Proceedings of the 3rd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/2025.22725

Research and Application of Key Techniques of Multimodal Sentiment Analysis

Haozhi Luo

Institute of Science, University of New South Wales, Sydney, Australian
z5602722@ad.unsw.edu.au

Abstract: Multimodal sentiment analysis is an important research direction in the field of natural language processing and artificial intelligence, aiming to improve the accuracy of emotion recognition by using multimodal data such as text, speech and image. Under the background of today's big data era, the technology of multi-modal sentiment analysis is becoming more and more important, and it is also widely used in all aspects of people's lives. This paper first introduces the basic concept and research background of multimodal sentiment analysis. Then, a series of key technologies such as feature fusion, feature extraction, sentiment analysis, data fusion and deep learning are discussed in the current situation. Next, it analyzes the typical applications of multimodal sentiment analysis in the fields of social media and human-computer interaction. At the same time, it also analyzes and evaluates the future development direction of multimodal sentiment analysis since the limitations of the current application of the technology, and provides certain ideas for the future development of multimodal sentiment analysis.

Keywords: Multimodal sentiment analysis, Feature fusion, Feature extraction, Data fusion

1. Introduction

In the context of today's big data era, with the continuous development of artificial intelligence, big data science and language analysis, sentiment analysis has become more and more important, and this technology has been applied more and more widely, which can be applied in social media, human-computer interaction, mental health monitoring and other fields, and is closely related to People's Daily life [1-3]. The traditional sentiment analysis is based on a single model, which has certain limitations. For example, many single emotion analysis is only based on the user's text at that time or the change of the user's mood at that time, while ignoring the change of the user's tone and voice intonation and the change of facial subtle expression, the conclusion may be inaccurate and true. Since this method is difficult to capture the real emotions of the user at that time, multi-modal sentiment analysis is applied in this background. Multimodal sentiment analysis is a comprehensive analysis and research from many perspectives and aspects, including image, text, video, voice and other modes [4]. Therefore, the accuracy and robustness of sentiment analysis can be improved.

In recent years, with the development of large-scale data processing technology and deep learning technology, multi-modal sentiment analysis has also been a great development, and has made a breakthrough in technology. In order to improve the accuracy of emotion recognition, a large number of researches have adopted the method of text-speech-visual three-modal combination [5-6]. At the same time, many emerging technologies such as GNN and BERT are also applied to multimodal

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2025.22725

sentiment analysis. Gandhi et al. improved the accuracy of sentiment analysis results [4]. However, although multimodal sentiment analysis technology has been widely used in many fields, it still faces many difficulties and challenges. For example, the instability of mode fusion method, the inaccuracy of large-scale data processing, and the excessive resource consumption are a series of problems. The research objectives of this paper include a review of the current mainstream multi-modal sentiment analysis methods, focusing on feature extraction, modal alignment, feature fusion and emotion reasoning techniques. Figure 1 shows this Paper structure. The practice of multimodal sentiment analysis in different application scenarios is summarized and discussed. For example, this technology shows different practice situations in two different application scenarios: social media and mental health monitoring. Finally, the present challenges and future prospects of multimodal sentiment analysis are studied.

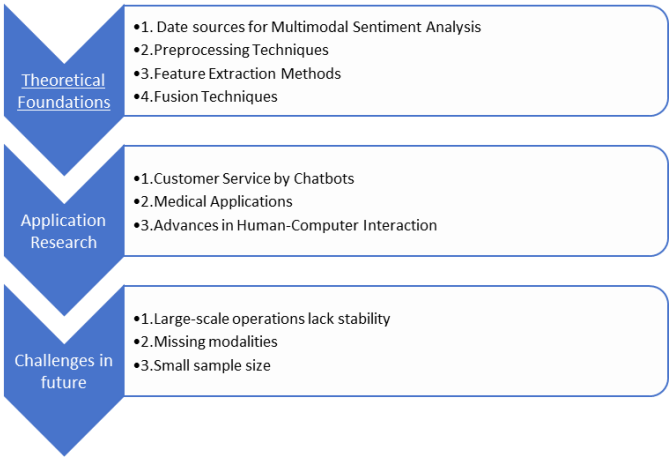


Figure 1: This paper structure flow chart

2. Theoretical foundations

2.1. Data sources for multimodal sentiment analysis

Multimodal sentiment analysis involves multiple data sources, including speech data, visual data, text data, physiological signal data and a series of other data sources. Among them, the three most important data sources are voice data, visual data and text data. Among them, voice data usually comes from recorded interviews, telephone calls, and some blog vlogs and other forms of voice program data [7]. Researchers typically use standard speech emotion datasets such as IEMOCAP and then extract the speaker's audio. Tone of voice. Speed. Forte and other features to identify the speaker's speech to express the emotional features. Secondly, in addition to voice data, visual data is also an extremely important one among multiple data sources for multimodal sentiment analysis. Visual data mainly includes facial expressions and body language. The data sources of visual data mainly include videos on social media, clips of film and television works, and some news interviews broadcast on TV. The researchers extracted visual features from these videos to analyze the emotional characteristics of the characters in the moment.

The last is text data, which is also an important part of multi-modal sentiment analysis data source. Most of them come from articles published on some social media platforms or user comments, such as Twitter, Weibo and other platforms. Researchers perform text feature extraction in order to understand the emotions expressed in the text.

DOI: 10.54254/2753-8818/2025.22725

2.2. Preprocessing techniques

Since multi-modal sentiment analysis involves a variety of different data sources, including a series of related data such as voice data, visual data and text data, and because the data features of different modes are different, a step of pre-processing, key features extraction and modal alignment should be carried out before sentiment analysis. This process can improve the accuracy and robustness of sentiment analysis.

2.2.1. Pre-processing of text data

Text data often has many spelling errors and improper punctuation, so a standardized pre-processing is required [8-9]. First, the original text is cleaned to remove unnecessary labels, symbols, and special characters to ensure that the text data is neat. Secondly, natural language processing tools are used to extract stems to reduce redundant information. Finally, the processed text data needs to be converted into vectors for representation.

2.2.2. Pre-processing of voice data

Speech data is greatly affected by ambient noise and audio quality, so signal processing is needed to improve the accuracy of speech sentiment analysis. First, the background noise needs to be removed by noise reduction. Second, the audio segmentation, that is, the long audio into a fixed length of short fragments. Finally, the most critical step is feature extraction. Researchers often use parameters such as the Meir cepstrum coefficient, pitch, pitch, and speech speed to capture emotional features in speech data. In addition, standardized preprocessing can reduce differences between individual data.

2.2.3. Pre-processing of visual data

Video data mainly involves the actions and facial expressions of the characters in the video, and the pre-processing usually includes face detection and feature extraction. Face detection is an important part of visual data processing. Researchers usually use MTCNN for face detection and Angle adjustment to ensure the stability of facial expressions. As for feature extraction, deep learning technology can be used to extract human facial expressions, and tools such as Open Pose can be used to extract human gestures and analyze human gestures. And in order to reduce computational costs, researchers usually down sample the number of video frames to improve the efficiency of data analysis.

2.2.4. Multi-modal data alignment

Because text data, speech data, and visual data are at different times, they need to be aligned to ensure that data of different modes can be entered into the model synchronously [10]. Common alignment methods include time alignment, where researchers align text, audio, and video to the same time scale through dynamic time warping or interpolation methods [5].

2.3. Feature extraction methods

In multimodal emotion analysis, feature extraction is the key step, which determines the model's ability to understand and express emotion information. And for different data sources, there will be different feature extraction methods. The core point of text feature extraction is to convert text into a computable numerical representation that can be processed by deep learning models. Firstly, the method of word vector representation is widely used. Methods such as Fast Text can map words to

DOI: 10.54254/2753-8818/2025.22725

high-dimensional vector Spaces, making similar words closer to each other, so as to improve the expression ability of text features.

In addition to vector representation, text feature extraction can also use emotion dictionaries, such as Senti Word Net, to score emotion words in text data, so as to obtain a more intuitive emotion feature. Next is the speech data, for the speech data, the common extraction methods are short-time Fourier transform and Maier frequency cepstrum coefficient. The short-time Fourier transform (STFT) can transform the audio signal into a time-frequency graph to analyze the change of speech over time [7]. As for the Meir frequency cepstrum coefficient, it is one of the most commonly used methods in speech emotion analysis to extract the spectral features of audio by imitating the way human ears perceive sound. Speech speed and pause are also important phonetic and emotional characteristics, and different speech speeds express different inner feelings. When people are in a state of excitement, they speak faster.

On the other hand, when people are in a sad state, they speak more slowly. In recent years, due to the development of deep learning technology, technology such as CNN can directly extract features from raw speech data, improving the effect of emotion recognition. Finally, visual feature extraction. Visual feature extraction is basically divided into the analysis of character emotion through the extraction of facial expression and the analysis of character emotion through the extraction of character action feature. In terms of analyzing facial expressions, researchers usually use OpenCV or MTCNN for face detection, and identify characters' emotions through the extraction of facial key points, such as changes in eyebrows and eyes [4]. As for the extraction of character action features, optical flow technology will be used. Optical flow technology can be used to detect motion information in videos and capture the dynamic changes in the body movements of characters.

2.4. Fusion technology

Table 1: The characteristics of different fusion methods

Fusion method	Definition	Advantages	Disadvantages
Early Fusion	Fusion is performed at the data level to transform raw data from different modes into a unified representation and serve as model input.	The information of each mode can be fully utilized, and the information loss at the feature level is less.	The data needs to be aligned to deal with the problem of data size mismatch between different modes.
Late Fusion	At the decision level, the independent models of each mode are trained respectively, and the strategies such as voting and weighted average are used to make decisions in the prediction stage.	Each modal model is trained independently, with high flexibility, which is suitable for the case of missing modal data.	Failure to make full use of the information interaction between modes may lead to information loss.
Hybrid Fusion	Combining early fusion and late fusion, fusion is carried out at the feature level and optimization is carried out at the decision level.	It can not only use the information interaction between modes, but also improve the robustness of the model.	The computational complexity is high and the fusion strategy design is difficult.
Model Fusion	Using deep learning technology to carry out information interaction and fusion of different modes.	It can automatically learn the relationship between modes and improve the accuracy of sentiment analysis.	Training is complex and requires a lot of data and computational resources.

DOI: 10.54254/2753-8818/2025.22725

Table 1 clearly shows and compares the characteristics of different fusion methods used in multimodal sentiment analysis. Therefore, the most suitable multimodal fusion strategy should be selected from the actual situation for different fusion.

3. Application research

3.1. Enhancements in customer service

In the business field, multimodal sentiment analysis is widely used in customer service, emotional marketing and user feedback analysis. Traditional forms of customer service mainly rely on voice system, it is difficult to accurately judge the real emotional changes of customers. The multi-modal sentiment analysis makes up for this disadvantage, because the intelligent customer service based on the multi-modal sentiment analysis technology can be combined with various modes such as voice, text and video to evaluate the customer's emotional changes and product satisfaction from many angles, making the assessment more accurate and thus providing more personalized services. In the telephone customer service of many industries such as banks and network mobile companies, the voice emotion analysis system can detect and analyze the user's emotions at that time in real time. When the system detects the user's anxious and irritable emotions, after connecting the manual customer service, the system will remind the customer service staff to adopt a more gentle communication method to improve the customer's experience.

3.2. Medical applications

In the medical field, multimodal sentiment analysis is mostly used in aspects such as health diagnosis and screening for depression. In traditional mental health diagnosis, a psychiatrist gives a verbal consultation with a patient and then comes up with a treatment plan based on the patient's self-report. In this case, the diagnosis results are often subjective, and sometimes can not accurately describe the patient's condition. The system based on multimodal emotion analysis can evaluate the symptoms of patients according to their movement characteristics, facial expressions, voice intonation, language expression and other aspects, and provide a more accurate and objective state assessment. In addition, with the continuous development of multi-modal emotion analysis technology, telemedicine has gradually become popular. With the help of multi-modal emotion analysis technology, telemedicine system can evaluate the current status of patients in real time through cameras and voice systems to help doctors better provide assistance to patients.

3.3. Advances in human-computer interaction

In the field of intelligent human-computer interaction, due to the application of multimodal sentiment analysis, the field of intelligent human-computer interaction has been greatly improved. Traditional AI voice assistants can only perform some basic voice questions and answers based on user questions and answers. The AI voice assistant based on the technology of multimodal emotion analysis can infer the user's psychological state according to the voice tone, facial expression and some body movements when the customer asks it, and make corresponding responses according to this psychological state, so as to improve the user's experience.

4. Challenges in multimodal sentiment analysis

Although multimodal sentiment analysis has made some breakthroughs in many technical fields, and has been widely used in many aspects of various fields of society, it still faces many challenges. Multimodal sentiment analysis relies on a large number of high-quality multimodal data such as text, voice, video, etc., but in real life, the acquisition of these data is often very difficult, and consumes a

DOI: 10.54254/2753-8818/2025.22725

lot of manpower, material and financial resources. And after obtaining the data, the data alignment is also a big difficulty. Because data of different modes often have different time granularity and information representation, it is a great challenge to align these data so that models can merge them efficiently. Secondly, different modes of emotional information may contain different levels of emotional expression, for example, text data can clearly express the mood of loss with text, but voice data and video data may judge that the user is in a happy state due to some special circumstances of the user at that time, such as a rise in tone, which will lead to information imbalance.

The next is the difficulty of feature extraction and feature fusion, and the feature extraction of different data will cost a lot of financial resources. For example, when extracting the feature of voice data, it is necessary to remove noise processing to improve the performance of the model. Finally, because multimodal sentiment analysis involves personal data such as voice and video, privacy security and ethical issues have attracted much attention. The collection of users' video, voice, facial expression and other data may involve the privacy leakage of users, causing certain troubles to users' daily life, which is common in some sensitive fields such as medical care and finance. Therefore, how to protect the privacy of users while collecting user data is extremely important for the future development of multimodal sentiment analysis technology.

5. Conclusion

This paper explored the key techniques and practical applications of multimodal sentiment analysis, a field that is rapidly evolving alongside the advancements in big data, artificial intelligence, and deep learning. This study has provided a comprehensive review of current methods, focusing on critical areas such as feature extraction, modal alignment, and feature fusion, while also highlighting the integration of state-of-the-art deep learning models like CNNs, GNNs, and transformers (e.g., BERT) into the multimodal framework. The research emphasizes that the traditional single-modal sentiment analysis methods often fail to capture the complete spectrum of human emotions, as they typically rely on text data alone. In contrast, multimodal sentiment analysis leverages the complementary nature of various data sources—text, audio, and visual—to achieve a more accurate and robust understanding of sentiment. By integrating diverse signals, such as the semantic content of text, vocal intonations in speech, and non-verbal cues from facial expressions and gestures, the proposed techniques enable a richer representation of emotional states. This comprehensive approach not only enhances the accuracy of emotion recognition but also improves the resilience of the analysis against noisy or incomplete data.

References

[1] *Kaur, R., & Kautish, S. (2022). Multimodal sentiment analysis: A survey and comparison. Research Anthology on Implementing Sentiment Analysis across Multiple Disciplines, 1846-1870.*

[2] *Das, R., & Singh, T. D. (2023). Multimodal sentiment analysis: a survey of methods, trends, and challenges. ACM Computing Surveys, 55(13s), 1-38.*

[3] *Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S. F., & Pantic, M. (2017). A survey of multimodal sentim ent analysis. Image and Vision Computing, 65, 3-14.*

[4] *Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A system atic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Inform ation Fusion, 91, 424-444.*

[5] *Zhu, L., Zhu, Z., Zhang, C., Xu, Y., & Kong, X. (2023). Multimodal sentiment analysis based on fusion methods: A survey. Information Fusion, 95, 306-325.*

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

[7] *Xu, H. (2023). Multimodal sentiment analysis. In Multi-Modal Sentiment Analysis (pp. 217-240). Singapore: Sprin ger Nature Singapore.*

DOI: 10.54254/2753-8818/2025.22725

[8] *Xu, N., & Mao, W. (2017, November). Multisentinet: A deep semantic network for multimodal sentiment analysis. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 2399-2402).*

[9] *Cambria, E., Hazarika, D., Poria, S., Hussain, A., & Subramanyam, R. B. V. (2018). Benchmarking multimodal se ntiment analysis. In Computational Linguistics and Intelligent Text Processing: 18th International Conference, CI CLing 2017, Budapest, Hungary, April 17–23, 2017, Revised Selected Papers, Part II 18 (pp. 166-179). Springer I nternational Publishing.*

[10] *Chen, M., Wang, S., Liang, P. P., Baltrušaitis, T., Zadeh, A., & Morency, L. P. (2017, November). Multimodal sent iment analysis with word-level fusion and reinforcement learning. In Proceedings of the 19th ACM International Conference on Multimodal Interaction (pp. 163-171).*