


# Exploring Multimodal Sentiment Analysis Models: A Comprehensive Survey

Phuong Q. Dao 

Foreign Trade University, HCMC Campus, Vietnam  
Email: daoquocphuong.cs2@ftu.edu.vn

Thien B. Nguyen-Tat 

University of Information Technology, Ho Chi Minh City, Vietnam  
Vietnam National University, Ho Chi Minh City, Vietnam  
Email: thientnb@uit.edu.vn

Mark Roantree 

Insight Centre for Data Analytics,  
School of Computing, Dublin City University, Ireland  
Email: mark.roantree@dcu.ie

Vuong M. Ngo  

Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam  
Email: vuong.nm@ou.edu.vn

**Abstract**—The exponential growth of multimodal content across social media platforms, comprising text, images, audio, and video, has catalyzed substantial interest in artificial intelligence, particularly in multi-modal sentiment analysis (MSA). This study presents a comprehensive survey of 30 research papers published between 2020 and 2024 by eminent publishers such as Elsevier, ACM, IEEE, Springer, and others indexed in Google Scholar. Our analysis primarily focuses on exploring multimodal fusion techniques and features, with specific emphasis on the integration of text and image data. Additionally, the article offers an overview of the evolution, definition, and historical context of MSA. It delves into the current challenges and potential advantages of MSA, investigating recent datasets and sophisticated models. Furthermore, the study provides insights into prospective research directions. Notably, this review offers valuable recommendations for advancing research and developing more robust MSA models, thus serving as a valuable resource for both academic and industry researchers engaged in this burgeoning field.

## I. INTRODUCTION

The expression used for both sentiment analysis and emotion analysis is the same. In this sense, keyboard and mouse input is not necessary for the sentiment analysis techniques to function. With the use of novel modalities like speech, gesture, messaging, and facial expression, it assesses opinion, emotions, and polarity. The modalities are subjective and can be positive, negative, neutral, joyous, wonderful, and many other things. Taking "Mai dislikes the battery of the ABC phones" as an example. Mai expresses her opinion in this statement, and she has an unfavorable view regarding the "battery" of the "ABC phone." Sentiment analysis research involving the extraction of sentiments from voice, text, and facial expressions, have been the subject of extensive research in recent years. For example, Nguyen et al. in [1] used an ontological method to determine entity ratings. The authors then run trials using these entity scores to categorize opinions or detect opinion spam. In an another paper [2], Tran et al. used a popular machine learning method (SVM) and the WEKA library to build a Java web program for sentiment analysis of English comments on dresses, handbags, shoes, and rings. Their system

was trained on 300 comments and tested on 400 comments, and got 89.3% for precision. With models for Vietnamese sentiment analysis, in the paper [3], Dang et al. proposed hybrid deep learning models and results show hybrid models achieve higher accuracy on Vietnamese datasets. However, because of the ambiguity and adaptability of the data, researchers have faced many obstacles to effectively solving sentiment problems. A single piece of evidence is typically insufficient to yield reliable information. Sentiment analysis methods, for instance, are unable to precisely categorize user attitudes in the cases of irony, subjectivity, tone, and sarcasm. There are numerous ways to write the same text and the right class cannot be determined using only a single data source. Certain situations, personality, cultural, gender, and situational differences may cause a change in a person's facial expression. As a result, developing precise forecasts have become more difficult in recent years. These new challenges center on an innovative and interactive technology that integrates many information sources to forecast more precise classification and enhance computation accuracy and dependability.

People are posting photographs and text together to communicate their thoughts and feelings, thanks to the rise of social media and mobile devices. Multimodal Sentiment Analysis (MSA) is an emerging field of study that aims to analyze and identify sentiments using data from several modalities. Applications for comprehending multimodal sentiment include opinion mining, tailored advertising, affective cross-modal retrieval, and decision-making, among others. A field of study called multimodal sentiment analysis combines data from several sources to more accurately categorize people's thoughts and emotions. Numerous applications, such as social media, navigation tools, and human-to-human contact, have already been implemented utilizing the multimodal framework. These applications have already proven MSA's viability and significance.

In a survey paper published in 2021 [4], the authors highlighted that multimodal representation learning, multimodal alignment, and multimodal information fusion are the three

primary issues in MSA. Information fusion is a primary challenge because: (1) modalities may not have their information temporally aligned; (2) fusion models may find it challenging to leverage complementarity between modalities; and (3) noise types and intensities may differ among modal data. Additionally, they concentrated on the deep learning (DL) MSA fusion techniques, such as CNNs, RNNs, LSTMs, Transformers, and Attention Machines. For example, with fusion of text and image, it is called static MSA, CNN-based method is used and showed accuracy is nearly 91%. In this modal, with predicting the sentiment of visual information, text analytics uses a hybrid Convolutional Neural Network (CNN) and picture analytics uses a support vector machine (SVM) classifier that was trained using a bag-of-visual words. Another method of fusing text, image, audio, and video is called dynamic MSA. It uses an LSTM-based approach and extracts text features using textCNN, audio features using openSMILE, visual features using 3D-CNN, and shared information among multimodal features using context-sensitive LSTM. The modal accuracy is 80.3%. Especially, CNN+LSTM based method, RNN based method are used in MSA in conversation. In a separate survey [5], the authors discussed using recurrent neural networks in sentiment analysis in textual, visual and multimodal inputs. This work also discussed how Textual SA extracts huge semantic information using DL models, RNN, LSTM, and their derivatives are used to extract features from a series of visual frames, while Visual SA uses deep CNN to extract more abstract features. A third study [6], on Multimodal Sentiment Analysis research focused primarily on SA with only a brief discussion of MSA.

**Contribution.** In contrast to existing survey papers, our survey aims to provide a comprehensive overview of MSA datasets and techniques, with a specific emphasis on multimodal features, multi-modal fusion and offer insights into MSA based on text and image data. The contributions are outlined as follows:

- 1) we offer a thorough examination of datasets and tasks specifically within the field of MSA;
- 2) we review and analyse Multimodal features and Multimodal fusion;
- 3) we present the challenges and research future development in MSA, addressing issues such as the cross-modal interactions, context-dependent interpretations, and the prospect of constructing knowledge graph of multimodal representation for semantic analytics.

**Paper Structure.** The rest of this paper is organised as follows: in Section II, we detail the methodology employed in selecting the papers from the literature; in Section III, current research is analysed on MSA datasets, multimodal features, multimodal fusion and the analysis/modelling techniques applied; Section IV provides a discussion of the main findings from the survey; and finally in Section V, we present conclusions and discuss further research for this topic.

## II. METHOD

We focus on reviewing papers employing machine learning or DL models for multi-modal sentiment analysis systems. We

base our entire technique on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses [7] in order to find relevant studies. The following standards were applied for choosing research papers:

- Published by Elsevier, ACM, IEEE, Springer, or Elsevier, with Springer Nature, Science Direct, IEEE Xplore, or ACM Digital Library as their corresponding libraries.
- Published from 01/01/2020 to 31/03/2024.
- Written in English, not discriminating by geographical area and dataset language.
- Title or keywords or abstract of each paper has keywords: ("Multimodal" or "Multimedia") AND ("Sentiment Analysis" OR "Opinion Mining") AND ("Machine Learning" OR "Deep Learning" OR Classification). The keywords are used in the Boolean search query based on the form requirements of each library.

TABLE I  
THE NUMBER OF RELATED LITERATURES OF EACH PUBLISHER

Publisher with search criteria	The number of literatures		
	After downloading	After reviewing title and abstract	After reviewing full text
<b>Springer</b>	955	40	<b>10</b> (Res <sup>a</sup> 9, Sur <sup>b</sup> 1)
<b>Elsevier</b>	276	18	<b>8</b> (Res: 8, Sur: 0)
<b>IEEE</b>	215	14	<b>3</b> (Res: 3, Sur: 0)
<b>ACM</b>	226	19	<b>8</b> (Res: 7, Sur: 1)
<b>Total</b>	1,672	91	<b>29</b> (Res: 27, Sur: 2)
<b>Others with unlimited publisher and pub. year</b>			<b>4</b> (Res: 3, Sur: 1)
		<b>Final Review</b>	<b>33</b> (Res: 30, Sur: 3)

<sup>a</sup> Res: the number of research papers.

<sup>b</sup> Sur: the number of survey papers.

Table I, which displays 1,672 downloaded papers using the advanced search features of the publisher libraries, is produced using the search criteria mentioned above. Upon scrutinizing the titles and abstracts of every publication, we narrowed down the corpus to 91 pertinent articles. We carefully selected 29 papers from among these 91 publications, evaluating each one's whole text to determine which was most pertinent to our paper's objectives. We did an unrestricted publisher and publication year search on Google Scholar to make sure we didn't overlook any other pertinent publications.. This search led us to find an extra 4 relevant papers, resulting in a total of 33 relevant papers. This final set contained 30 research papers and 3 review papers for the MSA topic.

In Section I, we introduced and compared three review papers: [4], [5], and [6]. Subsequent sections were dedicated to a comprehensive analysis, categorization, and discourse on the 30 research papers, as delineated in detail in Table II. Furthermore, 30 papers could potentially meet the requirements of certain conference regulations.

### III. CLASSIFICATION AND ANALYSIS

#### A. Datasets

1) **CMU-MOSI**: Ninety-three carefully chosen YouTube videos on a variety of subjects make up this dataset. In order to catch genuine expressions, these videos have a lone speaker facing the camera. There are 89 speakers in all, 48 men and 41 women, who talk only in English while providing remarks and presentations. There are no restrictions on the setting, distance, or camera model. 2,199 subjective opinion segments with sentiment intensity values ranging from -3 to 3 were retrieved from these videos and annotated. For researchers looking into multi-modal sentiment analysis, the CMU-MOSI<sup>1</sup> dataset is a useful resource.

2) **CMU-MOSEI**: The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI<sup>2</sup>) dataset stands as the most extensive collection for sentiment analysis and emotion detection. It encompasses nearly 23,500 videos comprising phrase utterances, sourced from 1,000 diverse YouTube speakers, where each video maintains a balanced representation of gender. Phrases are selected randomly from various thematic and monologue videos, all meticulously punctuated and transcribed.

3) **T4SA**: The Twitter for Sentiment Analysis (T4SA<sup>3</sup>) dataset integrates 1 million tweets and 1.5 million photos, encompassing both textual and visual content. Sentiment annotation is divided into three classes: positive, negative, and neutral, although the neutral class annotations were somewhat ambiguous due to poor quality in the original tests. Sentiment labels assigned to the textual content were used to annotate the corresponding photographs but this has led to some confusion between the neutral class and both positive and negative classes.

4) **DFMSD**: The domain-free multimedia sentiment dataset (DFMSD<sup>4</sup>) is a recent release for textual and visual sentiment analysis, designed for uncontrolled online social media and outdoor environments. It was gathered using the Twitter Stream API, distinguishing itself from previous datasets by eschewing predefined criteria. To ensure unbiased annotation, questions and annotators were meticulously selected by three professional psychologists. DFMSD consists of 14,488 tweets, including 10,244 photos, with 46% positive, 33% negative, and 21% neutral tweets. Among the images, 47% are positive, 10% negative, and 43% neutral.

5) **Fakeddit**: A dataset comprising one million multimodal records of false news, sourced from Reddit<sup>5</sup> between March 2008 and October 2019, is publicly available. It includes text, images, metadata, and comments. The dataset offers three labeling schemes: 2-way (real or fake), 3-way (entirely real, entirely fake, or mixed), and 6-way (categorized as Satire, True, Fake, Misleading content, Manipulated content, False content, or Imposter content).

6) **MuSe-CaR**: For the MuSe-Wilder and MuSeSent sub-challenges in MuSe 2021, the MuSe-CaR<sup>6</sup> dataset was employed. This extensive multimodal (text, audio, and visual) dataset comprises 291 YouTube videos featuring 70 host speakers providing automotive reviews. It's important to note that the MuSe-CaR dataset exhibits several *in the wild* attributes. For instance, the videos feature: 1) predominantly non-frontal face angles; 2) audio recordings with ambient noise; 3) speaker utterances containing domain-specific expressions and colloquialisms; and 4) instances of face occlusion, missing faces, and widely varying backgrounds.

7) **MVSA**: The Multimodal Sentiment Analysis (MVSA) dataset is accessible to the public and was gathered through Twitter, where users share messages containing text, photos, hashtags, and other elements. Every text-image pair has a unique sentiment label associated with it. The MVSA<sup>7</sup> dataset is manually labeled as positive, neutral, or negative. MVSA-Single (MVSA-S) comprises 4869 image-text pairings labeled by a single annotator with a single sentiment label, while MVSA-Multiple (MVSA-M) comprises 19,598 image-text pairs identified by three annotators with three sentiment labels. These two portions of the MVSA dataset are separated.

8) **ReactionGIF**: ReactionGIF<sup>8</sup> is an affective dataset of 30,000 English-language tweets together with their corresponding GIF responses. This is a novel dataset that focuses on two-turn discussions. Each entry in the dataset consists of a GIF reply in response to a merely textual root post. These tweets have all been categorized with the appropriate reaction category. Based on innovative reaction-to-emotions mapping, this category—which conveys a strong affective signal—has been used to provide an appropriate sentiment and emotion label to the item. We retrieved the reaction GIFs, gathered the metadata related to each tweet, and fetched the tweets using Tweepy.

#### B. Multimodal Features

In MSA, feature engineering, also known as feature extraction, is a crucial field for obtaining features from unprocessed data. First, we take a range of properties out of three widely used modalities: text, visual, and audio. The combination of two or more features for the SA is included in the hybrid features. Since individual features may not always be relevant, hybrid features can be used to create a high-rate sentiment classification method.

For instance, combining voice and visual elements improves the visualization of emotion analysis. Hazarika et al. employed Modality-Invariant and -Specific representations in their proposed framework MISA in [8] to collect Invariant and Specific information, which they then combined to predict emotional states. The cross-attention map and forget gate mechanism are coupled by the authors in [15], which is useful to obtain appropriate interaction among various modality pairings and

<sup>1</sup><http://multicomp.cs.cmu.edu/resources/cmu-mosi-dataset/>

<sup>2</sup><http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/>

<sup>3</sup><http://www.t4sa.it/>

<sup>4</sup><https://mcrlab.net/datasets/dfsmd/>

<sup>5</sup><https://fakeddit.netlify.app/>

<sup>6</sup><https://www.muse-challenge.org/>

<sup>7</sup><https://mcrlab.net/research/mvsa-sentiment-analysis-on-multi-view-social-data/>

<sup>8</sup><https://paperswithcode.com/dataset/reactiongif>

TABLE II  
DATASETS, MULTIMODAL FEATURES (MF), MULTIMODAL FUSION METHODS (MFM), MODELS AND ACCURACY OF THE 30 CURRENT RESEARCH

No	Study	Year	Datasets <sup>1</sup> (Name)	MF <sup>2</sup>	MFM	Models and Accuracy
1.	Hazarika et al. [8]	2020	3rd+O CMU-MOSI,CMU-MOSEI	T+V+A	Late	MISA (Acc: 83.4, 85.5, 70.61)
2.	Dong Zhang et al. [9]	2020	3rd+O CMU-MOSI,CMU-MOSEI	T+V+A	Late	Bi-modal (acc: 70.9, F1: 70.9), Tri-modal (acc: 71.2, F1: 71.2)
3.	Sun et al. [10]	2021	3rd+O MuSe-CaR	T+V+A	Late	temporal model (Acc: 0.5549)
4.	Dong Liu et al. [11]	2021	Self+NO	V+A	Early	(SIFT, CNN) for Face + LIBSVM for fusion (acc: 90.89%)
5.	K.Vasanth et al. [12]	2022	Self+NO	T+V+A	Early	N/A
6.	Garcia et al. [13]	2022	Self+NO	T+V+A	Late	HERA framework
7.	Palani et al. [14]	2022	3rd+O Politifact,Gossipcop	T+V	Early	CB-Fake (Acc: 0.93)
8.	Jiang et al. [15]	2023	3rd+O CMU-MOSI,CMU-MOSEI	T+V+A	Late	CMGA (Acc: 53.03)
9.	Wu et al. [16]	2023	3rd+O CMU-MOSI,CMU-MOSEI	T+A	Late	HG-BERT model (Acc: 83.82)
10.	Perti et al. [17]	2023	Self+NO	T+V	Late	Auc: 0.8420
11.	Grosz et al. [18]	2023	3rd+NO	T+V+A	Late	Auc: 0.8420
12.	Sun et al. [19]	2023	3rd+O CMU-MOSI,CMU-MOSEI	T+V+A	Late	GEAR (Acc: 84.39%)
13.	Bryan Smith et al. [20]	2023	Self+NO	T+V	Late	CLIP-based model (Precision: 0.624, Recall: 0.607)
14.	Meena et al. [21]	2023	3rd+O CK+,FER2013,JAFFE	V	N/A	CNN-based Inception-v3 (Acc: 99.57%, 73.09%, 86%)
15.	Nadeem et al. [22]	2023	3rd+NO	T+V	Early	Proposed SSM (Acc: 96.90)
16.	Tong Zhu et al. [23]	2023	Self+NO	T+V	Late	ITTN (Acc: 0.7519)
17.	Alzamzami et al. [24]	2023	3rd+O T4SA,FER-2013,DFMSD	T+V	Late	parallel (Acc: 0.82)
18.	Uppada et al. [25]	2023	3rd+O Fakeddit	T+V	Late	Fine-tuned BERT and fine-tuned Xception (Acc: 91.94%)
19.	Fu et al. [26]	2023	3rd+O CMU-MOSI,CMU-MOSEI	T+V+A	Late	LMR-CBT
20.	Jain et al. [27]	2023	Self+NO	T+V+A	Late	MTCNN Model, NLP model, SVM model, Google API
21.	Volkanovska et al. [28]	2023	3rd+O	T+V	N/A	using NLP tools to enrich corpus (meta)data
22.	Huiyu Wang et al. [29]	2023	3rd+O MVSA-single,HFM	T+V	Early	BERT + BiLSTM, CNN and CBAM attention
23.	Aggarwal et al. [30]	2023	3rd+O ReactionGIF	T+V	Late	BERT, OCR, VGG19
24.	Shi et al. [31]	2024	3rd+O CMU-MOSI,CMU-MOSEI	T+V	Late	CoASN model based on CMMC and AMAG
25.	Zheng et al. [32]	2024	3rd+O CMU-MOSI,CMU-MOSEI	T+V+A	Late	DJMF framework
26.	Ayetiran et al. [33]	2024	Self+NO	T+V+A	Early	Acc: 0.94
27.	Lu et al. [34]	2024	3rd+O CMU-MOSI,CMU-MOSEI	T+V+A	Late	sentiment-interactive graph (Acc: 86.5%, 86.1%)
28.	Yifeng Wang et al. [35]	2024	3rd+O CMU-MOSI,CMU-MOSEI	T+V+A	Late	MTAMW (multimodal adaptive weight matrix)
29.	Wang et al. [36]	2024	3rd+O Twitter-2015, Twitter-2017	T+V	Late	GLFFCA + BERT (Acc: 74.07%, 68.14%)
30.	Kumar et al. [37]	2024	Self+NO	V+A	Late	ParallelNet (Acc: 89.68%)

<sup>1</sup> 3rd: third party, Self: build-self, O: open, NO: no open

<sup>2</sup> T: text, V: visual, A: audio.

retain the instrumental signals to represent the multimodal input. In [16], the authors present a hierarchical multi-head self attention mechanism that uses a progressive number of heads to extract features by utilizing the differences in feature extraction capabilities of different BERT network layers. Moreover, the other authors employed an n-gram-based word embeddings approach [17] to get the machine-level word representations, the idea of N-gram-based word embeddings was used to create the vector representation of tweets from Twitter and discovered that the ensemble technique yielded the best results.

The forward sequential selection (FSS) technique proposed in [18], selects the most informative feature iteratively and adds it to the list of optimal features. The authors of [19] pointed out that while previous work on multimodal sentiment analysis (MSA) uses multimodal data for prediction, it inevitably suffers from fitting the false correlations between sentiment labels and multimodal features. For instance, if the majority of the films in a dataset have positive labels, the model will rely on these correlations for prediction even though "blue background" is not a sentiment-related attribute. To address this problem, the authors constructed a general debiasing MSA process in their study.

### C. Multimodal Fusion Methods

Multimodal data are more informative than single-modal data because they depict objects from several angles. It's possible that different modalities of data information complement one another well. Significant and challenging difficulties in multimodal sentiment analysis are: maintaining the modalities'

semantic integrity, producing a good fusion between modalities, and fusing data features between modalities. It can be summed up as feature-based multimodal fusion in the early stages and decision-based multimodal fusion in the latter stages, depending on the various types of modal fusion.

1) *Early Fusion*: Shallow fusion is carried out by early feature-based multimodal fusion algorithms following the first feature extraction phase. At the shallow level of the model, combining the characteristics of several modalities is equal to integrating the features of various single modalities into the same parameter space. Features may contain a large amount of duplicate information because different modalities have distinct information. To get rid of the extraneous data, dimensionality reduction techniques are usually required. To finish feature extraction and prediction, features that have undergone dimensionality reduction are added to the model.

Early feature fusion aims to include input data from several modalities and start feature modeling as soon as feasible. Unfortunately, the method of integrating numerous distinct parameter spaces in the input layer frequently fails to produce the intended effect since different modes have different parameter spaces. This kind of model can handle robust and accurate multimodal sentiment analysis tasks. However, due to their very intricate structures, these models require a large amount of training data in order to perform well, and training takes longer. The attention mechanism in [22] is applied by the authors through the use of the multimodal fusion module. This mechanism gives more weight to important features such as the physical and semantic



properties of the picture and text.

2) *Late Fusion*: Information from many modalities is combined using a decision level fusion approach. The practice of training models independently on input from many modalities in order to combine outputs from numerous modalities into the final result is known as decision-level fusion. Typically, learnable models, majority voting, averaging, and weighing are used in decision fusion to combine modalities, where variants are often lightweight and flexible. In the event that any modality is unavailable, decision making uses the remaining modalities.

3) *Text and Image Fusion*: The self-attention technique was employed by the authors in [16] to facilitate multi-mode fusion. Using visual information, the CLIP-based model in [20] takes into account both text and images, which can help close the gap between AI and human raters. The authors of [23] state that as multimodal sentiment analysis analyzes the latent alignment information between picture regions and text words, the relationship between image affective regions and the related text is crucial. The suggested cross-modal gating module can be used to further filter the negative effects of misaligned region-word pairs. Image regions and sentence words are intended to be aligned in the embedding space using the Cross-modal Alignment Module. This module finds the most appropriate textual information for each region by attending to sentence words in relation to each image region using a cross-modal attention technique. The Cross-modal Gating Module deconstructs messages that flow between the two modalities and produces the most relevant word-level data for each location.

In [25], the authors presented a new framework in which pre-trained Xception models are used to analyze visual data, which has two properties relating to image manipulation and image polarity, while pre-trained BERT is used to assess textual data. In order to categorize the social postings as Real or Fake, the features gathered from these branches are ultimately fused using fusion models such as Concatenate and Maximum. In [29], the authors employed a novel feature extractor called BERT + BiLSTM to identify long-distance connections in sentences and to take into account the location data of input sequences in order to provide richer text features. After splicing text features and picture features, CNN and Convolutional Block Attention Module (CBAM) were applied to improve the feature representation ability by removing redundant data and making the network focus more on the relationship between text and picture attributes.

#### D. MSA Frameworks and Methodologies

Figure 1 presents an MSA framework, illustrating both feature-level and early fusion. The first step, Multimodal Data, represents the inputs to the framework, the second step is a process of Feature extraction from inputs, while the subsequent step depends on type of multimodal fusion method. The final step will always be the result of sentiment analysis.

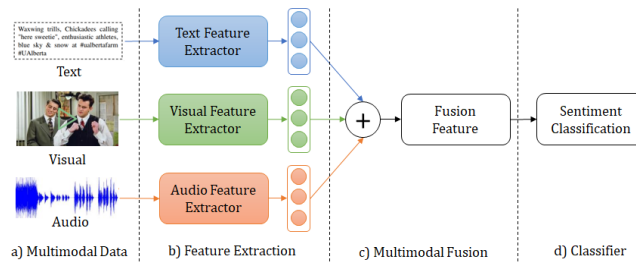


Fig. 1. MSA framework both feature-level and early fusion

A semi-supervised learning approach to multi-modal sentiment categorization was proposed by the authors in [9]. This approach can capture both independent knowledge inside a single modality and interaction knowledge among distinct modalities. In [16], gate channel is employed in place of the Feed Forward layer in the BERT model to realize noise filtering, and the authors used hierarchical multi-head self-attention to realize hierarchical extraction of data features. The optimized BERT model with a gate channel and hierarchical multi-attention mechanism is a newly proposed framework called the HG-BERT model. A tensor fusion model based on self-attention realizes information exchange between models as regarding feature fusion.

In [18], the authors suggested self-supervised pre-trained models, relevant Sub-spaces of BERT, Wav2Vec 2.0, ELECTRA and ViT Embedding for prediction. In addition, The learning rate, the number of iterations, the number of hidden layers, the number of hidden units, and the selection of the activation function are just a few of the hyper-parameters that the authors took into consideration when developing their DL or transfer learning model in [21]. In [22], the Hyperbolic Hierarchical Attention Network was developed, a model initially trained with textual data, which combined news title and body, in order to identify the hidden patterns of fake news. The article's summary and title are the subjects of the second comparison. The similarity between the two illustrates how a news headline and a summary of its content relate to each other. Thirdly, the semantic similarity between written and visual content is ascertained by extracting image semantic features and comparing them to the summary.

In [24], the authors employed learning performance, measured by the F-score and accuracy, to show significant improvements when the threshold-moving technique and the transformer architecture are combined. In [26], three components make up the new CB-Transformer framework: global self-attention representations, cross-modal feature fusion, and local temporal learning. The transformer encoder and the residual-based cross-modal fusion, which are represented by TransEncoder and CrossModal, are the two key elements of this module. In [30], the system is split into two training paths: the first one uses text data for perceived sentiment analysis, while the second path uses video data for induced sentiment analysis. In [33], a unified DL framework based on an inter-modal attention mechanism is developed by the authors using the unified modalities.

## IV. RESULTS AND DISCUSSION

### A. Publication year

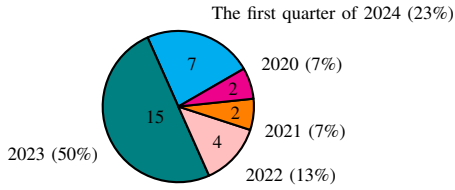


Fig. 2. The number of prior research were published through years

Figure 2 shows the number of the research papers as increasing year on year. Specifically, MSA could be considered essential research in 2023 with approximately 50% of the selected research papers, due to the rapid growth of social medias and MSA as a potential tool to detect and interpret public sentiment.

### B. Datasets

Table III provides an overview of the datasets utilized in the selected literature. Approximately 63% of the papers utilized generated third-party and open data, accessible to the research community. A slightly smaller proportion, around 30%, relied on self-generated and closed data, not publicly available. A minority of papers, roughly 7%, utilized datasets generated by a third-party but kept private.

TABLE III  
THE NUMBER OF PRIOR RESEARCH WERE CLARIFIED IN TERMS OF USED DATASETS, FUSION METHODS, AND MODELS

Terms	Classification and Percentage (papers)		
	3rd+NO	Self+NO	3rd+O
Datasets	7% (2p)	30% (9p)	63% (19p)
Fusion Methods	Early	Late	N/A
	20% (6p)	73% (22p)	7% (2p)
Models	Improve	Experiment	New model
	17% (5p)	17% (5p)	66% (20p)

Table II serves as a comprehensive overview of two distinct types of data sources: Self (self-built by authors) and 3rd (third-party supplies). Beyond merely identifying the source, the table offers valuable insights into two critical properties of these data sources, namely NO (no Open for access) and O (Open for access). This classification provides an important context for understanding the availability and accessibility of the datasets used in the discussed multimodal sentiment analysis studies. One of key components of MSA is the dataset. A multi-modal sentiment analysis model with excellent generalization and widespread application could be trained on a vast and diverse dataset, considering the diversity of languages and ethnicities in many nations. Furthermore, researchers must label multimodal datasets more precisely because they now have low annotation accuracy and have not reached absolute continuous values. The majority of multimodal data available now only include text, voice, and visual modalities; they do not include modal

information paired with physiological signals like pulses and brain waves.

### C. Fusion Methods

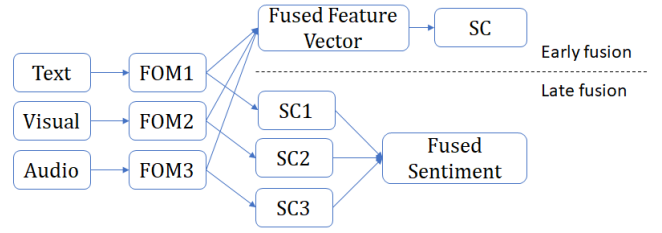


Fig. 3. Fusion methods

Table III also outlines fusion methods, including Late (~73%) and Early (~20%). A much smaller percentage of papers do not address fusion methods (~7%). Early and Late fusions are illustrated in Figures 3 with some components such as Feature of Modality (FOM) and Sentiment Classifier (SC), respectively. Especially, The text and image fusion mainly involves late fusion, where each modal input is handled by a model. During the decision phase, combination technologies are utilized to generate an output..

### D. Models

In terms of model development, the models in MSA include 'new-model' (~66%), 'improvement based-line' (~17%), 'experiment' (~17%), as shown in Table III. Recent years have seen the development of almost entirely new models, with a particular emphasis on multimodal fusion and features.

A new framework called the HG-BERT model combines a gate channel and hierarchical multi-attention mechanism to optimize the BERT model. Wav2Vec 2.0, ELECTRA, ViT Embedding, and BERT's relevant Sub-spaces are self-supervised pre-trained models that are used for prediction. The new proposed Hyperbolic Hierarchical Attention Network model showed semantic similarity between written and visual content when compared to the summary. Continuously, a new proposed CB-Transformer framework: global self-attention representations, cross-modal feature fusion, and local temporal learning. The transformer encoder and the residual-based cross-modal fusion, which are represented by TransEncoder and CrossModal, are the two key elements of this module. The other system consists of two training paths: the first path uses textual data for perceived sentiment analysis, and the second path uses video data for induced sentiment analysis. A unified DL framework is constructed based on the unified modalities, an inter-modal attention mechanism. Furthermore, the BERT model was nearly utilized for text classification, while CNN was employed for image classification in text-image fusion. Additionally, a newly suggested model, such as CNN and CBAM, may concentrate on the relationship between text and image.

MSA models need to be examined further in light of increasing accuracy or other metrics, or they might be used to create a new modal based on sophisticated temporal models and fusion techniques. Regarding the features, models could be created with time-dependent interactions in mind. They could also make use of social context features like user profiles and propagation patterns, as well as invariant feature learning to help learn how to better distinguish biased features and facilitate bias estimation.

Network training can be used to get parameters for the feature's distribution, combine features from different sources to create more relevant multimodal characteristics, and explore additional feature types that could help learn more about online sentiment behavior. Transfer learning approaches are a focus model technique that has gained popularity recently. Additionally, MSA models have the potential to track a user's credibility by utilizing metadata and comments in conjunction with user-related data. Additionally, they can leverage adversarial learning and knowledge graphs to enhance the effectiveness of unified inter-modal attention approaches. In addition, models can investigate the relationship between the relative importance of modalities and capture complicated relations. The interpretability of emotion identification in the aforementioned modalities is investigated through additional methodologies, crossmodal linkages, and filtering mechanisms.

## V. CONCLUSION

The importance of multimodal sentiment analysis approaches has been acknowledged by scholars across multiple domains, positioning it as a primary area of study in the domains of features and fusion. We go into great detail in this review to cover the definition, history of research, and evolution of multimodal sentiment analysis, among other topics. We also present a summary of frequently used benchmark datasets in Table II, and we examine and contrast the most current iterations of multimodal sentiment analysis models. Finally, we discuss the difficulties that the multimodal sentiment analysis field faces and speculate about potential future advancements, like the use of transfer learning techniques to enhance certain model metrics. Moreover, due to redundant information across modalities, the fusion process remains a significant challenge. Although several frameworks with optimized classifiers have been proposed, no single model can be universally applicable to all features; its effectiveness relies on the specific context of its application. Nevertheless, the MSA holds the promise of addressing these challenges in the future.

## ACKNOWLEDGEMENT

This research was conducted with the financial support of Science Foundation Ireland [12/RC/2289\_P2] at Insight the SFI Research Centre for Data Analytics at Dublin City University.

- [1] H. L. Nguyen, H. T. N. Pham, and V. M. Ngo, "Opinion spam recognition method for online reviews using ontological features," *CoRR*, vol. abs/1807.11024, 2018. [Online]. Available: <http://arxiv.org/abs/1807.11024>
- [2] T. N. T. Tran, L. K. N. Nguyen, and V. M. Ngo, "Machine learning based english sentiment analysis," 2019. [Online]. Available: <https://arxiv.org/abs/1905.06643>
- [3] C. N. Dang, M. N. Moreno-García, F. De la Prieta, K. V. Nguyen, and V. M. Ngo, "Sentiment analysis for vietnamese – based hybrid deep learning models," in *Hybrid Artificial Intelligent Systems*, P. García Bringas, H. Pérez García, F. J. Martínez de Pisón, F. Martínez Álvarez, A. Troncoso Lora, Á. Herrero, J. L. Calvo Rolle, H. Quintián, and E. Corchado, Eds. Cham: Springer Nature Switzerland, 2023, pp. 293–303.
- [4] L. Xudong, et al., "Multimodal sentiment analysis based on deep learning: Recent progress," in *ICEB 2021 Proceedings*, 2021.
- [5] J. V. Tembhurne and T. Diwan, "Sentiment analysis in textual, visual and multimodal inputs using recurrent neural networks," *Multimedia Tools and Applications*, vol. 80, no. 5, pp. 6871–6910, Feb 2021.
- [6] R. Das and T. D. Singh, "Multimodal sentiment analysis: A survey of methods, trends, and challenges," *ACM Comput. Surv.*, vol. 55(13s), 2023.
- [7] V. M. Ngo, et al., "Investigation, detection and prevention of online child sexual abuse materials: A comprehensive survey," in *Proceedings of the 16th IEEE-RIVF*, 2022, pp. 707–713.
- [8] D. Hazarika, et al., "Misa: Modality-invariant and -specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, p. 1122–1131.
- [9] D. Zhang, et al., "Multi-modal sentiment classification with independent and interactive knowledge via semi-supervised learning," *IEEE Access*, vol. 8, pp. 22 945–22 954, 2020.
- [10] L. Sun, et al., "Multimodal emotion recognition and sentiment analysis via attention enhanced recurrent model," in *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, 2021, p. 15–20.
- [11] D. Liu, et al., "Speech expression multimodal emotion recognition based on deep belief network," *J. of Grid Computing*, vol. 19(2), p. 22, 2021.
- [12] K. Vasanth, et al., "Dynamic fusion of text, video and audio models for sentiment analysis," *Procedia Computer Science*, vol. 215, pp. 211–219, 2022, the 4th Int. Conf. on IDCT&A.
- [13] J. M. Garcia-Garcia, et al., "Building a three-level multimodal emotion recognition framework," *Multimedia Tools and Applications*, vol. 82, no. 1, pp. 239–269, 2022.
- [14] B. Palani, et al., "Cb-fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and bert," *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 5587–5620, 2022.
- [15] M. Jiang and S. Ji, "Cross-modality gated attention fusion for multimodal sentiment analysis," 2022.
- [16] J. Wu, et al., "A optimized bert for multimodal sentiment analysis," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, no. 2s, 2023.
- [17] A. Perti, et al., "Cognitive hybrid deep learning-based multi-modal sentiment analysis for online product reviews," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 2023.
- [18] T. Grósz, et al., "Discovering relevant sub-spaces of bert, wav2vec 2.0, electra and vit embeddings for humor and mimicked emotion recognition with integrated gradients," in *Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation*, 2023, p. 27–34.
- [19] T. Sun, et al., "General debiasing for multimodal sentiment analysis," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, p. 5861–5869.
- [20] L. Bryan-Smith et al., "Real-time social media sentiment analysis for rapid impact assessment of floods," *Computers & Geosciences*, vol. 178, p. 105405, 2023.
- [21] G. Meena et al., "Sentiment analysis on images using convolutional neural networks based inception-v3 transfer learning approach," *Int. J. of Information Management Data Insights*, vol. 3, no. 1, p. 100174, 2023.
- [22] M. I. Nadeem et al., "Ssm: Stylometric and semantic similarity oriented multimodal fake news detection," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 5, p. 101559, 2023.
- [23] T. Zhu, et al., "Multimodal sentiment analysis with image-text interaction network," *IEEE Transactions on Multimedia*, vol. 25, pp. 3375–3385, 2023.

- [24] F. Alzamzami and A. E. Saddik, "Transformer-based feature fusion approach for multimodal visual sentiment recognition using tweets in the wild," *IEEE Access*, vol. 11, pp. 47 070–47 079, 2023.
- [25] S. K. Uppada, et al., "An image and text-based multimodal model for detecting fake news in osn's," *Journal of Intelligent Information Systems*, vol. 61, no. 2, pp. 367–393, Oct 2023.
- [26] Z. Fu, et al., "Lmr-cbt: learning modality-fused representations with cb-transformer for multimodal emotion recognition from unaligned multimodal sequences," *Frontiers of Comp. Science*, vol. 18, no. 4, 2023.
- [27] R. Jain, et al., "Real time sentiment analysis of natural language using multimedia input," *Multimedia Tools and Applications*, vol. 82, no. 26, pp. 41 021–41 036, Nov 2023.
- [28] E. Volkanovska, et al., "The insightsnet climate change corpus (iccc)," *Datenbank-Spektrum*, vol. 23, no. 3, pp. 177–188, Nov 2023.
- [29] H. Wang, et al., "Exploring multimodal sentiment analysis via cbam attention and double-layer bilstm architecture," 2023.
- [30] A. Aggarwal, D. Varshney, and S. Patel, "Multimodal sentiment analysis: Perceived vs induced sentiments," 2023.
- [31] P. Shi, et al., "Deep modular co-attention shifting network for multimodal sentiment analysis," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 20, no. 4, jan 2024.
- [32] Y. Zheng et al., "Djmf: A discriminative joint multi-task framework for multimodal sentiment analysis based on intra- and inter-task dynamics," *Expert Systems with Applications*, vol. 242, p. 122728, 2024.
- [33] E. F. Ayetiran and Özlem Özgöbek, "An inter-modal attention-based deep learning framework using unified modality for multimodal fake news, hate speech and offensive language detection," *Information Systems*, vol. 123, p. 102378, 2024.
- [34] Q. Lu et al., "Coordinated-joint translation fusion framework with sentiment-interactive graph convolutional networks for multimodal sentiment analysis," *Information Processing & Management*, vol. 61, no. 1, p. 103538, 2024.
- [35] Y. Wang et al., "Multimodal transformer with adaptive modality weighting for multimodal sentiment analysis," *Neurocomputing*, vol. 572, p. 127181, 2024.
- [36] S. Wang, et al., "Aspect-level multimodal sentiment analysis based on co-attention fusion," *Int. J. of Data Science and Analytics*, 2024.
- [37] P. Kumar et al., "Interpretable multimodal emotion recognition using hybrid fusion of speech and image data," *Multimedia Tools and Applications*, vol. 83, no. 10, pp. 28 373–28 394, 2024.