm agreement in the literature as to which is the 'correct' or 'best' emotion model (Ekman, 2016). The emotion models used in this work (Ekman's 6 basic emotions, and Plutchik's Wheel of Emotion) are the two most popular in affective NLP literature. We built on the mapping schema of (Bostan and Klinger, 2018) for emotion and explained our choices where that mapping schema was inadequate. Similarly, we followed the mapping schema of (Chaumartin, 2007) for mapping from emotions to sentiment where appropriate.

The maximum input sequence length of BERT is 512 tokens and articles longer than this are truncated, which may negatively affect performance on the synthetic text detection task (Ippolito et al., 2020). However, we also saw that increasing the input sequence length may actually contribute to poorer performance (§3.3).

Datasets derived from user-generated content raise ethical concerns, such as harmful content and privacy, and this applies to language models which use such content in training (whether pretraining or fine-tuning). It also applies to potential harmful content in synthetic text (Gehman et al., 2020) which are among the dangers which are the subject of ongoing discussion and research (Bender et al., 2021). Ethical concerns are highlighted with our published datasets and language models with the following advisory:

> Care must be taken when using these language models (`emoPLMsynth` and `PLMsynth`), and datasets (*NEWSsynth* and *ChatGPT100*) as they may produce or contain ethically problematic content. Data scraped

> from the web may contain content which is ethically problematic such as adult content, bias, toxicity etc. and web-scraped data is used in the pre-trained language models such as BERT, BLOOM and Grover. `PLMsynth` and `emoPLMsynth` are based on BERT or BLOOM PLMs, while *NEWSsynth* was generated by Grover. Consequently, `emoPLMsynth` and `PLMsynth` could produce text which is ethically problematic, while *NEWSsynth* may contain ethically problematic content. As a result, any use of the language models (`emoPLMsynth`, `PLMsynth`) or the datasets (*NEWSsynth* or *ChatGPT100*) should employ appropriate checks and test regimes to handle potential harmful content.

The results from RQ4 show a similarity between the affective profile of human and synthetic text in relation to affective coherence and affective incoherence. This similarity could be because the overarching hypothesis, i.e. "*the task of synthetic text detection can be improved by exploiting emotion*", is false. However, our results from RQ1 support the hypothesis for different size models, different datasets, and in different domains. Another possibility is that the hypothesis is true, but our results do not reflect the full extent of the differences between the affective profiles of human and synthetic text i.e. the differences are in fact larger than our results indicate. It could also be the case that the hypothesis is true and our results are accurate, i.e. the differences are small, but our analysis of the results did not adequately expose the differences. In both of these scenarios, exploring alternative methods for generating and analysing the affective profile may improve the results and the conclusions we can draw from them. We reiterate that the small differences may be exploitable using machine learning techniques (we expand on this in the future work section below). We further reiterate that, despite appearing similar, the differences between human and synthetic text are statistically significant ($p < .001$). The statistical significance was demonstrated with a chi-squared (χ^2) analysis for emotion n-grams (§6.3) and for sentiment n-grams (§6.5.2); and it was demonstrated with a Kolmogorov-Smirnov analysis of the distribution for human and synthetic text of the 'percent of emotion maximum changes per article' (§6.4.2) and 'percent of sentiment maximum changes per article' (§6.5.3).

In §1.1 we introduced the concepts of *affective coherence* and *affective incoherence* and hypothesised that human text is affectively coherent whereas synthetic text may be affectively incoherent. We provided some justification for this hypothesis and described how it may be investigated using the datasets, detectors, and metrics created in this work. We also highlighted that human text may be incoherent, and synthetic text may be coherent. We reiterate these points here, and we will refer to them in the next section - Future Work.

Resource usage (carbon, compute, memory, time) were reduced by using smaller models where possible and using random search rather than exhaustive grid search; but this can come at a performance cost. Large Language Models (LLMs) did not exist when this research began, but they are now widely available. While we have used some LLMs, such as BLOOM and GPT 3.5, much of this research uses the PLMs which were state-of-the-art at the time. Having addressed some of the limitations of our work, we now consider future work.

7.4 Future Work

In this section, we consider some of the ways we can build upon the work presented in this thesis, including addressing some of the above limitations.

We mentioned our use of monolingual English datasets. This could be expanded to include other specific languages and/or multilingual datasets. Another consideration, regarding language choice, is that the performance of specific languages can vary. For example, Bareiß et al. (2024) reported that, in relation to zero-shot NLI for emotion detection, prompting in English tends to give better results overall - regardless of the target language. We also mentioned our primary domain was news, this could also be expanded to include other domains of interest - for example, social media sources such as X (formerly Twitter), reddit etc. A key consideration for language and domain is the availability of suitable datasets.

7.4.1 RQ1

We saved resources by using smaller models where possible. RQ1 experiments could be repeated using other PLMs for detection such as RoBERTa (Liu et al., 2019) which we use in RQ3 and RQ4, along with the latest version of the LLMs we used in these experiments i.e. BLOOM (Scao et al., 2023) and GPT 4.o (OpenAI, 2022), in addition to the latest open source LLMs like LLaMa (Touvron et al., 2023b), LLaMA 2 (Touvron et al., 2023a), Vicuna (Chiang et al., 2023), Falcon (Penedo et al., 2023), and Mistral (Jiang et al., 2023). Additional gains may be found by hyperparameter tuning, including prompt engineering (Liu et al., 2023). However, hyperparameter tuning can be resource hungry. Where supported by the language model, we could allow for a longer input sequence length as this can improve the performance of synthetic text detectors (Ippolito et al., 2020).

7.4.2 RQ2

We used δ values where $\delta \in \mathbb{N}_0$ i.e. the Natural numbers including zero. However, where a continuous loss function $\phi(\delta)$ is used (as is the case for all the example loss functions we described (§4.4), with the exception of the Step function) it is possible to specify floating point numbers for δ values i.e. $\delta \in \mathbb{R}$. This could prove very powerful, practical and flexible for users who may have varying distances (and therefore δ values) between classes. For example, if Horse and Donkey are $\delta = 1$ apart it may suit the user to specify that a Mule is $\delta = 0.5$ from both. Further work could investigate and demonstrate practical use-cases, in NLP and beyond, where $\delta \in \mathbb{R}$ could prove advantageous. Similarly, investigating and demonstrating this could be done with the $\phi(\delta)$ loss functions specified in (§4.4) in addition to other potential loss functions.

It may be possible to further develop δ (and therefore $\phi(\delta)$ and ψ) to work on a `Ratio` scale as shown in Table 2.3 by using Plutchik's 8 emotions plus 'Neutral' - where neutral is the "absolute zero" value as described by Stevens (1946).

In RQ1 and RQ3 we fine-tuned PLMs using the standard cross-entropy loss function. We suggest using the ψ metric, and more specifically the $\phi(\delta)$ loss function, as the loss function when fine-tuning a PLM. This has the potential to improve classifier performance because it could give more nuanced and fine-grained loss information to the fine-tuning algorithm. In addition, the ψ metric can be used to evaluate the classifier performance (as we did in §5.3).

In RQ4 we used the ψ metric in a novel way, i.e. to measure the *affective coherence* of articles - based on their *affective profile*. As such, we used ψ as an additional method to compare the affective coherence of human and synthetic text. Once again, further work could investigate and demonstrate other practical applications, in NLP and beyond, for ψ.

7.4.3 RQ3

We used fine-tuning and zero-shot NLI for emotion classification in RQ3, and this could be extended to use LLMs for emotion classification. The LLMs listed in Future Work for RQ1 would be suitable candidates, and could include prompt-engineering with the prompt-types defined by Plaza-del Arco et al. (2022). When balancing the GAB dataset (§5.2), we noted that we would not use synthetic text to create additional examples for the dataset because this could interfere with our downstream task i.e. comparing affective profiles of human and synthetic text. We also noted that future work could investigate the effect of exposure bias (Ranzato et al., 2016) when fine-tuning on the task of emotion classification. For example, an ablation experiment using datasets labelled with emotions and containing human-only text, human-and-synthetic text, and synthetic-only text, could shed light on the impact of incorporating synthetic text when fine-tuning PLMs in the task of emotion classification. The experiment could be extended to use these fine-tuned models as `emoPLM` models and, similar to our RQ1 experiments, fine-tune them in the task of synthetic text detection i.e. `emoPLMsynth` models. This may shed light on the impact of exposure bias on the task of synthetic text detection. Arora et al. (2022) reported that exposure bias negatively impacts the quality of generated synthetic text, so it is worth investigating if exposure bias negatively impacts the performance of classifiers - particularly those that deal with human and synthetic text.

Additional NLI experiments (§5.4.3), could include:

- Using few-shot techniques to improve NLI performance over the zero-shot method we used.
- Similar to the LLM prompt-tuning above, prompt-tuning could result in a better `hypothesis` and improve NLI performance, leveraging the prompt-types defined by Plaza-del Arco et al. (2022).
- Creating ensembles of NLI models which may perform better than individual models. Basile et al. (2021) reported that Bayesian ensembles worked better than the best individual model in several domains.

The experiments involving `emoPLM` models from RQ1 could be repeated using Plutchik's 8 emotions, rather than Ekman's 6 emotions. They could also use $\phi(\delta)$ as a loss function, and ψ as an evaluation metric. This could provide a performance improvement on the task of synthetic text detection. It could also provide an interesting comparison between the more conventional cross-entropy loss function and validation on $F1_{\mu}$, versus our $\phi(\delta)$ loss function and validation on ψ.

7.4.4 RQ4

We generated *affective profiles* for 10k human and synthetic articles from NEWSsynth. From these affective profiles, we created n-grams of emotions which, in turn, we used to calculate the 'percent emotion change' in the n-gram. An additional experiment could directly compare the actual n-grams of emotion (e.g. count occurrences etc.) between human and synthetic articles (rather than the derived 'percent emotion change'). This experiment (using the datasets, detectors, and metrics created in this work) can form the basis of modelling the structure of emotion patterns in human and synthetic text. Furthermore, there are many affective features of an article (e.g. counts for each permutation of emotion n-gram, 'percent emotion changes', count of each emotion, δ-profile etc.) which we can derive from the affective profile, and could be used as features in classic machine learning models. These features can be used in concert with more typical NLP features which (Alswaidan and Menai, 2020) have shown to be optimal in the task of emotion recognition using classic machine learning models like Support Vector Machines (Boser et al., 1992). These classic machine learning models may be able to exploit the consistent patterns of results we saw in RQ4, where we noted "The results from RQ4 show differences in the *affective profile* of human and synthetic text which, while slight to human eyes, could be exploited by machine learning algorithms". This possibility of exploiting these differences is supported by the statistical significance of the difference between human and synthetic text in relation to both emotion and sentiment, for n-grams and for 'percent of maximum emotion / sentiment changes'. Results showed very weak correlation between human prompt articles and the corresponding synthetic article. We mentioned in §6.5.3 that one could investigate the correlation between combinations of article length, ψ-score, 'percent emotion or sentiment change'. For example, combining article length and ψ-score etc., in human prompt articles and the corresponding synthetic article and measuring the Pearson correlation. We could include features of the *affective profile* as listed above in the prompt to a zero-shot or few-shot NLI model, or to an LLM. Prompt-tuning could yield an optimal prompt for the task of synthetic text detection, which may include some of these affective features. Finally, as mentioned in §6.3, explainable AI techniques could shed light on the extent to which the affective profile (encompassing affective coherence and affective incoherence) is a factor in the improved performance of emotionally-aware models over standard models. Additionally, classical machine learning models and their features could be used in ablation studies to selectively determine the impact of specific features relating to the affective profile of human and synthetic text.

7.5 Concluding Remarks

When this research began, in October 2019 (a few months before the start of the Covid pandemic), the NLP world was besotted with BERT (Google) and its benchmark breaking performance. GPT-2 (OpenAI) had been met with excitement, but fear quickly followed due to concerns of its potential misuse, and it went through a phased release as its impact was monitored over a period of months. GPT-2 has 1.5B parameters.

In November 2022, ChatGPT (OpenAI) rocked the world with its easy-to-use web interface and super impressive, though not perfect, generative text. Fears of plagiarism quickly arose and educators around the world redesigned assessments. For many people, their fear relates to job security as new reports list the 'jobs that AI will replace'. Job losses due to AI are already happening. GPT-4 reportedly has 1,760B parameters.

So what of the future? Geoffrey Hinton and Yann LeCun both believe AI will surpass human intelligence. Hinton (who left Google) is concerned about the potential consequences for humanity, LeCun (who works at Meta) is more optimistic. Yoshua Bengio (University Professor) acknowledges the risks and calls for rational discourse.

"*The future has not been written. There is no fate but what we make for ourselves.*" So says John Connor, leader of the human resistance against 'Skynet' - a malign artificial super intelligence in a dystopian future filled with human-hunting cyborgs - in the movie Terminator. Setting aside the timeless philosophical conundrum of free-will versus determinism, if Connor is correct, then the choice is ours as to how we use artificial intelligence. Hopefully our choices and our shared future will reflect the best of us.

Bibliography

Acheampong, F. A., Wenyu, C., and Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/eng2.12189.

Adelani, D. I., Mai, H., Fang, F., Nguyen, H. H., Yamagishi, J., and Echizen, I. (2019). Generating sentiment-preserving fake online reviews using neural language models and their human- and machine-based detection. *arXiv:1907.09177 [cs]*. arXiv: 1907.09177.

Afiki, Y. and Bar, M. (2020). Our need for associative coherence. *Humanities and Social Sciences Communications*, 7(1):1–11. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Palgrave Subject_term: Psychology;Science, technology and society Subject_term_id: psychology;science-technology-and-society.

Agrawal, A., An, A., and Papagelis, M. (2018). Learning emotion-enriched word representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 950–961, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alswaidan, N. and Menai, M. E. B. (2020). A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*.

An, S., Ji, L.-J., Marks, M., and Zhang, Z. (2017). Two sides of emotion: Exploring positivity and negativity in six basic emotions across cultures. *Frontiers in Psychology*, 8:610.

Aristotle (0350BCE). *Rhetoric*. Aristotle. Issued: -0350BCE.

Arora, K., Asri, L. E., Bahuleyan, H., and Cheung, J. C. K. (2022). Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. arXiv:2204.01171 [cs].

Badaro, G., Jundi, H., Hajj, H., and El-Hajj, W. (2018). EmoWordNet: Automatic expansion of emotion lexicon using english WordNet. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 86–93, New Orleans, Louisiana. Association for Computational Linguistics.

Bareiß, P., Klinger, R., and Barnes, J. (2024). English prompts are better for NLI-based zero-shot emotion classification than target-language prompts. arXiv:2402.03223 [cs].

Barry, J., Wagner, J., Cassidy, L., Cowap, A., Lynn, T., Walsh, A., Ó Meachair, M. J., and Foster, J. (2022). gaBERT — An irish language model. In *Proceedings of the Language Resources and Evaluation Conference*, pages 4774–4788, Marseille, France. European Language Resources Association.

Basile, A., Pérez-Torró, G., and Franco-Salvador, M. (2021). Probabilistic ensembles of zero- and few-shot learning models for emotion classification. In Mitkov, R. and Angelova, G., editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 128–137, Held Online. INCOMA Ltd.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., and Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, 5(4):323–370. Publisher: SAGE Publications Inc.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Berger, J. and Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, 49(2):192–205. Publisher: SAGE Publications Inc.

Bhat, M. M. and Parthasarathy, S. (2020). How effectively can machines defend against machine-generated fake news? An empirical study. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 48–53, Online. Association for Computational Linguistics.

Bohn-Gettler, C. M. and Rapp, D. N. (2014). Emotion during reading and writing. In *International handbook of emotions in education*, Educational psychology handbook series, pages 437–457. Routledge/Taylor & Francis Group, New York, NY, US.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *arXiv:1607.04606 [cs]*. arXiv: 1607.04606.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29:4349–4357.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, COLT '92, pages 144–152, New York, NY, USA. Association for Computing Machinery.

Bostan, L. A. M., Kim, E., and Klinger, R. (2020). GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.

Bostan, L.-A.-M. and Klinger, R. (2018). An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Brand, A. (1991). Social cognition, emotions, and the psychology of writing. *Journal of Advanced Composition*, 11(2):395–407. Publisher: JAC.

Brand, A. G. (1985). Hot cognition: Emotions and writing behavior. *Journal of Advanced Composition*, 6:5–15. Publisher: JAC.

Brand, A. G. (1987). The why of cognition: Emotion and the writing process. *College Composition and Communication*, 38(4):436–443. Publisher: National Council of Teachers of English.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *arXiv:2005.14165 [cs]*. arXiv: 2005.14165.

Camerer, C., Loewenstein, G., and Prelec, D. (2005). Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature*, 43(1):9–64.

Carstensen, L. L. and Mikels, J. A. (2005). At the intersection of emotion and cognition: Aging and the positivity effect. *Current Directions in Psychological Science*, 14(3):117–121. Publisher: SAGE Publications Inc.

Chaumartin, F.-R. (2007). UPAR7: A knowledge-based system for headline sentiment tagging. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 422–425, Prague, Czech Republic. Association for Computational Linguistics.

Chen, C. Y., Hung, T. M., Hsu, Y.-L., and Ku, L.-W. (2023). Label-aware hyperbolic embeddings for fine-grained emotion classification. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10947–10958, Toronto, Canada. Association for Computational Linguistics.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. (2023). Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2(3):113–124. Conference Name: IRE Transactions on Information Theory.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., and Smith, N. A. (2021). All that's 'human' is not gold: Evaluating human evaluation of generated text. *arXiv:2107.00061 [cs]*. arXiv: 2107.00061.

Combs, B. and Slovic, P. (1979). Newspaper coverage of causes of death. *Journalism Quarterly*, 56(4):837–849. Publisher: SAGE Publications.

Crowley, S. (2023a). Irish Times editor apologises for 'breach of trust'. Section: News.

Crowley, S. (2023b). Irish Times takes down article amid AI suggestions. Section: News.

Damasio, A. R., Everitt, B. J., Bishop, D., Roberts, A. C., Robbins, T. W., and Weiskrantz, L. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 351(1346):1413–1420. Publisher: Royal Society.

Darwin, C. (1872). *The expression of the emotions in man and animals, 3rd ed.* Oxford University Press, New York, NY, US.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805 [cs]*. arXiv: 1810.04805.

Dijksterhuis, A., Bos, M. W., Nordgren, L. F., and van Baaren, R. B. (2006). On making the right choice: The deliberation-without-attention effect. *Science*, 311(5763):1005–1007. Publisher: American Association for the Advancement of Science.

Dijksterhuis, A. and Nordgren, L. F. (2006). A theory of unconscious thought. *Perspectives on Psychological Science*, 1(2):95–109. Publisher: SAGE Publications Inc.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3/4):169–200.

Ekman, P. (1999). Basic emotions. In *Handbook of Cognition and Emotion*, pages 45–60. John Wiley & Sons, Ltd.

Ekman, P. (2016). What scientists who study emotion agree about. *Perspectives on Psychological Science*, 11(1):31–34.

Evans, J. S. B. T. and Stanovich, K. E. (2013). Dual-process theories of higher cognition: advancing the debate. *Perspectives on Psychological Science*, 8(3):223–241. Publisher: SAGE Publications Inc.

Eysenck, M. W. and Keane, M. T. (2020). *Cognitive psychology: A student's handbook.* Taylor & Francis Group, Milton, UNITED KINGDOM.

Firth, J. (1957). A synopsis of linguistic theory 1930–1955. In *Studies in linguistic analysis.*, Philological Society (Great Britain). Blackwell, Oxford. OCLC: 1805360.

Galtung, J. and Ruge, M. H. (1965). The structure of foreign news: The presentation of the congo, cuba and cyprus crises in four norwegian newspapers. *Journal of Peace Research*, 2(1):64–90. Publisher: SAGE Publications Ltd.

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., and Pearson, A. T. (2022). Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. Pages: 2022.12.23.521610 Section: New Results.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Gehrmann, S., Strobelt, H., and Rush, A. (2019). GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5):602–610.

Gupta, S., Nguyen, H. H., Yamagishi, J., and Echizen, I. (2020). Viable threat on news reading: Generating biased news using natural language models. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 55–65, Online. Association for Computational Linguistics.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814–834.

Hall, N.-A. (2021). Understanding brexit on facebook: Developing close-up, qualitative methodologies for social media research. *Sociological Research Online*, page 13607804211037356. Publisher: SAGE Publications Ltd.

Hartmann, J., Heitmann, M., Siebert, C., and Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.

Haugen, F. (2021). Public hearing on whistle-blower's testimony on the negative impact of big tech companies' products on user: Opening statement by Frances Haugen. URL: https://multimedia.europarl.europa.eu/it/public-hearing-on-whistle-blowers-testimony-on-the-negative-impact-of-big-tech-companies-products-on-user-frances-haugen-opening-statements_I213108-V_v.

Hayes, J., Schriver, K., Blaustein, A., and Spilka, R. (1986). If it's clear to me, it must be clear to them: How knowledge makes it difficult to judge. In *American Educational Research Association (AERA) Conference, San Francisco.*

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, 29(3):369–388. Publisher: SAGE Publications Inc.

Hegel, A., Rao, S., Celikyilmaz, A., and Dolan, B. (2020). Substance over style: Document-level targeted content transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6485–6504, Online. Association for Computational Linguistics.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, 466(7302):29–29. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7302 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group Subject_term: Psychology and behaviour;Scientific community Subject_term_id: psychology-and-behaviour;scientific-community.

Hochreiter, S. and Schmidhuber, J. (1996). LSTM can solve hard long time lag problems. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*, NIPS'96, pages 473–479, Cambridge, MA, USA. MIT Press.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2020). The curious case of neural text degeneration. In *ICLR2020*.

Horne, B. D. and Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *2nd International Workshop on News and Public Opinion at ICWSM*. arXiv: 1703.09398.

Hovy, D. (2016). The enemy in your own camp: How well can we detect statistically-generated fake reviews – an adversarial study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 351–356, Berlin, Germany. Association for Computational Linguistics.

Hovy, D. and Spruit, S. L. (2016). The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Huang, C., Zaïane, O., Trabelsi, A., and Dziri, N. (2018). Automatic dialogue generation with expressed emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 49–54, New Orleans, Louisiana. Association for Computational Linguistics.

Ippolito, D., Duckworth, D., Callison-Burch, C., and Eck, D. (2020). Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Isaak, J. and Hanna, M. J. (2018). User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer*, 51(8):56–59. Conference Name: Computer.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7B. arXiv:2310.06825 [cs].

Jooste, W., Haque, R., and Way, A. (2022). Knowledge distillation: A method for making neural machine translation more efficient. *Information*, 13(2):88. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58(9):697–720.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291. Publisher: [Wiley, Econometric Society].

Kanouse, D. and Hanson, L. R. (1972). Negativity in evaluations. In Jones, E. E., editor, *Attribution: perceiving the causes of behavior*. General Learning Press, Morristown, N.J. OCLC: 516505.

Kaplan, S. (2001). Meditation, restoration, and the management of mental fatigue. *Environment and Behavior*, 33(4):480–506. Publisher: SAGE Publications Inc.

Knaller, S. (2017). Emotions and the process of writing. In Jandl, I., Knaller, S., Schönfellner, S., and Tockner, G., editors, *Lettre*, pages 17–28. transcript Verlag, Bielefeld, Germany, 1 edition.

Kratzwald, B., Ilic, S., Kraus, M., Feuerriegel, S., and Prendinger, H. (2018). Deep learning for affective computing: text-based emotion recognition in decision support. *Decision Support Systems*, 115:24–35. arXiv: 1803.06397.

Kuran, T. and Sunstein, C. R. (1999). Availability cascades and risk regulation. *Stanford Law Review*, 51(4):683–768.

Kusal, S., Patil, S., Choudrie, J., Kotecha, K., Vora, D., and Pappas, I. (2023). A systematic review of applications of natural language processing and future challenges with special emphasis in text-based emotion detection. *Artificial Intelligence Review*, 56(12):15129–15215.

Labat, S., Hadifar, A., Demeester, T., and Hoste, V. (2022). An emotional journey: Detecting emotion trajectories in dutch customer service dialogues. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 106–112, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Laurer, M., Atteveldt, W. v., Casas, A., and Welbers, K. (2024). Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. *Political Analysis*, 32(1):84–100.

Leroy, S. (2009). Why is it so hard to do my work? The challenge of attention residue when switching between work tasks. *Organizational Behavior and Human Decision Processes*, 109(2):168–181.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., and Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory*, 4(6):551–578. Place: US Publisher: American Psychological Association.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):195:1–195:35.

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:1907.11692 [cs]*. arXiv: 1907.11692.

Lorandi, M. and Belz, A. (2023). How to control sentiment in text generation: A survey of the state-of-the-art in sentiment-control techniques. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, Toronto, Canada.

Massarelli, L., Petroni, F., Piktus, A., Ott, M., Rocktäschel, T., Plachouras, V., Silvestri, F., and Riedel, S. (2020). How decoding strategies affect the verifiability of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 223–235, Online. Association for Computational Linguistics.

Merchant, A., Rahimtoroghi, E., Pavlick, E., and Tenney, I. (2020). What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In Bengio, Y. and LeCun, Y., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Miller, R. A. and Albert, K. (2015). If it leads, it bleeds (and if it bleeds, it leads): Media coverage and fatalities in militarized interstate disputes. *Political Communication*, 32(1):61–82. Publisher: Routledge _eprint: https://doi.org/10.1080/10584609.2014.880976.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., and Finn, C. (2023). DetectGPT: zero-shot machine-generated text detection using probability curvature. arXiv:2301.11305 [cs] version: 1.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8640.2012.00460.x.

Morewedge, C. K. and Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in Cognitive Sciences*, 14(10):435–440. Publisher: Elsevier.

Neisser, U. (1963). The imitation of man by machine. Neisser, Ulric. "The Imitation of Man by Machine." Science, vol. 139, no. 3551, 1963, pp. 193–197. JSTOR, www.jstor.org/stable/1710006. Accessed 12 Nov. 2020.

OpenAI (2022). Introducing ChatGPT. URL:https://openai.com/blog/chatgpt.

Ortony, A., Clore, G. L., and Collins, A. (1988). *The cognitive structure of emotions*. Cambridge University Press, Cambridge.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. In *NeurIPS*. arXiv. Version Number: 1.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., and Launay, J. (2023). The RefinedWeb dataset for falcon LLM: Outperforming curated corpora with web data, and web data only. arXiv:2306.01116 [cs].

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv:1802.05365 [cs]*. arXiv: 1802.05365.

Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., and Harchaoui, Z. (2021). MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems*.

Plato (0370BCE). *Phaedrus*. Plato. Issued: -0370BCE URL: https://www.gutenberg.org/files/1636/1636-h/1636-h.htm.

Plaza-del Arco, F. M., Martín-Valdivia, M.-T., and Klinger, R. (2022). Natural language inference prompts for zero-shot emotion classification in text across corpora. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6805–6817, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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.

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350. Publisher: Sigma Xi, The Scientific Research Society.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. (2016). Sequence level training with recurrent neural networks. In *ICLR 2016*.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., and Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Rashkin, H., Smith, E. M., Li, M., and Boureau, Y.-L. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv:1811.00207 [cs]*. arXiv: 1811.00207.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2021). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Rozin, P. and Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, 5(4):296–320. Publisher: SAGE Publications Inc.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019*. arXiv. arXiv:1910.01108 [cs].

Santus, E., Lu, Q., Lenci, A., and Huang, C.-R. (2014). Taking antonymy mask off in vector space. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pages 135–144, Phuket,Thailand. Department of Linguistics, Chulalongkorn University.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del

Moral, A. V., Ruwase, O., Bawden, R., Bekman, S., McMillan-Major, A., Beltagy, I., Nguyen, H., Saulnier, L., Tan, S., Suarez, P. O., Sanh, V., Laurençon, H., Jernite, Y., Launay, J., Mitchell, M., Raffel, C., Gokaslan, A., Simhi, A., Soroa, A., Aji, A. F., Alfassy, A., Rogers, A., Nitzav, A. K., Xu, C., Mou, C., Emezue, C., Klamm, C., Leong, C., van Strien, D., Adelani, D. I., Radev, D., Ponferrada, E. G., Levkovizh, E., Kim, E., Natan, E. B., De Toni, F., Dupont, G., Kruszewski, G., Pistilli, G., Elsahar, H., Benyamina, H., Tran, H., Yu, I., Abdulmumin, I., Johnson, I., Gonzalez-Dios, I., de la Rosa, J., Chim, J., Dodge, J., Zhu, J., Chang, J., Frohberg, J., Tobing, J., Bhattacharjee, J., Almubarak, K., Chen, K., Lo, K., Von Werra, L., Weber, L., Phan, L., allal, L. B., Tanguy, L., Dey, M., Muñoz, M. R., Masoud, M., Grandury, M., Šaško, M., Huang, M., Coavoux, M., Singh, M., Jiang, M. T.-J., Vu, M. C., Jauhar, M. A., Ghaleb, M., Subramani, N., Kassner, N., Khamis, N., Nguyen, O., Espejel, O., de Gibert, O., Villegas, P., Henderson, P., Colombo, P., Amuok, P., Lhoest, Q., Harliman, R., Bommasani, R., López, R. L., Ribeiro, R., Osei, S., Pyysalo, S., Nagel, S., Bose, S., Muhammad, S. H., Sharma, S., Longpre, S., Nikpoor, S., Silberberg, S., Pai, S., Zink, S., Torrent, T. T., Schick, T., Thrush, T., Danchev, V., Nikoulina, V., Laippala, V., Lepercq, V., Prabhu, V., Alyafeai, Z., Talat, Z., Raja, A., Heinzerling, B., Si, C., Taşar, D. E., Salesky, E., Mielke, S. J., Lee, W. Y., Sharma, A., Santilli, A., Chaffin, A., Stiegler, A., Datta, D., Szczechla, E., Chhablani, G., Wang, H., Pandey, H., Strobelt, H., Fries, J. A., Rozen, J., Gao, L., Sutawika, L., Bari, M. S., Al-shaibani, M. S., Manica, M., Nayak, N., Teehan, R., Albanie, S., Shen, S., Ben-David, S., Bach, S. H., Kim, T., Bers, T., Fevry, T., Neeraj, T., Thakker, U., Raunak, V., Tang, X., Yong, Z.-X., Sun, Z., Brody, S., Uri, Y., Tojarieh, H., Roberts, A., Chung, H. W., Tae, J., Phang, J., Press, O., Li, C., Narayanan, D., Bourfoune, H., Casper, J., Rasley, J., Ryabinin, M., Mishra, M., Zhang, M., Shoeybi, M., Peyrounette, M., Patry, N., Tazi, N., Sanseviero, O., von Platen, P., Cornette, P., Lavallée, P. F., Lacroix, R., Rajbhandari, S., Gandhi, S., Smith, S., Requena, S., Patil, S., Dettmers, T., Baruwa, A., Singh, A., Cheveleva, A., Ligozat, A.-L., Subramonian, A., Névéol, A., Lovering, C., Garrette, D., Tunuguntla, D., Reiter, E., Taktasheva, E., Voloshina, E., Bogdanov, E., Winata, G. I., Schoelkopf, H., Kalo, J.-C., Novikova, J., Forde, J. Z., Clive, J., Kasai, J., Kawamura, K., Hazan, L., Carpuat, M., Clinciu, M., Kim, N., Cheng, N., Serikov, O., Antverg, O., van der Wal, O., Zhang, R., Zhang, R., Gehrmann, S., Mirkin, S., Pais, S., Shavrina, T., Scialom, T., Yun, T., Limisiewicz, T., Rieser, V., Protasov, V., Mikhailov, V., Pruksachatkun, Y., Belinkov, Y., Bamberger, Z., Kasner, Z., Rueda, A., Pestana, A., Feizpour, A., Khan, A., Faranak, A., Santos, A., Hevia, A., Unldreaj, A., Aghagol, A., Abdollahi, A., Tammour, A., HajiHosseini, A., Behroozi, B., Ajibade, B., Saxena, B., Ferrandis, C. M., Contractor, D., Lansky, D., David, D., Kiela, D., Nguyen, D. A., Tan, E., Baylor, E., Ozoani, E., Mirza, F., Ononiwu, F., Rezanejad, H., Jones, H., Bhattacharya, I., Solaiman, I., Sedenko, I., Nejadgholi, I., Passmore, J., Seltzer, J., Sanz, J. B., Dutra, L., Samagaio, M., Elbadri, M., Mieskes, M., Gerchick, M., Akinlolu, M., McKenna, M., Qiu, M., Ghauri, M., Burynok, M., Abrar, N., Rajani, N., Elkott, N., Fahmy, N., Samuel, O., An, R., Kromann, R., Hao, R., Alizadeh, S., Shubber, S., Wang, S., Roy, S., Viguier, S., Le, T., Oyebade, T., Le, T., Yang, Y., Nguyen, Z., Kashyap, A. R., Palasciano, A., Callahan, A., Shukla, A., Miranda-Escalada, A., Singh, A., Beilharz, B., Wang, B., Brito, C., Zhou, C., Jain, C., Xu, C., Fourrier, C., Periñán, D. L., Molano, D., Yu, D., Manjavacas, E., Barth, F., Fuhrimann, F., Altay, G., Bayrak, G., Burns, G., Vrabec, H. U., Bello, I., Dash, I., Kang, J., Giorgi, J., Golde, J., Posada, J. D., Sivaraman, K. R., Bulchandani, L., Liu, L., Shinzato, L., de Bykhovetz, M. H., Takeuchi, M., Pàmies, M., Castillo, M. A., Nezhurina, M., Sänger, M., Samwald, M., Cullan, M., Weinberg, M., De Wolf, M., Mihaljcic, M., Liu, M., Freidank, M., Kang, M., Seelam, N., Dahlberg, N., Broad, N. M., Muellner, N., Fung, P., Haller, P., Chandrasekhar, R., Eisenberg, R., Martin, R., Canalli, R., Su, R., Su, R., Cahyawijaya, S., Garda, S., Deshmukh, S. S., Mishra, S., Kiblawi, S., Ott, S., Sang-aroonsiri, S., Kumar, S., Schweter, S., Bharati, S., Laud, T., Gigant, T., Kainuma, T., Kusa, W., Labrak, Y., Bajaj, Y. S., Venkatraman, Y., Xu, Y., Xu, Y., Xu, Y., Tan, Z., Xie, Z., Ye, Z., Bras, M., Belkada, Y., and Wolf, T. (2023). BLOOM: A 176B-parameter open-access multilingual language model. arXiv:2211.05100 [cs].

Schuster, T., Schuster, R., Shah, D. J., and Barzilay, R. (2020). The limitations of stylometry for detecting machine-generated fake news. *arXiv:1908.09805 [cs]*. arXiv: 1908.09805.

Seyeditabari, A., Tabari, N., Gholizade, S., and Zadrozny, W. (2019). Emotional embeddings: Refining word embeddings to capture emotional content of words. *arXiv:1906.00112 [cs]*. arXiv: 1906.00112.

Seyeditabari, A. and Zadrozny, W. (2017). Can word embeddings help find latent emotions in text? Preliminary results. In *The Thirtieth International Flairs Conference*.

Shah, S., Reddy, S., and Bhattacharyya, P. (2023). Retrofitting light-weight language models for emotions using supervised contrastive learning. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3640–3654, Singapore. Association for Computational Linguistics.

Shen, J. H., Lapedriza, A., and Picard, R. W. (2019). Unintentional affective priming during labeling may bias labels. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 587–593. ISSN: 2156-8111.

Slovic, P., Finucane, M. L., Peters, E., and MacGregor, D. G. (2007). The affect heuristic. *European Journal of Operational Research*, 177(3):1333–1352.

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.

Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J., Hing, C., Kwok, C. S., Pang, C., and Harvey, I. (2010). Dissemination and publication of research findings : An updated review of related biases. *Health Technology Assessment*, 14(8):1–220.

Spielberger, C. D. (1972). Anxiety as an emotional state. In Spielberger, C. D., editor, *Anxiety: Current trends and research*, volume 1. New York Academic Press. Book Title: Anxiety.

Stanovich, K. E. and West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23(5):645–665. Publisher: Cambridge University Press.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684):677–680. Publisher: American Association for the Advancement of Science.

Strapparava, C. and Mihalcea, R. (2007). SemEval-2007 Task 14: Affective Text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

Strapparava, C. and Mihalcea, R. (2008). Learning to identify emotions in text. In *SAC '08*.

TenHouten, W. (2017). Social dominance hierarchy and the pride-shame system. *Journal of Political Power*, 10:1–21.

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, 1(1):39–60.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023a). LLaMA: Open and efficient foundation language models. arXiv:2302.13971 [cs].

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023b). Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288 [cs].

Troiano, E., Padó, S., and Klinger, R. (2021). Emotion ratings: How intensity, annotation confidence and agreements are entangled. *arXiv:2103.01667 [cs]*. arXiv: 2103.01667.

Turing, A. M. (1950). I.—Computing machinery and intelligence. *Mind*, LIX(236):433–460. Publisher: Oxford Academic.

Tversky, A. and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76(2):105–110. Place: US Publisher: American Psychological Association.

Tversky, A. and Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2):207–232.

Tzu, H. (0250BCE). *Xunzi*. Hsün Tzu. Issued: -0250BCE.

Uchendu, A., Ma, Z., Le, T., Zhang, R., and Lee, D. (2021). TuringBench: A benchmark environment for turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wang, S., Maoliniyazi, A., Wu, X., and Meng, X. (2020a). Emo2Vec: Learning emotional embeddings via multi-emotion category. *ACM Transactions on Internet Technology*, 20(2):13:1–13:17.

Wang, Y., Zhang, J., Ma, J., Wang, S., and Xiao, J. (2020b). Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 186–195, 1st virtual meeting. Association for Computational Linguistics.

Wardle, C. and Derakhshan, H. (2018). Thinking about 'information disorder': Formats of misinformation, disinformation, and mal-information. In *Journalism, 'Fake News' & Disinformation*, pages 43–54. UNESCO, Paris. Ireton, Cherilyn; Posetti, Julie.

Wason, P. C. and Evans, J. S. B. T. (1974). Dual processes in reasoning? *Cognition*, 3(2):141–154.

Weiss, M. (2019). Deepfake bot submissions to federal public comment websites cannot be distinguished from human submissions. *Technology Science*.

Wemmer, E., Labat, S., and Klinger, R. (2024). EmoProgress: Cumulated emotion progression analysis in dreams and customer service dialogues. In *Proceedings of The Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, Torino, Italy. European Language Resources Association.

Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell.

Wundt, W. (1896). Grundriss der Psychologie. *Philosophical Review*, 5(n/a):331. Publisher: Duke University Press.

Yang, L., Song, Y., Ren, X., Lyu, C., Wang, Y., Liu, L., Wang, J., Foster, J., and Zhang, Y. (2023). Out-of-distribution generalization in text classification: Past, present, and future. arXiv:2305.14104 [cs].

Zajonc, R. (2001). Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science*, 10(6):224–228. Publisher: SAGE Publications Inc.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, 9(2, Pt.2):1–27. Place: US Publisher: American Psychological Association.

Zeigarnik, B. (1927). Das Behalten erledigter und unerledigter Handlungen [On the retention of completed and uncompleted transactions]. *Psychologische Forschung*, 9:1–85.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. (2019). Defending against neural fake news. In Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. d., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 9054–9065. Curran Associates, Inc.

Zhong, W., Tang, D., Xu, Z., Wang, R., Duan, N., Zhou, M., Wang, J., and Yin, J. (2020). Neural deepfake detection with factual structure of text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2461–2470, Online. Association for Computational Linguistics.

Appendix A

Reproducibility

All code, models, and datasets (including *NEWSsynth* and *ChatGPT100*) are available at `https://github.com/alanagiasi/emoPLMsynth`.

Hyper-parameters used for generating synthetic text with Grover

Grover$_{\text{BASE}}$ was used for generating synthetic text news articles in *NEWSsynth*. Full contextual metadata was used, in addition to a top-p value of 0.95 because both can make discrimination more difficult. According to Zellers et al. (2019) contextual data decreased perplexity by 0.9 points for Grover$_{\text{BASE}}$, and a top-p value in the range 0.92 to 0.98 is a Goldilocks zone where discrimination is hardest (so we chose top-p=0.95 as it is in the middle of this difficult detection zone). Source code, installation, and generation instructions for Grover can be found on the Grover github. [1]

- Model: GROVER$_{\text{BASE}}$
- Model parameters: 124M
- Top-p = 0.95
- Metadata: Full contextual metadata (from RealNews-Test dataset)
- Time to generate 20k synthetic articles is approximately 90 hours on a single GPU (Tesla K40, or RTX2080ti) with 30GB RAM.

Metrics

Accuracy, Precision, Recall, F1, (and F1$_\mu$ for emotion classification), mean absolute error, and root mean square error were calculated using scikit-learn. [2] Pearson ρ and Spearman ρ correlation were calculated using pandas. [3] ψ metric is novel and defined in Chapter 4.

---

[1] `https://github.com/rowanz/grover`

[2] scikit-learn describes the metrics: `https://scikit-learn.org/stable/modules/model_evaluation.html#common-cases-predefined-values`. As noted in §3.2.3 when fine-tuning `emoBERT` on emotions: micro averaging over a single-label multi-class evaluation means that Accuracy, Precision, Recall and F1 all have the same value. `https://scikit-learn.org/stable/modules/model_evaluation.html#multiclass-and-multilabel-classification`

[3] `https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html`

Datasets

NEWSsynth We release *NEWSsynth* - a dataset comprising 40k English language human and synthetic news articles. The experiments in this paper use the first 20k of these articles, an additional 20k articles are provided in the dataset. The human articles are taken from the RealNews-Test dataset (Zellers et al., 2019) so they have not been seen by Grover - which generated the synthetic news articles as described in §3.2.3 and earlier in this Appendix.

ChatGPT100 We release *ChatGPT100*, a dataset comprising 100 English language articles in various non-news domains (Science, Entertainment (Music, Movies), Sport, Business, and Philosophy). 50 articles are human written, and 50 articles are generated by ChatGPT. The 100 articles have all been manually curated and do not contain toxic content. Furthermore, ChatGPT has a content filter which flags potentially harmful content.

The 50 human articles contained in *ChatGPT100* were gathered between 16-24 March 2023 from the domains shown in Table A.1. The 50 synthetic articles contained in *ChatGPT100* were generated using ChatGPT 3.5 (March 14 2023 version: gpt-3.5-turbo-0314) on dates between 16-24 March 2023.

RealNews and RealNews-Test These datasets were released with Grover and are described there in detail (Zellers et al., 2019).

Emotion and Sentiment Datasets GoodNewsEveryone is described in detail (Bostan et al., 2020) with modifications made to the dataset for this work described in §3.2.2 and Chapter 6. AffectiveText was released as part of SemEval 2008 and is described in detail (Strapparava and Mihalcea, 2008), while the SST-2 sentiment dataset is described in detail (Socher et al., 2013).

Domain	Count
britannica.com	9
investopedia.com	6
plato.stanford.edu	6
fandom.com	2
forbes.com	2
olympics.com	2
allmusic.com	1
arpansa.gov.au	1
arsenal.com	1
atptour.com	1
bbc.com	1
bhf.org.uk	1
bleacherreport.com	1
cambridge.org	1
canarahsbclife.com	1
empireonline.com	1
gaa.ie	1
hotpress.com	1
kaspersky.com	1
laureus.com	1
mayfieldclinic.com	1
oah.org	1
oceanservice.noaa.gov	1
open.lib.umn.edu	1
phys.org	1
science.nasa.gov	1
sixnationsrugby.com	1
slf.rocks	1
u2.com	1
Total:	50

Table A.1: Domains used for human text in *ChatGPT100* dataset released with this paper. Articles were gathered between 16-24 March 2023.

Appendix B

Hyperparameters Used for Fine-Tuning

The hyperparameters used for PLM fine-tuning are listed below. If not specifically listed, the hyperparameter value used was the default using HuggingFace Transformer libraries. [1] The $\text{BERT}_{\text{BASE}}$-cased, $\text{BERT}_{\text{LARGE}}$-cased, and BLOOM560M models were downloaded from HuggingFace. [2]

`emoBERT`, `BERTsynth`, and `emoBERTsynth` were all trained using freely available Google Colab with a single GPU (Tesla K80 or Tesla T4) with no guarantee on available RAM [3] or an NVIDIA GeForce RTX3090 GPU with 24GB RAM.

All models were trained and evaluated for 5 runs using different seeds for each of the 5 runs. The seeds used are listed below.

B.1 `BERTsynth`, `emoBERTsynth`

- Model: $\text{BERT}_{\text{BASE}}$-cased — $\text{BERT}_{\text{LARGE}}$-cased
- Model parameters: 110M — 355M
- Input sequence length: 512 tokens padded
- Train-Val-Test split size: 10k, 2k, 8k
- Epochs: 4 — 5
- Batch Size: 7. [4]
- Batch Gradient Accumulation: 8
- Warmup steps: 500
- Weight decay: 0.01
- Seeds = [179, 50, 124, 253, 86]. 5 seeds = 1 seed per training run.

---

[1] `https://huggingface.co/transformers/`

[2] `https://huggingface.co/bert-base-cased`, `https://huggingface.co/bert-large-cased`, `https://huggingface.co/docs/transformers/model_doc/bloom`

[3] GPU time was typically limited to 6 hours or less which limited the number of epochs the PLMs could be trained to 4.

[4] Experiments showed 7 was the largest batch size possible given a 512 input sequence length with the RAM available, and is similar to that reported by Google on BERT github: `https://github.com/google-research/bert`

- Data seeds = [17, 38, 5, 91, 59] #n, n-6, n+6 for train-val-test seeds respectively
- Metric for best model: Accuracy
- Training + Validation time: 150mins (for 4 epochs)
- Inference time: 10mins (for 8k examples)

B.2 `emoBERT`

- Model: $\text{BERT}_{\text{BASE}}$-cased — $\text{BERT}_{\text{LARGE}}$-cased
- Model parameters: 110M — 355M
- Input sequence length: 512 tokens padded
- Train-Val-Test split size: 2.5k, 0.5k, 2k for GNE
- Epochs: 10
- Batch Size: 7
- Batch Gradient Accumulation: 8
- Warmup steps: 500
- Weight decay: 0.01
- Seeds = [179, 50, 124, 253, 86]. 5 seeds = 1 seed per training run.
- Data seeds = [17, 38, 5, 91, 59] #n, n-6, n+6 for train-val-test seeds respectively
- Metric for best model: $F1_\mu$
- Training + Validation time: 11mins (for 10 epochs)
- Inference time: 22s (for 2k examples)

B.3 BLOOM

- Model: BLOOM
- Model parameters: 560M
- Input sequence length: 512 tokens padded
- Train-Val-Test split size: 10k, 2k, 8k
- Epochs: 4 — 5
- Batch Size: 7
- Batch Gradient Accumulation: 8
- Warmup steps: 500
- Weight decay: 0.01
- Seeds = [179, 50, 124, 253, 86]. 5 seeds = 1 seed per training run.
- Data seeds = [17, 38, 5, 91, 59] #n, n-6, n+6 for train-val-test seeds respectively
- Metric for best model: Accuracy
- Training + Validation time: 83mins (for 5 epochs) on GPU: NVIDIA GeForce RTX 3090
- Inference time: 241s (4 mins) (for 8k examples) on GPU: NVIDIA GeForce RTX 3090

Appendix C

Emotions and Content in News Articles and *NEWSsynth*

Regarding the emotional content and journalistic content of news articles in *NEWSsynth*: Previous authors have specifically chosen the news domain because of its high emotional content (Strapparava and Mihalcea, 2007; Bostan et al., 2020). It is long established that different emotions lead to different actions (Spielberger, 1972) including what we write (Brand, 1985). Emotion can be exploited, for example "engagement based ranking" tends to favour content that evokes anger (Haugen, 2021). While some journalistic reporting is objective, opinion editorials (op-ed) are opinions pushing an agenda and, for example, tabloids tend to specifically exploit emotion. The 10k news articles in the *NEWSsynth* training split, for example, come from 150 online sources which also include: movie reviews and entertainment such as rollingstone.com, hollywoodlife.com, bollywoodhungama.com and mashable.com; and tabloids such as thedailymail.co.uk, dailystar.co.uk, thedailystar.net etc. which cover many types of news including journalism, op-eds, reviews, opinions etc. In short, *NEWSsynth* is not limited to non-emotional objective fact reporting, it contains a broad spectrum of journalistic styles and content.

Appendix D

Static Word Embeddings and Dimensional Emotions

We investigate the semantic relationships captured by static word embeddings (for words which represent emotions) and compared them to the affective relationships between those words - as described by Plutchik's Wheel of Emotion (Plutchik, 2001) shown in Figure D.1.

The GloVe word embeddings (Pennington et al., 2014) for Wikipedia 2014 and Gigaword 5 were converted to word2vec embeddings (Mikolov et al., 2013) and visualised using the Gensim word vector visualizer.[1] The scikit-learn principal component analysis tool reduced the word representations from 300D to 2D which allows us to visualise the simplified spatial relationship between word embedding vectors, as shown in Figure D.2. The comparative semantic relationship between words, or so called analogies, were calculated using vector addition and subtraction, and cosine similarity on the full 300D word embeddings. An analogy takes the form: `joy : sadness :: fear : X`. The analogy is calculated using word vectors: **X** = **sadness** + **fear** - **joy**. Cosine similarity is then used to find the closest word vector to **X**. The closest word vector represents the word representing the opposing word to fear. We can then determine if the opposing word to fear is the same as the opposing emotion to fear which, according to Plutchik's Wheel, should be `anger`. The calculation is: $X = argmin\{D_c(sadness + fear - joy)\}$ where X, sadness, fear, and joy are word embedding vectors, and $\mathrm{D_c}$ is cosine distance.

Analogy (using Semantic Relationship)	Plutchik's Model	Word Embedding (`X`)
1. joy : sadness :: fear : ____	anger	anger
2. anticipation : surprise :: disgust : ____	trust	exasperation

Table D.1: Analogy predictions of Plutchik's Model and Word Embedding. Given the analogy on the left, what does Plutchik's model specify, and what do the Word Embedding vector calculations return e.g. $X = argmin\{D_c(sadness + fear - joy)\}$.

The affective analogies from Plutchik's Wheel and the word analogies from

---

[1] `https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/materials/Gensimwordvectorvisualization.html`

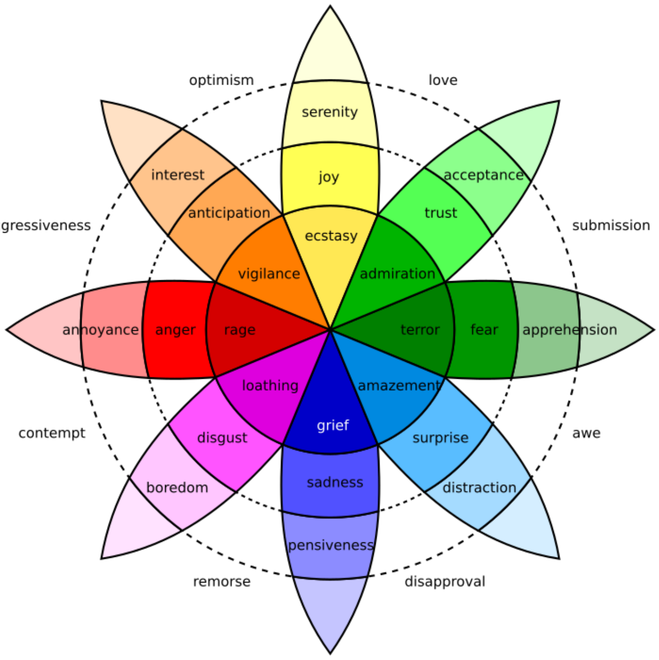


Figure D.1: Plutchik Wheel of Emotion. The middle ring includes Ekman's 6 emotions plus trust & anticipation. Similar emotions lie on adjacent spokes e.g. anger-disgust, while opposing emotions are placed on opposing spokes e.g. joy-sadness. Image from (Plutchik, 2001), downloaded from `https://en.m.wikipedia.org/wiki/File:Plutchik-wheel.svg`.

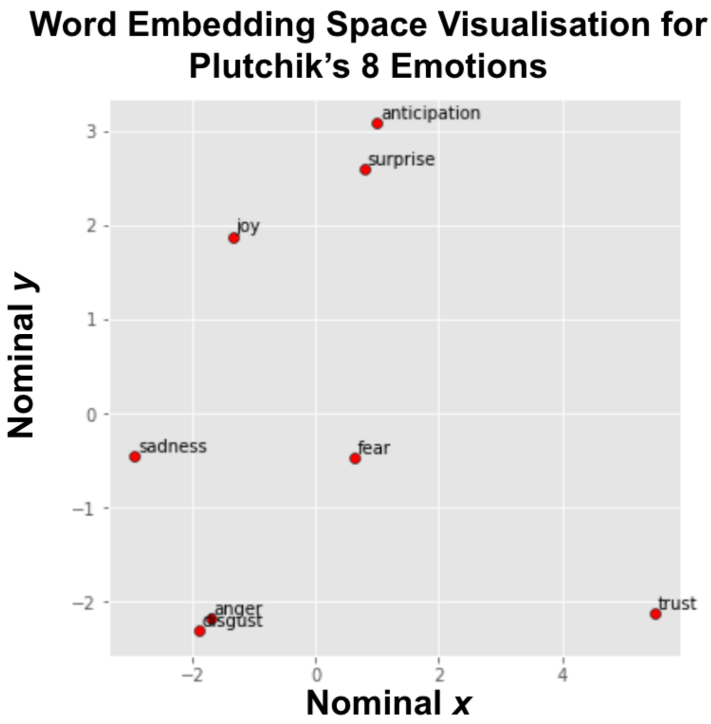


Figure D.2: Visualisation of the word embedding space for Plutchik's 8 Emotions indicating the semantic relationship between words. Note this is 300D reduced to 2D for illustration.

the word embeddings are shown in Table D.1 columns 2 & 3 respectively. We see that in example (1) the word embedding for the opposing word ‘anger’ agrees with the opposing emotion ‘anger’ as specified by Plutchik’s model. However, in example (2) we see the word embedding for the opposing word ‘exasperation’ is returned where Plutchik’s model specifies the opposing emotion ‘trust’. Thus, there is an inconsistency between Plutchik’s model and the word embeddings. We hypothesise this inconsistency is due to the inability of static (non-contextualised) word embeddings like GloVe and word2vec to fully represent the different senses of polysemous words. While ‘trust’ is an emotion it is also a financial instrument and the latter sense determined the word embedding location. This was confirmed by finding the 10 closest words to trust - all 10 were financial.

This experiment demonstrated some advantages (e.g. measuring word similarity using cosine similarity, capturing semantic relationships) and some disadvantages (e.g. cosine similarity weak at measuring antonyms, a single word embedding not adequately representing multiple senses of a word) of distributed non-contextualised (static) word embeddings. The experiment also highlighted the effect of training data in determining word embedding vectors, for example if more affective data were present we might expect the embedding for the word ‘trust’ to better reflect its affective sense (possibly at the cost of its financial sense). Language models like BERT and GPT-2 produce contextualised word embeddings which allow multiple representations for multiple senses of a word. Furthermore, such language models allow for transfer learning and fine-tuning which can improve performance on specific tasks such as emotion detection and synthetic text detection as we demonstrated in Chapters 3 and 5.

Appendix E

Human Cognition and Emotion & Sentiment Distribution in News Headlines

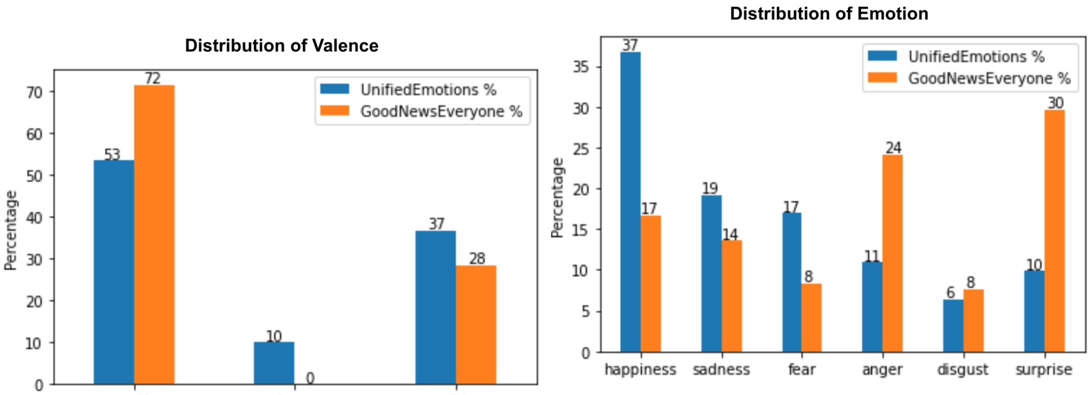
We discussed in Chapter 2 how surprise draws our attention, negativity has a greater impact and is more contagious than positive and neutral sentiments, and anger drives engagement. We compare GoodNewsEveryone (Bostan et al., 2020) to the UnifiedEmotions composite dataset which contains over 100k examples from 12 different datasets in multiple domains including news, blogs, twitter, questionnaires, conversations, and fairytales (Bostan and Klinger, 2018). Valence was obtained from the emotion labels following a similar mapping schema to (Chaumartin, 2007).[1] The distribution of valence and emotion in both datasets is illustrated in Figure E.1. The disparities in the distribution of both valence and emotions between news headlines and general media may be evidence of the Surprise Negativity Anger Paradigm (*SNAP*). They could also support the thesis that the triumvirate of surprise, negativity, and anger are tuned to exploit consumers' emotions, cognition and cognitive biases. Observations:

1. Number of negative news headlines (72%) is 250% that of positive news headlines (28%)
2. Percentage of negative news headlines is 19 points higher (72%) than general media (53%)
3. Percentage of positive news headlines is 9 points lower (28%) than general media (37%)
4. Surprise occurs 300% as often in news headlines as in general media.
5. Anger occurs more than 200% as often in news headlines as in general media.
6. Happiness occurs less than 50% as much in news headlines as in general media.[2]

---

[1] "surprise" from UnifiedEmotions was treated as "unknown" since its valence cannot be correctly determined. "Surprise" from GoodNewsEveryone was easily mapped since it was labelled as positive or negative in the dataset.

[2] UnifiedEmotions specifically included content from `Positive.News` to increase the news headlines with positive emotions.



(a) Distribution of valence for UnifiedEmotions and GoodNewsEveryone

(b) Distribution of emotions for UnifiedEmotions and GoodNewsEveryone. UE counts reported by (Bostan and Klinger, 2018)

Figure E.1: The distribution of valence and emotions in online news headlines (GoodNewsEveryone (Bostan et al., 2020)) and media in general (UnifiedEmotions (Bostan and Klinger, 2018))

Appendix F

Affective Profile - Statistical Significance Tests

Emotion n-gram Statistical Significance: The chi-squared (χ^2) test for independence [1] was used to determine the statistical significance of the difference between human and synthetic text - in the distribution of the number of emotion changes in emotion n-grams (Figures 6.1 to 6.4, from §6.3). The results are shown in Table F.1.

The null hypothesis and alternative hypothesis are stated as follows:

- Null hypothesis (H0) : the distribution of emotion changes in emotion n-grams is the same for human and synthetic text.
- Alternative hypothesis (HA) : the distribution of emotion changes in emotion n-grams is significantly different between human and synthetic text.

If the null hypothesis is true it indicates that any observed differences in the distributions are due to random chance rather than any significant distinction. If the null hypothesis is false, then the alternative hypothesis is taken to be true, which indicates that any observed differences in the distributions are due to real differences between human and synthetic text in how emotions change in emotion n-grams. A common threshold for accepting the null hypothesis is a p-value $p \geq .05$, whereas a p-value $p < .05$ indicates that the null hypothesis can be rejected and the alternative hypothesis can be accepted.

Table F.1 shows the χ^2 and the p-values for the number of emotion changes in emotion n-grams. All of the p-values are $p < .05$. Thus, we can reject the null hypothesis in all cases (i.e. n=2, 3, 4 & 5) and state that **the differences in the number of emotion changes in emotion n-grams between human and synthetic text are statistically significant**.

Sentiment n-gram Statistical Significance: Moving from emotion n-grams to sentiment n-grams, the chi-squared (χ^2) test for independence was used to determine the statistical significance of the difference between human and synthetic text - in the distribution of the number of sentiment changes in sentiment n-grams (Figures 6.10 to 6.13, from §6.5.3). The results are shown in Table F.2.

---

[1] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html`

n	χ^2	df	p-value	statistically significant ($p < .05$)
2	29.43	1	5.80x10^{-8}	yes
3	50.13	2	1.30x10^{-11}	yes
4	71.64	3	1.90x10^{-15}	yes
5	91.41	4	6.60x10^{-19}	yes

Table F.1: Chi-squared results for the distribution of number of emotion changes for emotion n-grams calculated from the *NEWSsynth* (train, 10k) dataset. n is the size of the n-gram. χ^2 is the chi-squared value. df is the degrees of freedom. p-value is the probability that the null hypothesis is true, a p-value $p < .05$ is considered statistically significant. All of the results are well below the .05 threshold, indicating that the difference between human and synthetic distributions of emotion changes in emotion n-grams is statistically significant.

n	χ^2	df	p-value	statistically significant ($p < .05$)
2	69.74	1	6.78x10^{-17}	yes
3	130.91	2	3.74x10^{-29}	yes
4	170.01	3	1.27x10^{-36}	yes
5	204.71	4	3.65x10^{-43}	yes

Table F.2: Chi-squared results for the distribution of number of sentiment changes for sentiment n-grams calculated from the *NEWSsynth* (train, 10k) dataset. n is the size of the n-gram. χ^2 is the chi-squared value. df is the degrees of freedom. p-value is the probability that the null hypothesis is true, a p-value $p < .05$ is considered statistically significant. All of the results are well below the .05 threshold, indicating that the difference between human and synthetic distributions of sentiment changes in sentiment n-grams is statistically significant.

Table F.2 shows the χ^2 and the p-values for the number of sentiment changes in sentiment n-grams. All of the p-values are $p < .05$. Thus, we can reject the null hypothesis in all cases (i.e. n=2, 3, 4 & 5) and state that **the differences in the number of sentiment changes in sentiment n-grams between human and synthetic text are statistically significant**. This is consistent with the statistical significance between human and synthetic text for emotions.

Emotion Change and Sentiment Change Statistical Significance A Kolmogorov-Smirnov (K-S) test [2] was conducted to measure the statistical significance between the human and synthetic distributions for 'percent of emotion maximum changes per article', from Figure 6.5. Similarly for sentiment, a K-S test was conducted to measure the statistical significance between the human

---

[2] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kstest.html`

and synthetic distributions for ‘percent of sentiment maximum changes per article’, from Figure 6.14. The results are shown in Table F.3.

Category	K-S statistic	p-value	statistically significant ($p < .05$)
Emotion	0.0536	1.15x10^{-6}	yes
Sentiment	0.0684	1.36x10^{-10}	yes

Table F.3: Kolmogorov-Smirnov Test results for the distribution of ‘percent of emotion/sentiment maximum changes per article’ calculated from the *NEWSsynth* (train, 10k) dataset. Both of the results are well below the $p < .05$ threshold, indicating that the difference between human and synthetic distributions is statistically significant - for both emotion and sentiment.

Table F.3 shows the K-S statistic and the p-values for the ‘percent of emotion/sentiment maximum changes per article’. The p-values are $p < .05$ for both emotion and for sentiment. Thus, we can state that: **the differences in the ‘percent of emotion maximum changes per article’ between human and synthetic text are statistically significant**; and that **the differences in the ‘percent of sentiment maximum changes per article’ between human and synthetic text are statistically significant**. These results are consistent with our results on statistical significance for emotion n-grams and sentiment n-grams.

Reproducibility We provide the following data to facilitate reproducibility of the statistical significance results.

Table F.6 contains the emotion change counts for emotion 2-grams to emotion 5-grams respectively. These tables are the basis for the χ^2 and p-value calculations shown in Table F.1. However, readers should note that the tables are normalized, and the normalized tables are used for calculating χ^2 and p-values. Normalization removes the difference between count totals from interfering with the calculation of statistical significance. The normalization re-scales the emotion change counts (columns) so that the human and synthetic totals (rows) are equal. The re-scaled row totals equal the mid-point of the original human and synthetic row totals. For example, Table F.4 shows the raw counts for emotion changes in emotion 2-grams, while Table F.5 shows the normalized counts.

Table F.7 contains the raw counts of the number of changes in sentiment n-grams. These are the values used in the n-gram distributions shown in Figures 6.10 - 6.13. The same process was uesd for sentiment n-grams as was used for emotion n-grams: the human and synthetic counts were normalized so that the counts were approximately the mid-point of the raw human and synthetic totals. Contingency tables, containing row and column totals, were then used to calculate the p-values which are shown in Table F.2.

Category	0 changes	1 change	Total
Human	56,759	84,900	141,659
Synthetic	36,075	56,495	92,570
Total	92,834	141,395	234,229

Table F.4: Distribution of number of emotion changes for emotion 2-grams calculated from the *NEWSsynth* (train, 10k) dataset. These are the raw figures which were shown as percentage values in Figure 6.1

Category	0 changes	1 change	Total
Human (Normalized)	46,925	70,190	117,115
Synthetic (Normalized)	45,640	71,474	117,114
Total	92,565	141,664	234,229

Table F.5: Normalized distribution of number of emotion changes for emotion 2-grams calculated from the *NEWSsynth* (train, 10k) dataset. These are the values from Table F.4 above, but normalized to the approximate mid-point (117,115) of the Human and Synthetic totals. This normalization ensures the sample size variation does not impact the chi-squared results, and consequently normalized tables were used to calculate χ^2 and p-values shown in Figure F.1.

Category	0 changes	1 change	2 changes	3 changes	4 changes	n-gram
Human	56,759	84,900	-	-	-	2-gram
Synthetic	36,075	56,495	-	-	-	2-gram
Human	28,557	52,360	55,743	-	-	3-gram
Synthetic	17,510	33,158	36,908	-	-	3-gram
Human	16,334	29,501	50,055	35,771	-	4-gram
Synthetic	9,682	17,878	31,446	23,582	-	4-gram
Human	10,113	16,782	35,683	40,363	23,722	5-gram
Synthetic	5,792	9,772	21,389	25,088	15,562	5-gram

Table F.6: Distribution of number of emotion changes for emotion n-grams calculated from the *NEWSsynth* (train, 10k) dataset. These are the raw figures which were shown as percentage values in Figures 6.1 - 6.4

Category	0 changes	1 change	2 changes	3 changes	4 changes	n-gram
Human	94,216	47,443	-	-	-	2-gram
Synthetic	60,050	32,520	-	-	-	2-gram
Human	66,251	49,052	21,357	-	-	3-gram
Synthetic	40,351	32,773	14,452	-	-	3-gram
Human	48,807	40,818	33,932	8,104	-	4-gram
Synthetic	28,490	26,075	22,620	5,403	-	4-gram
Human	36,975	31,752	38,112	16,148	3,676	5-gram
Synthetic	20,723	19,267	24,430	10,825	2,358	5-gram

Table F.7: Distribution of number of sentiment changes for sentiment n-grams calculated from the *NEWSsynth* (train, 10k) dataset. These are the raw figures which were shown as percentage values in Figures 6.10 - 6.13