Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/112/2024.17918

Exploration of the Application of Multimodal Model in Psychological Analysis

Weihan Wang[1,a,*]

[1]*School of International Digital Economy, Minjiang University, Fuzhou, China*
a. 3222901154@stu.mju.edu.cn
**corresponding author*

Abstract: Multimodal sentiment analysis is one of the important research areas in the field of artificial intelligence today. Multimodal sentiment analysis is to extract features from various human modalities such as facial expressions, body movements, and voice information, perform modal fusion, and finally classify and predict emotions. This technology can be used in multiple scenarios such as stock prediction, product analysis, movie box office prediction, etc., especially psychological state analysis, and has important research significance. This paper introduces two important datasets in multimodal sentiment analysis, namely CMU-MOSEI and IEMOCAP. It also introduces the feature-level fusion, model-level fusion, decision-level fusion and other fusion methods in multimodal fusion methods, and also introduces the semantic feature fusion neural network and sentiment word perception fusion network in multimodal sentiment analysis related models. Finally, the application of multimodal sentiment analysis models in depression and other related mental illnesses and the challenges of multimodal sentiment analysis models in the future are introduced. This paper hopes that the above research will be helpful for multimodal sentiment analysis.

Keywords: Multimodal Model, Modal Fusion, Psychological Analysis.

1. Introduction

With the rapid development of society, in the post-epidemic era, the pressure in life or work of both teenagers and adults continues to increase, which to some extent leads to increased psychological pressure. Excessive psychological pressure may cause inconvenience in study or life, and trigger a series of psychological problems. Having good mental health and well-being can stimulate our potential and enable us to face more challenges[1]. According to the World Health Organization (WHO), nearly 1 billion people worldwide suffer from mental disorders, and reports indicate that mental disorders are the leading cause of disability, which brings great harm to each individual and their family. Therefore, the analysis of psychological state is becoming increasingly important.

More and more researchers are using artificial intelligence to analyze psychological states, including sentiment analysis technology, which is an emerging research direction. The goal is to use various sentiment data sets to train a multimodal sentiment analysis model, analyze, infer and classify human emotions, and ultimately be able to analyze and identify emotions like humans [2]. This technology is not only used in psychological state analysis, but also has wide applications in business, politics, entertainment and other fields.

DOI: 10.54254/2755-2721/112/2024.17918

The research object of sentiment analysis is mainly the characteristics that change during human-computer interaction and human-to-human conversations, such as text information, voice strength, facial expressions, and other behaviors [3]. In the past, sentiment analysis was mainly focused on a single modality, such as text, speech, facial expressions, etc. However, unimodal sentiment analysis has limitations. For example, when it comes to text analysis, the word "great" has an encouraging and positive meaning in most scenarios, but in some ironic scenarios, a single text mode may result in incorrect analysis. At this time, the multimodal sentiment analysis model is proposed to solve this problem. If the voice and text modalities are combined, this problem may be avoided. Therefore, multimodal models will improve the accuracy of unimodal models for sentiment analysis.

Many researchers have conducted in-depth research on multimodal fusion models. Feiran Huang et al. proposed a deep multimodal fusion mechanism model (DMAF) for joint analysis of images and texts. The role of this model is to automatically discover the most emotionally discriminative features in image and text data and use the attention mechanism to focus on emotionally relevant image regions and text words. Through intermediate fusion and late fusion methods, data information from different modalities is integrated for sentiment classification[4]. Qiuchi Li et al. proposed a quantum theory-inspired multimodal fusion framework, mainly used for video sentiment analysis. The model combines information from different modalities through the concepts of superposition and entanglement in quantum mechanics to better predict the emotional state in the video[5].

This paper will be carried out based on the fusion method of multimodal models, relevant data sets and sentiment analysis related models. The specific application of multimodal sentiment analysis in psychological state analysis is also discussed. Finally, the challenges of multimodal models in sentiment analysis and future application directions are summarized.

2. Multimodal fusion method

Table 1: Comparison of multimodal fusion methods.

Fusion Method	Specific methods	Advantage	Disadvantage
Feature-level fusion	Tensor Fusion	Ability to learn intra-modal and inter-modal dynamics	Greatly increases the dimension of the feature vector, making the model too large and difficult to train
	Compact Bilinear Pooling Fusion	Allows multiplicative interactions between all elements, and can capture complex relationships between different modes	When determining the output dimension, it is necessary to find the optimal value through experiments, which increases the difficulty of model training and debugging
	Low Rank Fusion	Avoid creating high-dimensional tensors by decomposing the weight tensor into modality-specific low-rank factors	Once the feature is too long, it is still easy to have parameter explosion

DOI: 10.54254/2755-2721/112/2024.17918

Table 1: (continued).

Model-level fusion	Multi-core learning fusion	Use multiple kernel functions to adapt to the data characteristics of different modes and improve the expressiveness of features and the accuracy of classification.	If the number of samples is too large, the dimension of the kernel matrix will also be very large.
	ML-LSTM Fusion	Ability to effectively consider the relationship between discourses, leading to more robust feature extraction	A multi-layer LSTM structure is used, the model has more parameters and the computational complexity is relatively high.
Decision-level fusion	Weighted majority voting fusion	According to the characteristics of the specific problem and data set, select the appropriate classifier combination and adjust the weights to optimize performance	The choice of weights usually needs to be determined based on experience or experiments, and may be somewhat subjective.
	Selective additive fusion	It can effectively deal with the model generalization problem caused by limited data in multimodal sentiment analysis and reduce the confounding effect of the speaker's unique characteristics in the training data on sentiment classification.	Some parameters in the method (such as the sparse regularizer weights in the selection phase, the parameters of the Gaussian noise in the addition phase, etc.) have a significant impact on the model performance
Other Fusion	Attention Mechanism Fusion	Able to effectively integrate information from different modalities, capturing the unique features of each modality and their interrelationships	If the data of a certain modality has problems such as noise, missing values, or inaccurate annotations, it may affect the attention mechanism's correct focus and integration of the modality information.

Single-modal sentiment analysis often has major limitations, such as low recognition rate, low accuracy, and the influence of intra-modal noise. Therefore, more and more researchers adopt the method of multimodal fusion, extracting the features of each modality and fusing them into a multimodal feature. Common fusion methods nowadays include feature-level fusion, model-level fusion, and decision-level fusion. The paper also lists attention mechanism fusion in addition to the

DOI: 10.54254/2755-2721/112/2024.17918

above three fusion methods. Table 1 lists the specific methods of feature-level, model-level, decision-level, and attention-mechanism fusion as well as the advantages and disadvantages of each method.

2.1. Feature-level fusion

Feature-level fusion, also known as early fusion, refers to fusing the features of each modality into a feature vector through concatenation or other methods after the features are extracted. Before fusion begins, the vectors of each feature must be converted into a unified format. However, due to this early fusion, the modalities may suppress each other, so Amir Zadeh et al. proposed tensor fusion network (TFN). This fusion method mainly constructs a multimodal tensor. TFN can clearly model the interaction between single modality, dual modality and trimodality, and capture the relationship between modalities. Since tensor fusion involves the Cartesian product of multimodal features, the computational complexity is relatively high [6]. Subsequently, Zhun Liu et al. proposed a low-rank fusion method to avoid creating high-dimensional tensors by decomposing the weight tensor into modality-specific low-rank factors [7]. Akira Fukui et al. proposed a compact dual-line pooling fusion method. This method achieves modal fusion by calculating the outer product of two vectors and learning a linear model. Then, the count sketch projection function is used to solve the high-dimensional problem. Finally, the count sketch of the outer product of the two vectors is represented as the convolution of two count sketches, avoiding the direct calculation of the outer product and reducing the amount of calculation and the number of parameters [8].

2.2. Model-level fusion

Model-level fusion, also known as mid-term fusion, is to fuse the collected single-modal feature vectors through a fusion model in the middle layer of the model. Common model-level fusion methods include multi-core learning and neural networks. Soujanya Poria et al. used the extracted feature vectors and the corresponding sentiment polarity labels in the training set to train the classifier of the multi-kernel fusion algorithm to process heterogeneous data, and selected 5 RBF kernels and 3 polynomial kernels as the parameters of the classifier [9]. Weizhi Nie et al. combined a multi-layer network with the traditional Long Short-Term Memory(LSTM) model. They input text features into the first layer of LSTM to obtain the hidden layer state of neurons, and then input it and the audio features into the next layer of LSTM, and so on, finally obtaining the fusion result [10].

2.3. Decision-level fusion

Decision-level fusion, also known as late fusion. It is a method in which the data of each single modality is trained first, and then modal fusion is performed through different decision-making methods in the later stage. Common methods include weighted average, majority voting, etc. The advantage of this fusion method is its flexibility. For example, the missing data of a single modality will not have a significant impact on the fusion result. Azife Dimililer et al. used the weighted majority voting method to fuse sentiment features. The probability values of specific categories provided by six separate classifiers (NB, SGD, SVM, LR, DT, RF) were weighted and summed according to the importance of the classifiers. Finally, the weighted fusion was performed by voting. The category with the largest weighted sum was the final prediction result [11]. Haohan Wang et al. proposed selective additive method to improve the generalization ability of neural network in sentiment analysis. The selective additive method consists of two stages: the first is the selection stage, which discovers the identity-related confounding dimensions by optimizing the loss function. The second is the addition stage, which masks the confounding dimensions by adding gaussian noise [12].

DOI: 10.54254/2755-2721/112/2024.17918

2.4. Other fusion

The fusion of the attention mechanism is outside the above three categories. This fusion method uses a neural network to extract multi-level context features and then perform feature fusion. Fuyan Ma et al. proposed an attention selective fusion method, which can efficiently fuse local and global information by calculating local and global weights and has high flexibility. This fusion method has been experimented on different datasets, and the results show that it has high sentiment analysis accuracy and strong generalization ability[13].

3. Analysis of sentiment analysis related model examples

In the past few years, more and more researchers have trained many models for multimodal sentiment analysis and applied them in various fields. In this paper, two multimodal sentiment analysis models are listed, namely the semantic feature fusion neural network and the sentiment word perception fusion network.

3.1. Semantic feature fusion neural network

In each modality, semantic information is relatively stable. In the process of modal fusion, adding semantic information can improve the performance of the model. Weidong Wu et al. proposed a semantic feature fusion neural network (SFNN) model for sentiment analysis. The model is divided into four parts: visual feature extraction module, image semantic feature extraction module, text feature extraction module and multimodal feature fusion module. The advantage of the model is that it combines image text vectors and text semantic vectors in terms of semantics, which can better perform sentiment analysis. On the Yelp dataset, SFNN achieves an accuracy of 62.8%[14].

3.2. Sentiment word aware fusion network

Sentiment words contained in language modality are of great significance for sentiment analysis. Minping Chen et al. proposed a sentiment word aware network model for sentiment analysis. The model is divided into two parts: shallow fusion part and aggregation part. For the shallow fusion part, a collaborative attention mechanism is used, which can make bidirectional contextual connections between two modalities (such as sound and text) and obtain contextual information by calculating the weights between modalities to achieve shallow fusion of modalities. For the aggregation part, sentiment word-level classification task prediction is used to assist in achieving multimodal fusion. Many benchmark models may overfit on the training set, but this model demonstrates better generalization ability [15].

4. Multimodal sentiment analysis dataset

Table 2: Comparison between CMU-MOSEI dataset and IEMOCAP dataset.

	CMU-MOSEI	IEMOCAP
Data Source	Collected from the online video sharing website YouTube	The Speech Analysis and Interpretation Laboratory at the University of Southern California
Data content	250 topics 1000 speakers 23,453 annotated video clips	Recordings of 10 actors in dyadic conversations About 12 hours of video

DOI: 10.54254/2755-2721/112/2024.17918

Table 2: (continued).

Feature	Large data size Rich multimodal information	Each video is segmented into sentences with fine-grained sentiment annotations Covering a variety of emotions
Advantage	Has emotion strength marker (-3, 3) It is the largest multimodal sentiment and emotion recognition dataset	Facial expression and posture information obtained by sensors High reliability of annotation
limitation	Data labeling is subjective It may be difficult to label and identify complex emotions	Actors may exaggerate their emotions The data deviates from the actual situation

Establishing a complete database mainly includes four stages: data collection, data preprocessing, data post-processing, and data annotation. With the rapid development of the Internet, more and more researchers are collecting various data from the Internet, including videos, texts, actions and other information, and building their own multimodal sentiment analysis databases. The paper lists two popular databases with large data volumes, CMU-MOSEI and IEMOCAP, for a general analysis. Table 2 compares the two datasets based on five dimensions: data source, data content, feature, advantage, and limitation.

4.1. IEMOCAP dataset

IEMOCAP database was collected by the Speech Analysis and Interpretation Laboratory at the University of Southern California and released in 2008. The experimenters selected 10 actors and used the VICON motion capture system to place markers on the actors' faces, heads, and hands to record detailed facial and hand information. The total length of the video is 12 hours, inducing 10 emotions such as happiness, anger, sadness, etc. through improvisation based on hypothetical scenarios. It provides rich resources for the research of multimodal sentiment analysis[16].

4.2. CMU-MOSEI dataset

CMU-MOSEI is a large-scale dataset in multimodal sentiment analysis. It collected 3228 video data from YouTube, including 23453 annotated video clips from 1000 speakers, covering 250 different topics. Six emotions (happiness, sadness, anger, fear, disgust, and surprise) are labeled from extremely negative to extremely positive, providing a valuable dataset for multimodal sentiment analysis [17].

5. Multimodal model application

Multimodal model sentiment analysis has a wide range of applications and has achieved relatively good results. Especially in the field of psychology, for example, many researchers use multimodal sentiment analysis models for lie detection, movie box office prediction, stock prediction, etc., which has social value. In this paper, the application of multimodal sentiment analysis models in mental illness is described.

DOI: 10.54254/2755-2721/112/2024.17918

5.1. Depression detection

According to the WHO, an estimated 4.4% of the global population suffers from depression. Commonly used methods for detecting depression include the Hamilton Depression Rating Scale and doctor interviews with patients. If the multimodal sentiment analysis model is applied to the detection and analysis of patients with depression, it will improve the efficiency of treatment. Li Zhou et al. proposed a time-aware attention multimodal fusion network (TAMFN) for depression detection. TAMFN is divided into three parts: temporal convolutional network, inter-modal feature extraction, and time-aware multimodal fusion. Experiments on the D-Vlog dataset show that it has a good effect on the detection of depression. Compared with the baseline model, this model achieved the best performance in both accuracy and F1 score. Although many baseline models may overfit on the training set, this model demonstrates better generalization ability[18].

5.2. Alzheimer's disease diagnosis

Alzheimer's disease is a common brain disease that affects tens of millions of people worldwide. However, the disease is difficult to detect in the early stages and is a problem of widespread concern among many researchers. Michal Golovanevsky et al. proposed a multimodal deep learning framework that uses imaging, genetic, and clinical data to improve the diagnostic accuracy of Alzheimer's disease (AD) and mild cognitive impairment (MCI). Compared with many studies that only focus on the binary classification problem of AD or MCI, this model explores the three-classification task, especially distinguishing between two highly similar categories of MCI and AD, showing the potential of the model in complex diagnostic tasks[19].

6. Future challenges of multimodal sentiment models

6.1. Dataset

Currently, most datasets are built around the three modalities of images, text, and audio, while there are fewer datasets for EEG signals and physiological signals. If the data set in this area is increased, the accuracy of the multimodal sentiment analysis model will be improved.

6.2. Ethical issues

Mental state and emotions are personal issues. Using artificial intelligence to explore and mine personal information may infringe on personal privacy, which also raises moral and ethical issues. In this regard, it is worthwhile for researchers to think deeply about it in the future.

6.3. Invisible emotions

"I am very happy" may express a happy emotion, but in some specific contexts it may express a negative meaning, which is the so-called invisible emotion. At present, multimodal sentiment analysis models may not be able to detect hidden emotions such as irony and sarcasm, or the analysis effect may not be good. This is one of the challenges that multimodal sentiment analysis models will face in the future.

7. Conclusion

The paper summarizes the fusion methods, datasets, and related models of multimodal sentiment analysis, and introduces its application in psychological state analysis, especially depression detection and Alzheimer's disease detection. Finally, the future challenges of multimodal sentiment analysis

DOI: 10.54254/2755-2721/112/2024.17918

are summarized. Multimodal sentiment analysis is an emerging field that is of great help to various fields such as business, education, and medical care. Multimodal sentiment analysis will develop in a more accurate and personalized direction, and will be able to capture and understand complex emotional states more comprehensively. With the continuous advancement of artificial intelligence technology and sensor technology, the diversity and accuracy of data acquisition will be further improved, which will enable this technology to play a more extensive role in mental health monitoring, emotional interaction systems, and even human-computer interaction. At the same time, cross-cultural sentiment analysis will also become an important research direction, providing a new perspective for emotional understanding in the context of globalization.

References

[1] *World Health Organization. (2021). Comprehensive mental health action plan 2013–2030. World Health Organization.*

[2] *Zhu, T., Li, L., Yang, J., Zhao, S., Liu, H., & Qian, J. (2022). Multimodal sentiment analysis with image-text interaction network. IEEE transactions on multimedia, 25, 3375-3385.*

[3] *Han, W., Chen, H., Gelbukh, A., Zadeh, A., Morency, L. P., & Poria, S. (2021, October). Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In Proceedings of the 2021 international conference on multimodal interaction (pp. 6-15).*

[4] *Huang, F., Zhang, X., Zhao, Z., Xu, J., & Li, Z. (2019). Image–text sentiment analysis via deep multimodal attentive fusion. Knowledge-Based Systems, 167, 26-37.*

[5] *Li, Q., Gkoumas, D., Lioma, C., & Melucci, M. (2021). Quantum-inspired multimodal fusion for video sentiment analysis. Information Fusion, 65, 58-71.*

[6] *Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250.*

[7] *Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., & Morency, L. P. (2018). Efficient low-rank multimodal fusion with modality-specific factors. arXiv preprint arXiv:1806.00064.*

[8] *Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847.*

[9] *Poria, S., Chaturvedi, I., Cambria, E., & Hussain, A. (2016, December). Convolutional MKL based multimodal emotion recognition and sentiment analysis. In 2016 IEEE 16th international conference on data mining (ICDM) (pp. 439-448). IEEE.*

[10] *Nie, W., Yan, Y., Song, D., & Wang, K. (2021). Multi-modal feature fusion based on multi-layers LSTM for video emotion recognition. Multimedia Tools and Applications, 80, 16205-16214.*

[11] *Aziz, R. H. H., & Dimililer, N. (2020, December). Twitter sentiment analysis using an ensemble weighted majority vote classifier. In 2020 International Conference on Advanced Science and Engineering (ICOASE) (pp. 103-109). IEEE.*

[12] *Wang, H., Meghawat, A., Morency, L. P., & Xing, E. P. (2017, July). Select-additive learning: Improving generalization in multimodal sentiment analysis. In 2017 IEEE International Conference on Multimedia and Expo (ICME) (pp. 949-954). IEEE.*

[13] *Ma, F., Sun, B., & Li, S. (2021). Facial expression recognition with visual transformers and attentional selective fusion. IEEE Transactions on Affective Computing, 14(2), 1236-1248.*

[14] *Wu, W., Wang, Y., Xu, S., & Yan, K. (2020, September). SFNN: semantic features fusion neural network for multimodal sentiment analysis. In 2020 5th International*

[15] *Chen, M., & Li, X. (2020, December). Swafn: Sentimental words aware fusion network for multimodal sentiment analysis. In Proceedings of the 28th international conference on computational linguistics (pp. 1067-1077).*

[16] *Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42, 335-359.*

[17] *Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2018, July). 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) (pp. 2236-2246).*

[18] *Zhou, L., Liu, Z., Shangguan, Z., Yuan, X., Li, Y., & Hu, B. (2022). TAMFN: time-aware attention multimodal fusion network for depression detection. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31, 669-679.*

[19] *Golovanevsky, M., Eickhoff, C., & Singh, R. (2022). Multimodal attention-based deep learning for Alzheimer's disease diagnosis. Journal of the American Medical Informatics Association, 29(12), 2014-2022.*