Research Article

Multimodal Sentiment Analysis: A Survey

Songning Lai · **Haoxuan Xu** · **Xifeng Hu** · **Zhaoxia Ren** · **Zhi Liu**

Received: date / Accepted: date

Abstract Multimodal sentiment analysis has emerged as a critical area of research in the field of artificial intelligence. With the recent advancements in deep learning, the technology has reached new heights. It holds immense potential for applications and research, making it a popular topic of study. This review focuses on the definition, background, and development of multimodal sentiment analysis. It also covers recent data sets and advanced models, highlighting the challenges and future prospects of the technology. Finally, the article concludes with possible research directions for future development.

Keywords Multimodal Sentiment Analysis · Multimodal Fusion · Affective Computing · Computer Vision

1 Introduction

Emotion is the subjective reaction of the organism to the external value relationship [1, 2], and it is a form that the organism shows when it encounters a specific situation.

Humans have an extremely powerful capacity for sentiment analysis, and making it available to artificial agents is a current and highly exciting area of research [3].

Songning Lai (202000120172@mail.sdu.edu.cn), Haoxuan Xu (202020120237@mail.sdu.edu.cn), Xifeng Hu (2019425440@mail.sdu.edu.cn) and Zhi Liu (liuzhi@sdu.edu.cn) are with the School of Information Science and Engineering, Shandong University, Qingdao, China.
Zhaoxia Ren (renzx@sdu.edu.cn) is with Assets and Laboratory Management Department, Shandong University, Qingdao, China.
Corresponding author: Zhi Liu and Zhaoxia Ren. (liuzhi@sdu.edu.cn,renzx@sdu.edu.cn).

Sentiment analysis refers to the analysis of sentiment polarity through available information [4, 5]. With the rapid development of many fields such as artificial intelligence, computer vision, and natural language processing, it is gradually becoming possible for artificial agents to implement sentiment analysis. At the same time, sentiment analysis is also an interdisciplinary research area that includes the intersection of computer science, psychology, social science, and other disciplines [6–8].

Over the past decades, scientists have also been working to empower AI agents with sentiment analysis capabilities. This makes AI more like a "human". This is also a key component of human-like AI.

Sentiment analysis has extremely significant research value [9–12]. With today's rapid development of the Internet, the amount of data has exploded. Vendors try to use this evaluative data (reviews, review videos, and so on) to improve their products. Not only that, but there are countless research values, such as lie detection, interrogation, entertainment, etc. The application and research value of sentiment analysis will be elaborated in the following sections.

In the past, sentiment analysis has mostly focused on a single modality (visual modality, speech modality, or text modality) [13]. Sentiment analysis in a single modality has also been developed to some extent. Text-based sentiment analysis [14–16] has gone a long way in NLP. Vision-based sentiment analysis pays more attention to human facial expressions [17] and movement postures. Sentiment analysis based on speech mainly extracts features such as pitch, timbre and temperament in speech for sentiment analysis [18]. With the development of deep learning, these three modalities have gained some foothold in sentiment analysis.

Numerous researchers still use a single modality for sentiment analysis [19–22]. For a single modality, the

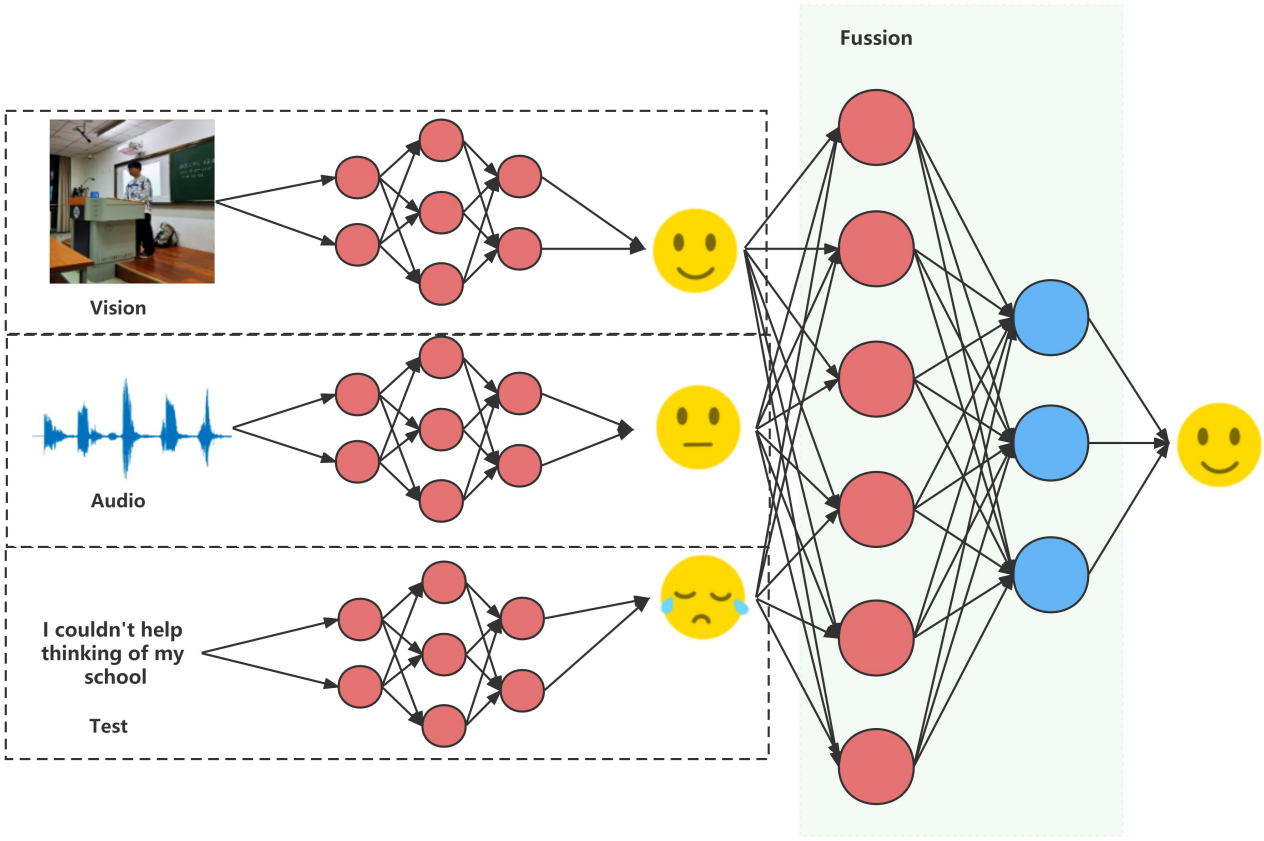

Fig. 1 Figure shows that only a single modality is considered, the results are all limited and it is difficult to make a correct analysis of the emotion of an action. Deeper emotional polarity can be obtained if information from multiple modalities is combined and analyzed together.

emotional information it contains is limited and incomplete, with numerous limitations. As shown in Fig. 1, if only a single modality is considered, the results are all limited and it is difficult to make a correct analysis of the emotion of an action. Deeper emotional polarity can be obtained if information from multiple modalities is combined and analyzed together.

Researchers have gradually realized the need for multi-modal sentiment analysis, and many multi-modal sentiment analysis models have emerged to accomplish this task. Text features dominate and play a key role in the analysis of deep emotions [23]. Visual modality extraction of expression and pose features can effectively aid text sentiment analysis and judgment [24]. On the one hand, speech modality can extract text features, and on the other hand, speech tone can be recognized to reveal the state of text at each time point [25]. Fig. 2 shows the model architecture for the more classical multi-modal sentiment analysis. The overall architecture consists of three parts: one part for feature extraction of individual modalities, one part for fusion of features of each modality, and one part for sentiment analysis of the fused features. These three parts are very important, and researchers have begun to optimize these three parts one by one [26].

In this review, we have given a small overview of the large field of multimodal sentiment analysis. Summary of the data sets in this region. In this paper, we focus on the most recent family of models in the field of multimodal sentiment analysis and compare and analyze these models. The application of multimodal sentiment analysis to various areas of society is discussed. Finally, the challenges and future developments of multimodal sentiment analysis are presented.

2 Multimodal Sentiment Analysis Datasets

With the growth of the Internet, an era of data explosion has been created [27–29]. Numerous researchers have widely collected these data from the Internet (videos, reviews, etc.) and built sentiment datasets according to their own needs. Tab. 1 summarizes the commonly

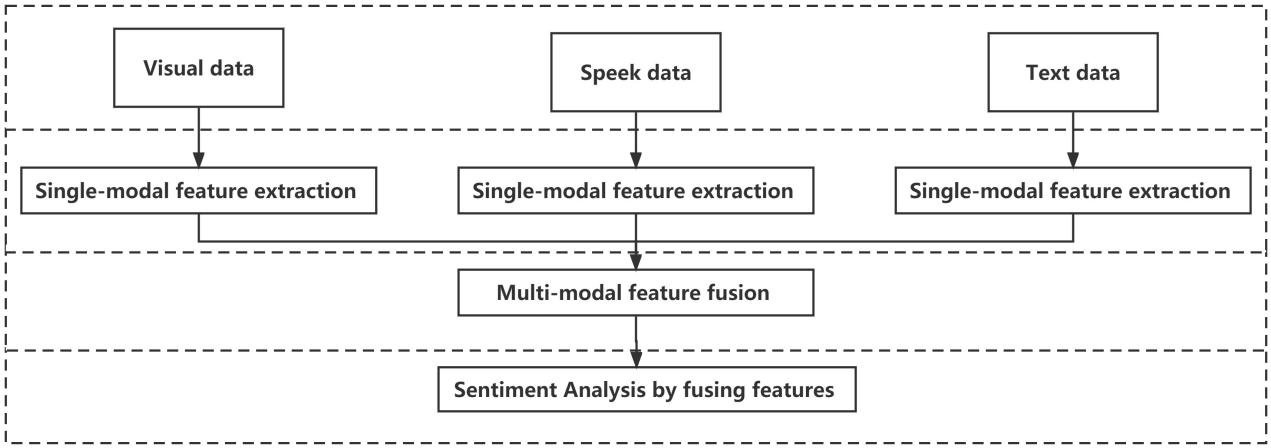

Fig. 2 Figure shows the model architecture for the more classical multi-modal sentiment analysis. The overall architecture consists of three parts: one part for feature extraction of individual modalities, one part for fusion of features of each modality, and one part for sentiment analysis of the fused features. These three parts are very important, and researchers have begun to optimize these three parts one by one

used multimodal datasets. The first column indicates the name of the data set. The second column is the year in which the sentiment data was released. The third column is the category of modalities included in the sentiment dataset. The fourth column is the platform from which the data set came. The fifth column is the language used by the dataset.

IEMOCAP [30]. This is a multi-modal sentiment analysis dataset released by the Speech Analysis and Interpretation Laboratory. In total, there are 1, 039 conversational segments and the total length of the video is 12 hours. Subjects participated in sessions of five scenarios separately and performed emotions according to a preset scenario. The data not only contains audio, video and text information, but also facial expression information and posture information (which is obtained through additional sensors). The data points are divided into 10 categories: neutral, happy, sad, angry, surprised, scared, disgusted, frustrated, excited, other.

DEAP [31]. This is a dataset for sentiment analysis based on physiological signals. The dataset examined EEG data from 32 subjects (1:1 male to female ratio). EEG signals were collected at 512Hz in the frontal, parietal, occipital and temporal lobes of the subject's brain. The subjects annotated the EEG signals by rating the three aspects (Valence, Arousal, Dominance) from 1 to 9 after watching the corresponding videos.

CMU-MOSI [32]. This dataset collects 93 critical YouTube videos that express different topics. The video files in the dataset are strictly required to have only one speaker, which in most cases is facing the camera and can adequately capture facial features. At the same time, there is no restriction on the model of the camera, the distance, or the speaker scene. The 93 videos contain 89 different speakers. There were 41 women and 48 men. All speakers made comments and presentations in English. The 93 videos were split into 2199 subjective opinion segments. For annotation work on the dataset, sentiment intensity annotations ranging from strongly negative to strongly positive (-3 to 3) were produced.

CMU-MOSEI [33]. The dataset consists of 3228 videos from YouTube. It contains data from three modalities: text, visual and sound. These 3228 videos are divided into 23453 segments. In total, 1,000 speakers participated and 250 different topics were covered. All speakers made comments and presentations in English. For label and annotation work on datasets, both sentiment and sentiment annotations are available. The emotion labels contain six categories: happy, sad, angry, scared, disgusted and surprised. It also contains sentiment intensity markers ranging from strongly negative to strongly positive (-3 to 3).

MELD [34]. The dataset is a collection of video clips from the television series Friends. It contains both textual information and audio and video information corresponding to the textual information. There are a total of 1400 videos, which are split into 13,000 individual

Name	Year	Modalities	Source	Language	Number
IEMOCAP	2008	A+V+T	N/A	English	10039
DEAP	2011	A+V+T A+V+T A+V+T	N/A	English	10039
CMU-MOSI	2016	A+V+T	YouTube	English	2199
CMU-MOSEI	2018	A+V+T	YouTube	English	23453
MELD	2019	A+V+T	The Friends	English	13000
Multi-ZOL	2019	V+T	ZOL.com	Chinese	5288
CH-SIMS	2020	A+V+T	N/A	Chinese	2281
CMU-MOSEAS	2021	A+V+T	YouTube	Spanish Portuguese German French	40000
FACTIFY	2022	V+T	Twitter	English	50000
MEMOTION	2022	V+T	Reddit Facebok	English	10000

Table 1 This table contains the used multimodal datasets. The first column indicates the name of the data set. The second column is the year in which the sentiment data was released. The third column is the category of modalities included in the sentiment dataset. The fourth column is the platform from which the data set came. The fifth column is the language used by the dataset.

segments. On the annotations of the dataset, there are seven categories of annotations: Anger, Disgust, Sadness, Joy, Neutral, Surprise and Fear. There are three sentiment annotations for each segment: positive, negative, and neutral).

Multi-ZOL [35]. This is a bimodal sentiment classification dataset for images and text. The researchers collected reviews of mobile phones on ZOL.com. It contains 5288 sets of multimodal data points covering multiple models of mobile phones from multiple brands. These data points are annotated with a sentiment intensity from 1 to 10 for six aspects.

CH-SIMS [36]. The dataset is a collection of 60 open-sourced videos on the web, split into 2281 video clips. A great feature of this dataset is that only Chinese (Mandarin) is considered. It is also ensured that all segments contain only the face and voice of one character. The entire dataset spans a wide range of scenes and different speaker ages. It is important to note that this dataset is not only multimodal labeled, but also individually labeled for each modality. The dataset annotations contain sentiment intensity annotations ranging from negative to positive (-1 to 1). In addition, the dataset also contains annotations for other attributes such as age and gender, which can help researchers to do further research.

CMU-MOSEA [37]. The best feature of this dataset is its extension to multiple languages, including Spanish, Portuguese, German, and French. This is a line of research that has a huge space to explore. The dataset contains 40, 000 sentence fragments, which is quite a large dataset. It covers 250 different topics and 1645 speakers. The annotations are split in two ways: sentiment intensity and binary. Each sentence is annotated with sentiment strength in the interval [-3,3]. The binary includes whether the speaker expressed an opinion or made an objective statement. Emotions are divided into six categories for each sentence: happiness, sadness, fear, disgust, surprise and annotation.

FACTIFY [38]. This dataset is a fake news detection dataset based on implementation validation. It contains data for both image and text modalities. The dataset contains 50,000 sets of data. Most of the data's claims refer to politics and government. The data are annotated into three categories: support, no evidence, and refutation.

MEMOTION [39]. This is a meme-based dataset where each meme typically contains two modalities: image and text. The researchers manually downloaded some popular memes (about politics, religion, and sports). The dataset contains 10, 000 data points in total, and the sub-tasks are Sentiment Analysis, Emotion Classi-

fication, and Scale/Intensity of Emotion Classes. The annotators annotated each data point differently under the different subtasks. Subtask one annotates each data point into three categories (negative, neutral, and positive).. Subtask two annotates each data point into four categories (humour, sarcasm, offense, motivation). Subtask three annotate each data point in the interval [0,4] to indicate the sentiment intensity.

3 Multimodal fusion

Multimodal data describe objects from different perspectives and are more informative than single-modal data. Data information from different modalities can be complementary to each other. In the task of multimodal sentiment analysis, it is a very important and challenging task to fuse the data features between different modalities, preserve the semantic integrity of the modalities, and achieve a good fusion between different modalities. According to the different modes of modal fusion, it can be summarized as feature-based multimodal fusion in the early stage, model-based multimodal fusion in the middle stage, and decision-based multimodal fusion in the late stage.

3.1 Early feature-based approaches for multimodal fusion

Early feature-based multimodal fusion methods perform shallow fusion after feature extraction in the early stage. The fusion of the features of different modalities at the shallow level of the model is equivalent to unifying the features of different single modalities into the same parameter space. Due to the differences in information between different modalities, features commonly contain a large amount of redundant information, and it is frequently necessary to use dimensionality reduction methods to remove the redundant information. Features after dimensionality reduction are fed into the model to complete feature extraction and prediction. Early feature fusion wants the model to consider input information from multiple modalities at the beginning of feature modeling. However, the method of unifying multiple different parameter spaces in the input layer may not achieve the desired effect due to the differences in the parameter spaces of different modes. Some representative models are:

THMM (Tri-modal Hidden Markov Model) [40]. The main idea is to represent the eigenvectors of multiple modalities as higher-order tensors and then use tensor decomposition methods to extract the hidden states and transition probabilities, thus enabling the modeling and analysis of multimodal sequences. The advantage is that it can effectively exploit the correlation and complementarity between multimodal data, while avoiding the curse of dimensionality and overfitting. The disadvantage is that the order and rank of the tensors and the number of hidden states need to be predetermined, and these parameters may affect the performance and efficiency of the model.

RMFN (Recurrent Multistage Fusion Network) [41]. The model uses multiple recurrent neural network layers to gradually fuse features from different modalities, from local to global, from low-level to high-level, and finally obtain a comprehensive sentiment representation.

RAVEN (Recurrent Attended Variation Embedding Network) [42]. The model uses an attention mechanism to adjust the location of features in semantic space for different modalities, such that the same word can exhibit different emotions under different nonverbal behaviors.

HFFN(Hierarchical feature fusion network with local and global perspectives for multimodal affective computing) [43]. In, a Hierarchical Fusion-Network is proposed for multi-modal sentiment analysis, which includes a local fusion module and a global fusion module. Local cross-modal fusion is explored through a sliding window, which effectively reduces the computational complexity.

MCTN (Multimodal Cyclic Translation Network) [44]. This model uses recurrent neural networks and adversarial learning to learn joint representations between different modalities, thereby improving the ability of single-modal representations and dealing with missing modalities or noise.

3.2 Mid-term model-based multimodal fusion method

A medium-term model-based multimodal fusion approach is to feed multimodal data into the network, and the intermediate layers of the model perform feature fusion between the modalities. Model-based modality fusion methods can select the location of modality feature fusion to achieve intermediate interactions. Model-based

fusion typically uses multiple kernel learning, neural networks, graph models, and alternative methods.

MKL (Multiple Kernel Learning) [45]. This model is a multiple kernel learning approach. It uses different kernel functions to represent different modal information and selects the optimal combination of kernel functions by optimizing an objective function to achieve the fusion of multi-modal information.

BERT-like (Self Supervised Models to Improve Multimodal Speech Emotion Recognition) [46]. This model is a Transformer-based multi-modal sentiment analysis method that can leverage self-attention mechanism to achieve alignment and fusion between text and image.

3.3 Multimodal model based on decision fusion in the later stage

A decision level fusion method is used to fuse information from different modalities. Decision-level fusion refers to training models separately on data from different modalities to incorporate outputs from different modalities into the final decision. Multimodal models based on decision fusion typically fuse modalities using methods such as averaging, majority voting, weighting, and learnable models. Such models are typically lightweight and flexible. When any modality is missing, the decision can be made by using the remaining modalities.

Deep Multimodal Fusion Architecture [47]. In this model, each modality has an independent classifier. The prediction results are output after averaging the confidence scores of each classifier.

SAL-CNN (Select-Additive Learning CNN) [48]. This model is a multimodal sentiment analysis model based on CNN and attention mechanism. It uses an adaptive attention mechanism to fuse text and image features, a spatial attention mechanism to extract text-related regions in images, and finally a completely connected layer to classify the output.

TSAM (Temporally Selective Attention Model) [49]. The proposed model is a time-selective attention model, which assigns weights through an attention mechanism to help the model choose the time step, and finally sends it to a distinct SDL loss function model for sentiment analysis. SDL is a multi-modal sentiment analysis method based on Self-Distillation Learning, which can exploit the complementarity between different modalities to improve the generalization ability and robustness of the model.

4 Latest Multimodal Sentiment Analysis Models

In recent years, the field of multimodal sentiment analysis has evolved into a huge system and many practical and efficient models and architectures have emerged. It's hard to cover all the models here. In this chapter, we present some recent and cutting-edge multimodal sentiment analysis models. Most of these models were used as benchmark models by later researchers for experimental reference. These models are summarized in Tab. 2. The first column is the name of the model. The second column is the year in which the model was published. The third column is the dataset used by the model. The fourth column is the accuracy under this dataset.

MultiSentiNet-Att [50]. This model uses an LSTM network to incorporate text information into word vectors. VGG is used to extract both target feature information and scene feature information of an image. The target and scene feature vectors are used to perform cross-modal attention mechanism learning with word vectors. That is, the target feature information and the scene feature information of the image are combined to assign special weights to the word vectors related to the sentiment in the text. The resulting features are fed into a multi-layer perceptron to complete the sentiment analysis task.

DFF-ATMF [51]. The proposed model mainly considers text modality and audio modality. The main contribution is to propose new multi-feature fusion strategies and multi-modal fusion strategies. Two parallel branches are used to learn features for text modality and features for audio modality. For the features of these two modalities, a multimodal attention fusion module is used to complete the multimodal fusion.

AHRM [52]. This model is mainly used to capture the relationship between text modality and visual modality. The authors propose an attention mechanism based heterogeneous relation model, which can well integrate the respective high-quality representation information of

Name	Year	Dataset	Acc
MultiSentiNet-Att	2017	MVSA	68.86%
DFF-TMF	2019	CMU-MOSI CMU-MOSEI	80.98% 77.15%
AHRM	2020	Flickr Getty Image	87.10% 87.80%
SFNN	2020	Yelp	62.90%
MISA	2020	MOSI	83.40%
MAG-BERT	2020	CMU-MOSI CMU-MOSEI	84.10% 84.50%
TIMF	2021	CMU-MOSI CMU-MOSEI	92.28% 79.46%
Auto-ML based Fusion	2021	B-T4SA	95.19%
Self-MM	2022	CMU-MOSI CMU-MOSEI CH-SIMS	84.00% 82.81% 80.74%
DISRFN	2022	CMU-MOSI CMU-MOSEI	83.60% 87.50%

Table 2 This table contains some of the most recent and top-performing multimodal sentiment analysis models. The first column is the name of the model. The second column is the year in which the model was published. The third column is the dataset used by the model. The fourth column is the accuracy under this dataset.

text modality and visual modality. This progressive dual attention mechanism can well highlight the channel-level semantic information of image and text information. To integrate social attributes, social relations are introduced to capture complementary information from the social context, and heterogeneous networks are constructed to integrate features.

SFNN [53]. The proposed model is a neural network based on semantic feature fusion. Convolutional neural networks and attention mechanisms are used to extract visual features. Visual features are mapped to text features and combined with text modality features for sentiment analysis.

MISA [54]. The proposed model presents a novel multimodal sentiment analysis framework. Each modality is mapped into two distinct feature spaces after feature extraction. One feature space mainly learns the invariant features of the modality and the other one learns the unique features of the modality.

MAG-BERT [55]. The authors propose a "multimodal" adaptation architecture and apply it to BERT. The model can receive input from multiple modalities during fine-tuning. MAG can be thought of as a vector embedding structure that allows us to input multimodal information and embed it as a sequence to BERT.

TIMF [56]. The main idea of this model is that each modality learns features separately and performs tensor fusion of the features of each modality. In the dataset fusion stage, the feature fusion for each modality is implemented by a tensor fusion network. In the decision fusion stage, the upstream results are fused by soft fusion to adjust the decision results.

Auto-ML based Fusion [57]. The authors propose to combine text and image individual sentiment analysis into a final fused classification based on AutoML. This approach combines individual classifiers into a final classification using the best model generated by Auto-ML. This is a typical model for decision-level fusion.

Self-MM [58]. In, the authors combine self-supervised learning and multi-task learning to construct a novel multi-modal sentiment analysis architecture. To learn the private information of each modality, the authors construct a single-modal label generation module ULGM based on self-supervised learning. The loss function corresponding to this module is designed to incorporate the private features learned by the three self-supervised learning subtasks into the original multi-modal senti-

ment analysis model using a weight adjustment strategy. The proposed model performs well, and the self-supervised learning based ULGM module also has the ability of single-modal label calibration.

DISRFN [59]. The model is a dynamically invariant representation-specific fusion network. The joint domain separation network is improved to obtain a joint domain separation representation for all modalities, so that redundant information can be effectively utilized. Second, a HGFN network is used to dynamically fuse the feature information of each modality and learn the features of multiple modal interactions. At the same time, a loss function that improves the fusion effect is constructed to help the model learn the representation information of each modality in the subspace.

5 Model comparison and suggestions

We have screened five state-of-the-art multimodal sentiment analysis models: DFF-ATMF, MAG-BERT, TIMF, Self-MM, and DISRFN. This subsection further analyzes these models and compares their performance under the same datasets (CMU-MOSI and CMU-MOSEI). DFF-ATMF does not consider visual modality, while the other models analyze sentiment from three modalities of audio, text, and vision.

DFF-ATMF uses feature vector fusion and multimodal attention fusion to learn more comprehensive sentiment information. However, there is a risk of overfitting due to the use of multi-layer neural networks and sophisticated fusion methods.

MAG-BERT adapts the interior of BERTs using multimodal adaptation gates, which employ a simple yet effective fusion strategy without changing the structure and parameters of BERTs. However, multimodal attention can only be performed within the same timestep but not across timesteps, which may ignore some temporal relationships. MAG-BERT requires freezing the parameters of BERT without being able to fine-tune BERT, which may result in a representation of BERT that is not adapted to a specific task or domain.

TIMF leverages the self-attention mechanism of Transformers to learn complex interactions between multimodal data and generate unified sentiment representations. Its disadvantage is that it may suffer from extreme computational complexity, long training times, and problems with large amounts of labeled data.

Self-MM is a self-supervised multi-modal sentiment analysis model that uses a multi-task learning strategy to learn both multimodal and unimodal emotion recognition tasks. Its advantage is that it can generate single-modal labels using a self-supervised approach, saving the cost and time of manual labeling. Its disadvantage is that there can be interference and imbalance between multiple tasks, requiring an appropriate weight adjustment strategy to balance the learning progress of different tasks.

DISRFN is a deep residual network-based multimodal sentiment analysis model that exploits the strategy of Dynamic Invariant-Specific Representation Fusion Network to improve sentiment recognition capability. Its advantage is that it can efficiently utilize redundant information to obtain joint domain-separated representations of all modalities through a modified joint domain separation network and dynamically fuse each representation through a hierarchical graph fusion network to obtain interaction information of multimodal data. Its disadvantage is that there can be interference and imbalance between multiple tasks, requiring a suitable weight adjustment strategy to balance the learning progress of different tasks.

Tab. 3 shows the performance metrics of these five models under the CMU-MOSI and CMU-MOSEI datasets. Based on the analysis and performance metrics of these five models, we recommend using the BERT model to extract features of text information. For video and audio modality information, LSTM is recommended to extract features as it captures modality information in the time series. DFF-ATMF does not consider visual modality, resulting in relatively low performance metrics. Visual information can provide information about human expressions, poses, scenes, etc., enhancing the information of text and speech modalities. Therefore, visual modality information deserves to be considered and explored in multimodal sentiment analysis. Overall, these findings can serve as a guide for building effective multimodal sentiment analysis models that leverage the strengths of each modality while addressing their limitations.

6 Challenges and Future Scope

With the development of deep learning, multimodal sentiment analysis techniques have also been rapidly developed [60–63]. However, multi-modal sentiment analysis still faces many challenges. This subsection analyzes the current state of research, challenges, and future developments in multimodal sentiment analysis.

6.1 Dataset

In multimodal sentiment analysis, the role of the dataset is very large. A large dataset in multiple languages is currently missing. That said, there are many languages

Model	CMU-MOSI				CMU-MOSEI			
	MAE	**Corr**	**Acc**	**F1-Score**	**MAE**	**Corr**	**Acc**	**F1-Score**
DFF-ATMF	–	–	80.9	81.3	–	–	77.2	78.3
MAG-BERT	0.712	0.796	–	86	0.623	0.677	82	82.1
TIMF	**0.373**	**0.93**	**92.3**	**92.3**	0.645	0.669	79.5	79.5
Self-MM	0.723	0.797	84.8	84.8	**0.534**	0.764	84.1	84.1
DISRFN	0.798	0.734	83.4	83.6	0.591	**0.78**	**87.5**	**87.5**

Table 3 The table shows the performance metrics of the DFF-ATMF, MAG-BERT, TIMF, Self-MM and DISRFN models under the CMU-MOSI and CMU-MOSEI datasets. The evaluation parameters included: MAE, Corr, Acc and F1-Score.

and races in many countries, and a very large dataset. Using this dataset, a multi-modal sentiment analysis model with strong generalization and wide usage can be trained. In addition, current multimodal datasets still have low annotation accuracy and have not yet reached absolute continuous values. This requires researchers to label multimodal datasets more finely. Most current multimodal data contain only visual, speech, and text modalities and lack modal information combined with physiological signals such as brain waves and pulses.

6.2 Detection of Hidden Emotions

There has always been a recognized difficulty in multimodal sentiment analysis tasks: the analysis of hidden emotions. Hidden emotions [64, 65] include: sarcastic emotions (such as sarcastic words), emotions that need to be concretely analyzed in context, and complex emotions [66,67] (such as a person's happiness and sadness). It is important to explore these hidden emotions. It's the gap between human and artificial intelligence [68],

6.3 Multiple forms of video data

Multi-modal sentiment analysis involving video data presents unique challenges. For example, speakers face the camera, videos from similar sources have similar sharpness and camera models, and the data resolution is maintained at a high level with less noise. In the past, high quality video data with similar data distributions made multimodal sentiment analysis tasks simpler. However, the actual situation is more complex and requires models to be robust against noise and applicable to low-resolution video data. Additionally, researchers should explore capturing micro-expressions and micro-gestures of speakers for sentiment analysis.

6.4 Multiform language data

In contrast, text data is typically single-formatted in multimodal sentiment analysis tasks. However, text data in online communities is cross-lingual, with reviewers using multiple languages to make more vivid comments. This remains a challenge in multimodal sentiment analysis tasks. Text data with mixed emotions also presents a challenge, as researchers must determine how to effectively analyze these mixed memes that contain strong emotional messages from reviewers. Transcribing text data directly from speech makes it particularly difficult to analyze emotions when multiple people are talking. Furthermore, cultural characteristics of different regions and countries may cause the same text data to reflect different emotions.

6.5 Future Prospectsn

The future of multimodal sentiment analysis techniques is extremely bright, and some of the future applications are listed below. Multimodal emotion analysis for real-time assessment of mental health [69–71]; multimodal criminal linguistic deception detection model [72]; offensive language detection; A human-like emotion-aware robot, etc. Multimodal emotion analysis is a technique for recognizing and analyzing emotions. Models that combine multi-modal information data for sentiment analysis can effectively improve the accuracy of sentiment analysis. In the future, multi-modal sentiment analysis techniques will be gradually improved. Perhaps one day there will be a multimodal sentiment analysis model with a large number of parameters that will have the same sentiment analysis capabilities as humans. It was a thing of rapture.

7 Conclusion

Here's my edited version of the article:

Multimodal sentiment analysis has become a central research topic in natural language processing and computer vision, with researchers recognizing its importance. In this review, we provide a detailed explanation of various aspects of multimodal sentiment analysis, including

its research background, definition, and development process. We summarize commonly used benchmark datasets in Tab. 1 and compare recent state-of-the-art multimodal sentiment analysis models. Finally, we present the challenges posed by the field of multimodal sentiment analysis and explore possible future developments. Many prospective works are currently being carried out and have even been largely implemented. In view of the existing challenges, we summarize meaningful research directions as follows:

Construct a large multimodal sentiment dataset in multiple languages. Solve the domain transfer problem of video, text, and speech modal data. Build a unified, large-scale multimodal sentiment analysis model with excellent generalization performance. Optimize the algorithm by reducing model parameters and complexity. Address the challenge of multilingual hybridness in multimodal sentiment analysis. Discuss the weight problem of modal fusion and provide the most reasonable weighting scheme for different modalities in different cases. Explore the correlation between modalities by separating shared and private information to improve model performance and interpretability. Develop a multimodal sentiment analysis model that can effectively capture hidden emotions.

Acknowledgements This work was supported in part by Joint found for smart computing of Shandong Natural Science Foundation under Grant ZR2020LZH013; open project of State Key Laboratory of Computer Architecture CARCHA202001; the Major Scientific and Technological Innovation Project in Shandong Province under Grant 2021CXG010506 and 2022CXG010504; "New University 20 items" Funding Project of Jinan under Grant 2021GXRC108 and 2021GXRC024.

References

1. Robert Plutchik. 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, 2001.
2. Julien Deonna and Fabrice Teroni. *The emotions: A philosophical introduction*. Routledge, 2012.
3. Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.
4. Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367–1373, 2004.
5. Erik Cambria, Björn Schuller, Yunqing Xia, and Catherine Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent systems*, 28(2):15–21, 2013.
6. Arshi Parvaiz, Muhammad Anwaar Khalid, Rukhsana Zafar, Huma Ameer, Muhammad Ali, and Muhammad Moazam Fraz. Vision transformers in medical computer vision—a contemplative retrospection. *Engineering Applications of Artificial Intelligence*, 122:106126, 2023.
7. Bo Zhang, Jun Zhu, and Hang Su. Toward the third generation artificial intelligence. *Science China Information Sciences*, 66(2):1–19, 2023.
8. Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.
9. Jireh Yi-Le Chan, Khean Thye Bea, Steven Mun Hong Leow, Seuk Wai Phoong, and Wai Khuen Cheng. State of the art: a review of sentiment analysis based on sequential transfer learning. *Artificial Intelligence Review*, 56(1):749–780, 2023.
10. Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, 2022.
11. Hui Li, Qi Chen, Zhaoman Zhong, Rongrong Gong, and Guokai Han. E-word of mouth sentiment analysis for user behavior studies. *Information Processing & Management*, 59(1):102784, 2022.
12. Ashima Yadav and Dinesh Kumar Vishwakarma. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6):4335–4385, 2020.
13. Ganesh Chandrasekaran, Tu N Nguyen, and Jude Hemanth D. Multimodal sentimental analysis for social media applications: A comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5):e1415, 2021.
14. Bernhard Kratzwald, Suzana Ilic, Mathias Kraus, Stefan Feuerriegel, and Helmut Prendinger. Decision support with text-based emotion recognition: Deep learning for affective computing. *arXiv preprint arXiv:1803.06397*, 2018.
15. Carlo Strapparava and Rada Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 70–74, 2007.
16. Yang Li, Quan Pan, Suhang Wang, Tao Yang, and Erik Cambria. A generative model for category text generation. *Information Sciences*, 450:301–315, 2018.
17. Rong Dai. Facial expression recognition method based on facial physiological features and deep learning. *Journal of Chongqing University of Technology (Natural Science)*, 34(6):146–153, 2020.
18. Zhu Ren, Jia Jia, Quan Guo, Kuo Zhang, and Lianhong Cai. Acoustics, content and geo-information based sentiment prediction from large-scale networked voice data. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–4. IEEE, 2014.
19. LIU Jiming, ZHANG Peixiang, LIU Ying, ZHANG Weidong, and FANG Jie. Summary of multi-modal sentiment analysis technology. *Journal of Frontiers of Computer Science & Technology*, 15(7):1165, 2021.
20. Feiran Huang, Xiaoming Zhang, Zhonghua Zhao, Jie Xu, and Zhoujun Li. Image–text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems*, 167:26–37, 2019.
21. Akshi Kumar and Geetanjali Garg. Sentiment analysis of multimodal twitter data. *Multimedia Tools and Applications*, 78:24103–24119, 2019.
22. Ankita Gandhi, Kinjal Adhvaryu, and Vidhi Khanduja. Multimodal sentiment analysis: review, application domains and future directions. In *2021 IEEE Pune Section International Conference (PuneCon)*, pages 1–5. IEEE, 2021.

23. Vaibhav Rupapara, Furqan Rustam, Hina Fatima Shahzad, Arif Mehmood, Imran Ashraf, and Gyu Sang Choi. Impact of smote on imbalanced text features for toxic comments classification using rvvc model. *IEEE Access*, 9:78621–78634, 2021.
24. Jia Li, Ziyang Zhang, Junjie Lang, Yueqi Jiang, Liuwei An, Peng Zou, Yangyang Xu, Sheng Gao, Jie Lin, Chunxiao Fan, et al. Hybrid multimodal feature extraction, mining and fusion for sentiment analysis. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pages 81–88, 2022.
25. Anna Favaro, Chelsie Motley, Tianyu Cao, Miguel Iglesias, Ankur Butala, Esther S Oh, Robert D Stevens, Jesús Villalba, Najim Dehak, and Laureano Moro-Velázquez. A multi-modal array of interpretable features to evaluate language and speech patterns in different neurological disorders. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 532–539. IEEE, 2023.
26. Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*, 37:98–125, 2017.
27. Sathyan Munirathinam. Industry 4.0: Industrial internet of things (iiot). In *Advances in computers*, volume 117, pages 129–164. Elsevier, 2020.
28. Esteban Ortiz-Ospina and Max Roser. The rise of social media. *Our world in data*, 2023.
29. Abdul Haseeb, Enjun Xia, Shah Saud, Ashfaq Ahmad, and Hamid Khurshid. Does information and communication technologies improve environmental quality in the era of globalization? an empirical analysis. *Environmental Science and Pollution Research*, 26:8594–8608, 2019.
30. Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.
31. Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.
32. Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.
33. AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.
34. Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.
35. Nan Xu, Wenji Mao, and Guandan Chen. Multi-interactive memory network for aspect based multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 371–378, 2019.
36. Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3718–3727, 2020.
37. Amir Zadeh, Yan Sheng Cao, Simon Hessner, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. Cmu-moseas: A multimodal language dataset for spanish, portuguese, german and french. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2020, page 1801. NIH Public Access, 2020.
38. Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. Factify: A multi-modal fact verification dataset. In *Proceedings of the First Workshop on Multimodal Fact-Checking and Hate Speech Detection (DE-FACTIFY)*, 2022.
39. Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, Parth Patwa, Amitava DaS, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. Memotion 2: Dataset on sentiment and emotion analysis of memes. In *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*, 2022.
40. Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176, 2011.
41. Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. Multimodal language analysis with recurrent multistage fusion. *arXiv preprint arXiv:1808.03920*, 2018.
42. Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223, 2019.
43. Sijie Mai, Haifeng Hu, and Songlong Xing. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 481–492, 2019.
44. Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899, 2019.
45. Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2539–2544, 2015.
46. Shamane Siriwardhana, Andrew Reis, Rivindu Weerasekera, and Suranga Nanayakkara. Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition. *arXiv preprint arXiv:2008.06682*, 2020.
47. Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 284–288, 2016.
48. Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. Select-additive learning: Improving generalization in multimodal sentiment analysis.

In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 949–954. IEEE, 2017.
49. Hongliang Yu, Liangke Gui, Michael Madaio, Amy Ogan, Justine Cassell, and Louis-Philippe Morency. Temporally selective attention model for social and affective state recognition in multimedia content. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1743–1751, 2017.
50. Nan Xu and Wenji Mao. Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2399–2402, 2017.
51. Feiyang Chen, Ziqian Luo, Yanyan Xu, and Dengfeng Ke. Complementary fusion of multi-features and multi-modalities in sentiment analysis. *arXiv preprint arXiv:1904.08138*, 2019.
52. Jie Xu, Zhoujun Li, Feiran Huang, Chaozhuo Li, and S Yu Philip. Social image sentiment analysis by exploiting multimodal content and heterogeneous relations. *IEEE Transactions on Industrial Informatics*, 17(4):2974–2982, 2020.
53. Weidong Wu, Yabo Wang, Shuning Xu, and Kaibo Yan. Sfnn: Semantic features fusion neural network for multimodal sentiment analysis. In *2020 5th International Conference on Automation, Control and Robotics Engineering (CACRE)*, pages 661–665. IEEE, 2020.
54. Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131, 2020.
55. Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access, 2020.
56. Jianguo Sun, Hanqi Yin, Ye Tian, Junpeng Wu, Linshan Shen, and Lei Chen. Two-level multimodal fusion for sentiment analysis in public security. *Security and Communication Networks*, 2021:1–10, 2021.
57. Vasco Lopes, António Gaspar, Luís A Alexandre, and João Cordeiro. An automl-based approach to multimodal image sentiment analysis. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2021.
58. Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10790–10797, 2021.
59. Jing He, Haonan Yanga, Changfan Zhang, Hongrun Chen, and Yifu Xua. Dynamic invariant-specific representation fusion network for multimodal sentiment analysis. *Computational Intelligence and Neuroscience*, 2022, 2022.
60. Mahesh G Huddar, Sanjeev S Sannakki, and Vijay S Rajpurohit. A survey of computational approaches and challenges in multimodal sentiment analysis. *Int. J. Comput. Sci. Eng*, 7(1):876–883, 2019.
61. Ramandeep Kaur and Sandeep Kautish. Multimodal sentiment analysis: A survey and comparison. *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines*, pages 1846–1870, 2022.
62. Lukas Stappen, Alice Baird, Lea Schumann, and Schuller Bjorn. The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements. *IEEE Transactions on Affective Computing*, 2021.
63. Anurag Illendula and Amit Sheth. Multimodal emotion classification. In *companion proceedings of the 2019 world wide web conference*, pages 439–449, 2019.
64. Donglei Tang, Zhikai Zhang, Yulan He, Chao Lin, and Deyu Zhou. Hidden topic–emotion transition model for multi-level social emotion detection. *Knowledge-Based Systems*, 164:426–435, 2019.
65. Petr Hajek, Aliaksandr Barushka, and Michal Munk. Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining. *Neural Computing and Applications*, 32:17259–17274, 2020.
66. Soonil Kwon. A cnn-assisted enhanced audio signal processing for speech emotion recognition. *Sensors*, 20(1):183, 2019.
67. Umar Rashid, Muhammad Waseem Iqbal, Muhammad Akmal Skiandar, Muhammad Qasim Raiz, Muhammad Raza Naqvi, and Syed Khuram Shahzad. Emotion detection of contextual text using deep learning. In *2020 4th International symposium on multidisciplinary studies and innovative technologies (ISMSIT)*, pages 1–5. IEEE, 2020.
68. Fereshteh Ghanbari-Adivi and Mohammad Mosleh. Text emotion detection in social networks using a novel ensemble classifier based on parzen tree estimator (tpe). *Neural Computing and Applications*, 31(12):8971–8983, 2019.
69. Zhentao Xu, Verónica Pérez-Rosas, and Rada Mihalcea. Inferring social media users' mental health status from multimodal information. In *Proceedings of the 12th language resources and evaluation conference*, pages 6292–6299, 2020.
70. Rahee Walambe, Pranav Nayak, Ashmit Bhardwaj, and Ketan Kotecha. Employing multimodal machine learning for stress detection. *Journal of Healthcare Engineering*, 2021:1–12, 2021.
71. Nujud Aloshban, Anna Esposito, and Alessandro Vinciarelli. What you say or how you say it? depression detection through joint modeling of linguistic and acoustic aspects of speech. *Cognitive Computation*, 14(5):1585–1598, 2022.
72. Safa Chebbi and Sofia Ben Jebara. Deception detection using multimodal fusion approaches. *Multimedia Tools and Applications*, pages 1–30, 2021.