# A Review of Research on Multimodal Sentiment Analysis

**Li Meng**

*Abstract*— **Multimodal emotion recognition refers to accurately recognizing an individual's emotional state by comprehensively analyzing multiple modal information closely related to human emotional expression, such as speech, vision and text. This research area has a significant research value in human-computer interaction, artificial intelligence, and affective computing, and has received close attention from a wide range of researchers. In view of the booming development of deep learning methods in recent years and their remarkable results in a variety of tasks, more and more deep neural networks have been applied to learn high-level emotional feature expressions to support the study of multimodal emotion recognition. In order to provide a comprehensive and systematic overview of the current research status of deep learning methods in the field of multimodal emotion recognition, we plan to conduct an in-depth analysis and generalization of the research literature on multimodal emotion recognition involving deep learning in recent years. In doing so, we will first describe the general framework of multimodal emotion recognition. Subsequently, we will focus on feature extraction techniques in multimodal sentiment analysis, which covers both traditional feature extraction methods and deep learning-based feature extraction strategies. Immediately after that, we will also elaborate on the fusion strategies of different modal information, which are crucial for improving the accuracy of multimodal sentiment recognition. Finally, we will analyze the main challenges and potential opportunities currently facing the field, and accordingly point out the direction for the future development of the field. Through this sorting and analysis, we expect to provide useful references and insights for the in-depth research on multimodal emotion recognition.**

*Index Terms*—**Sentiment analysis, Feature extraction, Multimodality, Modal fusion, Deep learning**

## I. Introduction

With the continuous progress and development of society, emerging technologies and concepts are increasingly integrated into our daily lives. Among them, cutting-edge technologies such as artificial intelligence and deep learning are gradually moving from the professional field to the public's field of vision, becoming an important force leading the development of the times. At the same time, social networks have become an important tool for people's daily communication and exchange of information with their unique charm. In the era of big data, a huge amount of information is like a huge treasure trove, containing endless value. In recent years, the rapid rise of short video social platforms, more and more people choose to record video as a multimodal form to publish their opinions and insights. For example, travelers will elaborate on the specific travel feelings of recent popular attractions or recommend niche travel locations by shooting travel vlogs; beauty bloggers will give their feelings on the use of recent hot single products by posting trial videos; and food enthusiasts will post shop-hunting videos to make specific evaluations of food. The amount of information generated by these social software programs is counted in billions every day, which contains rich and diverse data resources. How to better utilize these multimodal data to achieve the goals of artificial intelligence has become a hot area of current research.

Compared with multimodal sentiment analysis which usually uses two or three modalities for sentiment analysis, unimodal sentiment analysis mainly relies on data from a single modality such as text, image or speech to make sentiment judgments, and thus multimodal sentiment prediction will analyze sentiment more comprehensively and accurately. In unimodal sentiment analysis, such as relying only on words, phrases and their semantic relationships in the textual modality, there are indeed limitations that make it difficult to accurately capture complex sentiment information. This approach tends to ignore the multidimensionality in emotional expression, such as pitch, facial expression, and body language, which play equally important roles in conveying emotion. In contrast, an integrated analysis of multiple modalities, such as combining text, audio, and video, can provide a more comprehensive and rich description of emotion. This cross-modal approach can capitalize on the complementarities between different modalities to more accurately identify and interpret emotions. For example, text can reveal the basic meaning of a discourse, while audio can reflect emotional features such as the speaker's intonation and rhythm, and video can capture visual cues such as facial expressions and body language. Therefore, the integrated analysis of multiple modalities has significant advantages in emotion recognition, and can reflect human emotional states more realistically and comprehensively. This cross-modal approach has a broad application prospect in the field of emotion analysis and helps to improve the accuracy and reliability of emotion recognition.

## II. Fundamentals of Multimodal Sentiment Analysis

The basic principle of multimodal sentiment analysis lies in combining and processing information from different modalities (e.g., text, image, speech, etc.) for a more comprehensive and in-depth understanding and recognition of emotions. The realization of this principle relies on advan-

Manuscript received March 18, 2024
**Li Meng**, Tiangong University, Tianjin, China

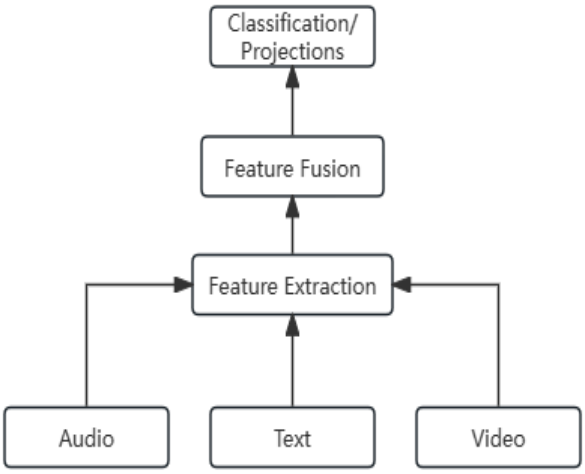


Figure. 1 Multimodal Sentiment Analysis Flowchart

ced feature extraction techniques, modality fusion strategies, and model training for machine learning or deep learning. As shown in Fig.1.

First of all, multimodal sentiment analysis requires preprocessing and feature extraction for each modal data. For text data, semantic features can be extracted by techniques such as word embedding and TF-IDF[1]; for image data, visual features such as facial expressions and gestures can be extracted using convolutional neural networks; while for speech data, acoustic features such as pitch and rhythm can be extracted by audio signal processing techniques. These feature extraction methods can capture the key information related to sentiment in the respective modality.

Second, multimodal sentiment analysis requires the fusion of features from different modalities. According to the step of modal fusion aims to construct a unified feature representation to fully utilize the complementarities among various modalities. There are various approaches to feature fusion, which can be as simple as feature splicing or as complex as feature interaction or transformation. For example, the attention mechanism in deep learning models can be used to assign different weights to the features of each modality to highlight the modalities that contribute more to sentiment recognition. Finally, multimodal sentiment analysis models utilize the fused features for sentiment classification or regression. This typically relies on machine learning algorithms or deep learning models such as support vector machines, random forests, or recurrent neural networks. These models are trained to learn mapping relationships from features to sentiment labels, thus enabling automatic identification and determination of sentiment states.

Overall, the basic principle of multimodal sentiment analysis lies in making full use of the information of different modalities to achieve comprehensive and accurate recognition of sentiment states through steps such as feature extraction, fusion and classification or regression. The realization of this principle not only relies on advanced technological means, but also requires an in-depth understanding of emotional expressions and human cognitive processes. With the continuous development of technology, multimodal sentiment analysis will play an important role in more fields and provide us with a smarter and more humanized interaction experience.

## III. MULTIMODAL EMOTION FEATURE EXTRACTION METHOD

In multimodal sentiment analysis, feature extraction has an extremely important position. It is not only a critical step in sentiment recognition, but also directly affects the training effect of the recognition model and the final classification or prediction results. Feature extraction is the core link to capture sentiment information from multimodal data. Whether it is emotional words in text, tones and rhythms in speech, or expressions and gestures in images, they all need to be accurately identified and analyzed by feature extraction techniques. These features not only reflect the direct expression of emotion, but may also contain deep emotional connotations, so their accuracy and comprehensiveness are crucial for emotion analysis.

In addition, with the development of deep learning and other technologies, feature extraction methods are constantly advancing and improving. New feature extraction methods are not only able to deal with more complex data, but also able to dig deeper into the connotations of sentiment information. The development of these technologies provides more powerful technical support for multimodal sentiment analysis and makes the role of feature extraction in sentiment analysis more prominent. In this section, the extraction methods of multimodal sentiment features are investigated and the related technical routes are analyzed.

### *A. Text Emotion Feature Extraction*

Text-based sentiment feature extraction is a key component in sentiment analysis, which aims to identify and extract features from text data that can represent sentiment tendencies. These features are crucial for subsequent tasks such as sentiment classification and sentiment recognition. In text-based sentiment feature extraction, common features mainly include word frequency, lexicality, syntactic structure, semantic roles and so on. These features can be extracted by different technical means, such as rule-based methods, statistical-based methods, and deep learning-based methods. Rule-based approaches usually rely on manually constructed sentiment dictionaries or rule sets. Sentiment dictionaries contain emotion-related words and their sentiment tendencies, while rule sets define a set of rules for extracting sentiment features from text. This approach is simple and intuitive, but is limited by the completeness and accuracy of the lexicon and rules. Statistical-based approaches, on the other hand, utilize statistical information to extract sentiment features. For example, word frequency statistics can reflect the frequency of occurrence of a certain word in the text, thus indirectly reflecting its emotional tendency. In addition, algorithms such as TF-IDF can be utilized to calculate the importance of words as part of the sentiment features. In recent years, with the development of deep learning technology, deep learning-based methods have made significant progress in text emotion feature extraction. Deep learning models, such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), etc., are able to automatically learn the

representation of text and capture the deep sentiment features in text. These models can learn rich emotional information by training on large-scale corpora and are used to extract emotional features from text.

The Attention mechanism[2] and the Transformer network structure proposed by Google augmented the word representation capability. And then, Devlin J et al.[3] came up with the idea of stacking the Encoder structure of Transformer in multiple layers, which resulted in the Bert model, a pre-trained model that allowed for significant improvements in various tasks in natural language processing. In order to solve the problem that Bert or Transformer can only learn dependencies in fixed-length contexts, Transformer-XL[4] was born.The overall structure of Transformer-XL is the same as that of Transformer, but it can better capture long-distance dependencies in text by introducing a looping mechanism when processing long text. relationships in the text. This improvement allows Transformer-XL to have better performance when processing long text tasks.XLNet[5] is a natural language processing model based on the Transformer-XL architecture, which was proposed by researchers at CMU and MIT in 2019.XLNet was proposed to address some of the problems in the BERT model, especially with autoregressive ( autoregressive) and self-encoding (autoencoding) methods related limitations.

OpenAI also launched a language model, GPT[6], which was also proposed based on the Transformer architecture.The GPT model can predict the next words for contextual information, and the model set records in several tasks. OpenAI then released GPT 2.0[7], a larger and more powerful version that uses far more web text data than its predecessor, and is already capable of generating high-quality text, even for simple tasks such as translating and refining text. Then after that, OpenAI released GPT 3.0[8] with more parameters and features. OpenAI then released the GPT-3.5[9] model, a version fine-tuned from GPT-3 with different training methods and more powerful features. The continuous emergence of these models has driven the continued progress and maturity of the NLP field, providing more powerful and effective tools for natural language processing tasks.

### *B. Video Emotion Feature Extraction*

Facial expression analysis of human faces is an in-depth and complex field that utilizes computer vision and pattern recognition techniques to parse changes in human facial expressions, and thus determine a person's emotional state,possible motivations, and reactions. This analysis not only helps us to understand human emotions, but can also be used in a variety of practical applications, such as security surveillance, healthcare services, etc.

Traditional face feature extraction requires manual design of feature extractors, which requires complex expertise. And the common types of features are appearance features and geometric features, etc., respectively. The extraction of appearance features usually adopts techniques such as directional gradient histogram[10], local binary pattern[11], Gabor transform[12], scale-invariant feature transform[13], etc., to extract features for the pixel characteristics of the face image, which shows the subtle changes of the localization of the face. Compared to the appearance features which show the subtle local changes of the face, the geometric features are more focused on the macro-structural changes of the face by recording the displacement information of the key points of the face to realize the feature extraction. Active appearance model, active shape model, and local constraint model are a few common feature extraction methods.

Deep learning based expression emotion feature extraction has made outstanding progress with the continuous development of deep neural networks. Deep feature extraction techniques are able to autonomously learn more essential feature representations from images.Compared with the traditional way, deep feature extraction techniques have stronger generalization ability, as well as more powerful autonomous learning ability, which is more suitable for practical business scenarios.CNN, RNN are common deep learning models.Shi et al.[14] attempted to utilize Convolutional Neural Networks (CNNs) to perform deep feature extraction, and achieved remarkable results. Siyue et al. [15]proposed a multilayer image feature extraction network based on CNN, which not only focuses on the local features of the image to highlight the details of the expression, but also extracts the high-level semantic features from a global perspective, and achieves more comprehensive expression feature extraction by combining the features at these two levels. With the development of the technology there are more models emerging, for example, 3DCNN[16] and so on.

### *C. Speech Emotion Feature Extraction*

The specific steps of speech emotion feature extraction usually involve several links, which are designed to extract the key information that can characterize the emotional state from the original speech signal. The first of these is signal preprocessing of the speech modal data, which is divided into two main steps: frame splitting, where the speech signal is a continuous signal stream, but is usually split into shorter frames (usually 20-30 ms) during processing. The reason for this is that the speech signal can be regarded as stable in a short period of time, so features can be extracted in each frame; windowing, after framing, the original speech signal becomes a finite signal, in order to avoid the leakage of high-frequency portions, a window function (such as a rectangular window or a Hamming window) is usually used to add a window to each frame.

The next step is to extract features from the processed signal information, which usually includes fundamental frequency, resonance peak features, MFCC, and energy and duration features. Among them, fundamental frequency (Pitch) and resonance peaks (Formants) are important features reflecting the emotional state of speech. The fundamental frequency can be extracted by various methods, such as the autocorrelation function method. The resonance peaks describe the resonance characteristics of the vocal tract, which are usually obtained through the analysis of speech signals, while the MFCC feature is a feature that simulates the auditory characteristics of the human ear, which is very useful for the recognition of the emotion of speech, and the extraction of the MFCC usually involves the steps of the Fourier Transform, the Mel Frequency Scale Conversion, and the Discrete Cosine Transform, etc. Moreover, the energy and articulatory time of the speech signal are also important

features. Moreover, the energy and articulatory duration of the speech signal are also key features for emotion recognition. The energy can be obtained by calculating the sum of squares of each frame of the speech signal, while the duration can be obtained by calculating the number of consecutive valid speech frames.

Subsequently, the features are normalized as well as downscaled according to the actual situation. Among them, normalization is to eliminate the differences of speech signals from different speakers or under different recording conditions, and it is usually necessary to normalize the extracted features. While dimensionality reduction is to reduce the computational complexity and avoid overfitting, sometimes it is necessary to reduce the dimensionality of the features, such as the use of principal component analysis (PCA) and other methods. Finally, the normalized features are obtained for feature integration, which is usually based on bag-of-words model. The extracted features are integrated in the form of bag-of-words model to get the emotional characterization of the whole speech segment. The bag-of-words model does not consider the order relationship between the features and only focuses on the frequency of occurrence of the features.

In recent years, deep learning techniques have been widely used in speech emotion recognition tasks. In contrast to the traditional approach that relies on hand-designed acoustic features, deep learning-based speech emotion feature extraction methods automatically learn feature representations through neural networks. Neural networks are able to extract hierarchical features from raw speech data that are more abstract and better able to capture emotional information in speech. Deep learning models, such as CNN, RNN, or LSTM, are able to automatically learn and optimize the feature extraction process, thus avoiding the tedious and subjective nature of manually designing features. In addition, deep learning-based sentiment feature extraction approach has a stronger data processing capability. It can build deep models by stacking multiple network layers to handle more complex and high-dimensional data. In addition, the deep learning model can suppress the influence of noise and background interference by learning the intrinsic laws and structure of the data to improve the accuracy of emotion recognition. And it has strong generalization ability. Through training on a large amount of data, deep learning models can learn more generalized feature representations, thus adapting to different application scenarios and data distributions.Mao et al[17]. proposed a CNN-based approach to speech emotion feature extraction, which utilizes unlabeled samples to learn locally invariant features through a sparse autoencoder; these locally invariant features are subsequently used as inputs to the salient discriminant feature analysis (SDFA), in order to further learn significant discriminative features. In addition, a novel multi-granularity feature extraction method was proposed by Jing Chen et al[18]. The method performs feature extraction based on different temporal units, including short-time frame granularity, mid-time granularity, and long-time window granularity, thus realizing a comprehensive analysis of speech emotion.

## IV. Multimodal Emotional Feature Fusion Strategies

The complexity of emotion, as a non-verbalized and dynamic psychophysiological process, poses significant challenges for emotion recognition. Although some research results have been achieved in unimodal emotion recognition tasks in recent years, several studies have demonstrated that multimodal emotion recognition tasks are significantly better than unimodal in terms of effectiveness. In order to improve the efficiency and accuracy of emotion recognition tasks, researchers are working on fusing signals from different modalities, including speech, vision, text, and other kinds of information. This chapter will focus on information fusion methods in multimodal emotion recognition. Among them, common fusion strategies include feature-level fusion, which integrates information from different modalities at the feature level; decision-level fusion, which fuses the outputs of each modality at the decision-making stage; and modality-level fusion, which realizes the synergy of multimodal information at the modality level. These methods provide effective ways to improve the performance of multimodal emotion recognition.

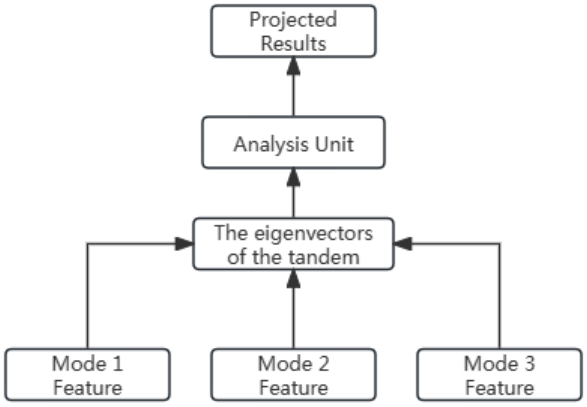


Figure.2 Early Fusion

### A. *Early Fusion*

The early fusion strategy [19]in multimodal sentiment analysis refers to the fusion of information from different modalities at the feature extraction stage to form a unified feature representation for subsequent classification or regression tasks. The core principle of this fusion approach is to make full use of the complementarities between modalities to enhance the accuracy of sentiment recognition through early information integration.

Specifically, the realization process of early fusion usually includes the following steps: first, preprocessing and feature extraction are performed on the data of each modality to obtain the feature vectors of the respective modality; then, these feature vectors are spliced or combined with weighting to form a comprehensive feature representation; finally, sentiment classification or regression is performed based on this comprehensive feature representation.

A concrete example is used to illustrate the application of early fusion. Suppose we have a multimodal sentiment analysis task based on text and speech. For the text modality,

we can use natural language processing techniques to extract features such as sentiment vocabulary, syntax, etc. in the text; for the speech modality, we can use audio processing techniques to extract features such as pitch, speech rate, etc. of the speech. Then, we splice the feature vectors of these two modalities to form a comprehensive feature vector containing text and speech information. Next, we can use machine learning algorithms or deep learning models to train this integrated feature vector to achieve the goal of sentiment classification or regression.

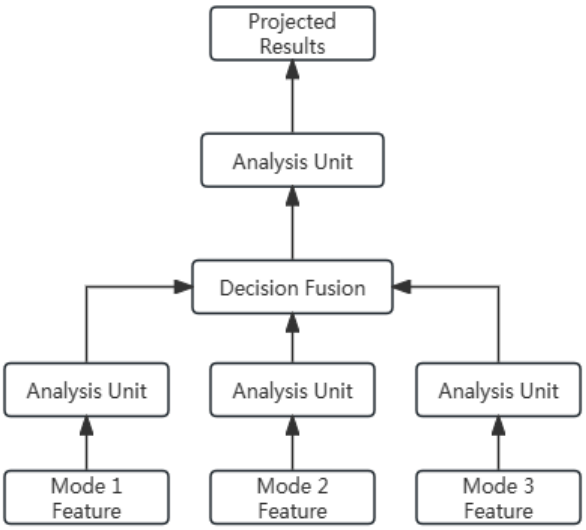


Figure.3 Late Fusion

The advantage of early fusion is that it can integrate the information of different modalities earlier, which enables the model to fully utilize the complementary nature of such information in the subsequent processing. However, this fusion approach also faces some challenges, such as how to choose the appropriate feature extraction method and how to determine the weights between different modal features. The flow of early fusion is shown in Fig.2

### *B. Late Fusion*

Late Fusion has also been made into integration at the decision-making level.Decision-level fusion for multimodal sentiment analysis, also known as late fusion, is a strategy that integrates and consolidates the outputs of the models of the respective modalities after they have completed their sentiment analysis[20]. Compared to early fusion, late fusion allows each modality's model to remain independent and more flexible, thus allowing for better handling of the uniqueness of different modal data.

First, the core of late fusion lies in the ability of models of each modality to perform sentiment analysis tasks independently. This means that for different modalities such as text, image, speech, etc., we can use their respective suitable model structures, feature extraction methods and training strategies. For example, for text modality, we can use models based on natural language processing such as Recurrent Neural Networks (RNN) or Transformer models; for image modality, we can use Convolutional Neural Networks (CNNs) to extract visual features such as facial expressions and gestures; and for speech modality, we can utilize audio signal processing techniques to extract acoustic features. These modality-specific models will generate their respective sentiment predictions.

Next, in the late fusion stage, we need to integrate these predictions from different modalities. The purpose of integration is to combine the information from each modality to get a more comprehensive and accurate sentiment judgment. This can be achieved through a variety of methods, such as weighted averaging, voting mechanisms, or rule-based logical combinations.

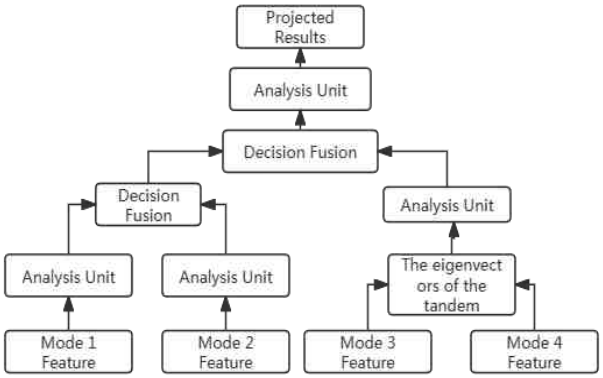


Figure.4 Hybrid Fusion

A common late fusion method is weighted average. In this method, we assign a weight to the predictions of each modality and then weighted average them according to the weights. The determination of the weights can be adjusted based on factors such as the importance of the modality, data quality, or model performance. For example, if a modality performs particularly well in a specific task, we can give it a higher weight. By weighted averaging, we can obtain a sentiment prediction result that combines information from multiple modalities.

Another late fusion method is the voting mechanism. In this approach, each modal model generates a sentiment label or classification result, and then we determine the final sentiment judgment by voting. This can be a simple majority vote or a weighted vote based on model performance. The benefit of the voting mechanism is that it is simple and intuitive and can easily handle the case of multiple categories or labels. In addition to weighted averaging and voting mechanisms, late fusion can also be performed using a rule-based logic combination approach. This method usually requires a series of rules or logical conditions based on specific tasks and data characteristics, and then combines the prediction results of different modalities according to these rules. This approach is more flexible and can be customized and optimized according to actual needs.

The advantage of late fusion is that it can fully utilize the independence and advantages of each modal model, and also flexibly deal with the complementarities and differences between different modalities. It allows us to choose the most suitable fusion method based on specific tasks and data characteristics to obtain more accurate sentiment analysis results. In addition, late fusion allows us to easily add or remove certain modalities to adapt to different application scenarios and requirements.

At the same time, late fusion has some challenges and

limitations. First, it requires that the model for each modality produces comparable and integrable outputs. This may require us to standardize or normalize the model outputs to ensure that they have the same scale and scope. Second, determining the weights or integration rules between different modalities is a key issue. This usually requires debugging and optimization based on experience or experiments to obtain the best fusion results. In addition, late fusion may also increase computational complexity and processing time, especially when dealing with large amounts of data or complex models. Late fusion is shown in Fig.3.

### C. *Hybrid Fusion*

The hybrid fusion[21] approach in multimodal sentiment analysis is an integrated approach that combines early fusion and late fusion strategies, aiming to fully utilize the complementary nature of different modal information and improve the accuracy of sentiment recognition. This approach combines the integration advantages of early fusion in the feature extraction stage and late fusion in the model output stage, which enables better fusion of multimodal information throughout the processing flow.

The basic idea of the hybrid fusion method is to perform modal fusion in both the feature extraction and model output stages. In the feature extraction stage, data from different modalities are preprocessed and then initially fused in some way (e.g., splicing, weighting, etc.) to form a comprehensive feature representation containing information from multiple modalities. This step is similar to the early fusion, but with the difference that the hybrid fusion approach allows for a certain amount of modal independence to be retained during the feature fusion process so that the subsequent processing stages can further utilize the specific information of each modality. In the model output phase, the models of the individual modalities are analyzed for sentiment based on the combined feature representations and their respective sentiment predictions are generated. These predictions are subsequently integrated by means of late fusion. Late fusion can use methods such as weighted averaging, voting mechanisms, or rule-based logical combinations to combine the prediction results from different modalities to obtain a final sentiment judgment. This step takes full advantage of the flexibility of late fusion in integrating model outputs, enabling the selection of appropriate integration strategies based on specific tasks and data characteristics.

The advantages of hybrid fusion methods are mainly reflected in the following aspects: first, hybrid fusion methods can fully utilize the complementary nature of different modal information. By combining early fusion and late fusion, hybrid fusion methods are able to consider the characteristics and prediction results of different modalities comprehensively throughout the processing flow, thus capturing emotion information more comprehensively. This helps to improve the accuracy and reliability of emotion recognition. Second, the hybrid fusion method has a certain degree of flexibility. It allows modal fusion to be performed in both the feature extraction and model output stages, so the fusion strategy can be adjusted according to specific tasks and data characteristics. This means that hybrid fusion methods can adapt to different application scenarios and needs, providing a more flexible and effective solution for sentiment analysis. In addition, hybrid fusion methods are able to utilize the independence of individual modal models. In the feature extraction stage, the hybrid fusion method retains the specific information of each modality, enabling the subsequent processing stage to further utilize this information for sentiment analysis. This independence helps to fully utilize the strengths of each modality and improve the accuracy of sentiment recognition.The flow of hybrid fusion is shown in Fig.4.

## V. SUMMARY AND OUTLOOK

As an important research direction in the field of artificial intelligence, multimodal sentiment analysis has made remarkable progress in recent years. However, with the continuous development of technology and the increasing complexity of application requirements, multimodal sentiment analysis still faces many challenges and opportunities. Looking ahead, multimodal sentiment analysis has the following possible development directions:

First, more refined sentiment recognition will become the focus of research. Currently, multimodal sentiment analysis mainly focuses on the recognition of basic emotions, such as happiness, anger, and sadness. However, in practical applications, emotions tend to be more complex and delicate, requiring more refined recognition. Therefore, future research will be devoted to the development of multimodal sentiment analysis models that can recognize more kinds and finer variations of emotions.

Second, cross-domain and cross-cultural sentiment analysis will become a hot spot of research. There may be differences in the expression and understanding of emotions in different domains and cultures, which brings challenges to multimodal sentiment analysis. Future research will focus more on cross-domain and cross-cultural sentiment analysis, and improve the applicability and accuracy of models in different scenarios by introducing domain knowledge and cultural factors.

In addition, the combination of multimodal sentiment analysis with other technologies will also become a research trend. For example, the combination with natural language processing, computer vision, speech recognition and other fields can further expand the application scope of multimodal sentiment analysis. Meanwhile, the combination with machine learning, deep learning and other technologies can continuously improve the performance and efficiency of multimodal sentiment analysis.

Finally, the landing and promotion of multimodal sentiment analysis in practical applications will also become the focus of research. At present, multimodal sentiment analysis has shown a broad application prospect in human-computer interaction, intelligent customer service, education, medical and other fields. In the future, with the continuous maturation and popularization of the technology, multimodal sentiment analysis will be applied in more fields, bringing more convenience and benefits to people's life and work.

In summary, multimodal sentiment analysis has a broad

development prospect and great application potential in the future. Through continuous research and innovation, we can expect multimodal sentiment analysis to make more significant progress in the areas of sentiment recognition, cross-domain application, technology combination and practical application landing.

## REFERENCES

[1] TRSTENJAK B, MIKAC S, DONKO D. KNN with TFIDF based framework for text categorization[J]. Procedia Engineering, 2014, 69: 1356-1364.
[2] Ashish V, Noam S, Niki P, Jakob U, Llion J, Aidan N G, Lukasz K, Illia P, et al. Attention Is All You Need.[J], Advances in neural information processing systems, 2017, 30(): 5998-6008.
[3] Jacob D, Ming-Wei C, Kenton L, Kristina T, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C], North American Chapter of the Association for Computational Linguistics, 2018, abs/1810.04805: 4171-4186.
[4] Zihang D, Zhilin Y, Yiming Y, Jaime C, Quoc V L, Ruslan S, et al. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context[C], Annual Meeting of the Association for Computational Linguistics, 2019, abs/1901.02860(): 2978-2988.
[5] Zhilin Y, Zihang D, Yiming Y, Jaime C, Ruslan S, Quoc V L, et al. Xlnet: Generalized Autoregressive Pretraining For Language Understanding[J], ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 32 (NIPS 2019), 2019, 32: 5754-5764.
[6] Alec R, Karthik N, Tim S, Ilya S, et al. Improving language understanding by generative pre training[J], URL https://s3-us-west-2. amazonaws.com/openai-assets/research-covers/language-unsupervised/language_ understanding_paper. pdf, 2018: 12.
[7] Alec R, Jeffrey W, Rewon C, David L, Dario A, Ilya S, et al. Language Models are Unsupervised Multitask Learners, arXiv preprint arXiv, 2019
[8] Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D M, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D, et al. Language Models are Few-Shot Learners[C], Conference on Neural Information Processing Systems, 2020, 33: 1877-1901.
[9] Long O, Jeff W, Xu J, Diogo A, Carroll L W, Pamela M, Chong Z, Sandhini A, Katarina S, Alex R, John S, Jacob H, Fraser K, Luke M, Maddie S, Amanda A, Peter W, Paul C, Jan L, Ryan L, et al. Training language models to follow instructions with human feedback[C], Conference on Neural Information Processing Systems, 2022
[10] Nigam S, Singh R, Misra A K. Efficient facial expression recognition using histogram of oriented gradients in wavelet domain[J]. Multimedia tools and applications, 2018, 77(21): 28725-28747.
[11] Yang, Biao, Cao, et al. Facial expression recognition based on dual-feature fusion and improved random forest classifier[J]. Multimedia Tools & Applications, 2018,77(16):20477-20499.
[12] Bengio, Yoshua, Courville, et al. Representation Learning: A Review and New Perspectives[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(8):1798-1828.
[13] Ren F , Huang Z . Facial expression recognition based on AAM–SIFT and adaptive regional weighting[J]. Ieej Transactions on Electrical & Electronic Engineering, 2015, 10(6):713-722.
[14] Shi B, Bai X, Yao C. Script identification in the wild via discriminative convolutional neural network[J]. Pattern Recognition, 2016(52): 448-458.
[15] Siyue, Xie, Haifeng, et al. Facial Expression Recognition Using Hierarchical Features With Deep Comprehensive Multipatches Aggregation Convolutional Neural Networks[J]. IEEE Transactions on Multimedia, 2019, 21(1):211-220.
[16] Du T, Lubomir B, Rob F, Lorenzo T, Manohar P, et al. Learning Spatiotemporal Features with 3D Convolutional Networks[J], 2015 IEEE International Conference on Computer Vision (ICCV), 2015: 4489-4497.
[17] MAO Q, DONG M, HUANG Z, et al. Learning salient features for speech emotion recognition using convolutional neural networks[J]. IEEE Transactions on Multimedia, 2014, 16(8): 2203-2213.
[18] CHEN J, LI H F, MA L, et al. Multi- granularity feature fusion for dimensional speech emotion recognition[J]. Journal of Signal Processing, 2017, 33(3): 374-382.
[19] Zadeh A, Zellers R, Pincus E, et al. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages[J]. IEEE Intelligent Systems, 2016, 31(6): 82-88.
[20] Sahoo S, Routray A. Emotion recognition from audio-visual data using rule based decision level fusion[C]//2016 IEEE Students' Technology Symposium (TechSym). IEEE, 2016: 7-12.
[21] Wöllmer M, Weninger F, Knaup T, et al. Youtube movie reviews: Sentiment analysis in an audio-visual context[J]. IEEE Intelligent Systems, 2013, 28(3): 46-53.

**Li Meng**

Graduated from the School of Software Engineering, Tianjin University of Technology, with a master's degree. During the postgraduate period, the research direction focused on multimodal sentiment analysis. Obtained the copyright for the BetterMan Online Music Audiovisual System in October 2023.