Fusion of Multimodal Information for Video Comment Text Sentiment Analysis Methods

Jing Han[1]*, Jinghua Lv[2]
School of Creative Art and Fashion Design of Huzhou Vocational & Technical College, Huzhou 313099, Zhejiang, China[1]
Department of Chinese Studies of Kyungsung University, Pusan 48434, Pusan, Korea[2]

***Abstract*—Sentiment analysis of video comment text has important application value in modern social media and opinion management. By conducting sentiment analysis on video comments, we can better understand the emotional tendency of users, optimise content recommendation, and effectively manage public opinion, which is of great practical significance to the push of video content. Aiming at the current video comment text sentiment analysis methods problems such as understanding ambiguity, complex construction, and low accuracy. This paper proposes a sentiment analysis method based on the M-S multimodal sentiment model. Firstly, briefly describes the existing methods of video comment text sentiment analysis and their advantages and disadvantages; then it studies the key steps of multimodal sentiment analysis, and proposes a multimodal sentiment model based on the M-S multimodal sentiment model; finally, the efficiency of the experimental data from the Communist Youth League video comment text was verified through simulation experiments. The results show that the proposed model improves the accuracy and real-time performance of the prediction model, and solves the problem that the time complexity of the model is too large for practical application in the existing multimodal sentiment analysis task of the video comment text sentiment analysis method, and the interrelationships and mutual influences of the multimodal information are not considered.**

Keywords*—Video commentary text sentiment analysis; multimodal information fusion; M-S multimodal sentiment model; convolutional neural network

I. Introduction

Due to the prevalence of the Internet and the advancement of social media, video platforms like YouTube and Jitterbug have emerged as significant avenues for users to express their views and sentiments [1]. The massive video comment data generated on these platforms provide rich materials for sentiment analysis. The foundation for sentiment analysis of video comments is laid by the advancement of sentiment analysis technologies, particularly the breakthrough in text sentiment analysis [2]. Although sentiment analysis of video comment language has gained a lot of attention lately, there are still several issues with the current approaches. Many of the current video commentaries sentiment analysis methods mainly rely on textual modality, while ignoring the rich visual and auditory information contained in the video [3]. This unimodal approach to analysis has obvious limitations because emotional information in videos is not only expressed through text, but also conveyed through multiple channels such as facial expressions, voice intonation, and body language [4]. This process usually involves sentiment lexicon methods, which determine sentiment polarity by comparing with sentiment lexicon, and machine learning methods, which improve the accuracy and adaptability of sentiment recognition by training machine learning models. Sentiment analysis has important applications in several fields, especially in opinion monitoring and marketing [5].

Currently Video Commentary Text Sentiment Analysis is a hot area of current research, and there are various methods and techniques, the dictionary-based method mainly relies on sentiment dictionary [6], which calculates the sentiment tendency by matching the sentiment words in the text. This method is simple and direct, but requires the support of a high-quality sentiment dictionary. Machine learning-based methods perform well on specific datasets by constructing feature vectors [7] and combining algorithms such as Support Vector Machines (SVMs) and Plain Bayes to perform sentiment classification, but this method requires the manual design of the features; deep learning methods automatically learn features through neural networks [8], which are widely used in text sentiment analysis, but the calculation is more complicated. Research indicates that multimodal sentiment analysis, which integrates visual and audio data, can markedly enhance the precision of sentiment recognition [9].

Aiming at the problems of sentiment polysemy, noise interference, dynamic change of sentiment vocabulary, cross-domain differences in sentiment, and insufficient deep semantic understanding in video comment text sentiment analysis [10], this paper proposes a M-S based multimodal sentiment model. The main main contributions of this paper are (1) analysing the main methods of video review text sentiment analysis and sorting out the related techniques; (2) designing an M-S based multimodal sentiment model; and (3) validating and analysing the evaluation model using related datasets.

II. Video Review of Textual Emotions Analysis Methodology

At present, text emotion classification studies usually classify text emotion tendency into positive and negative. Positive emotion means that the text has a positive attitude, and negative emotion means that the text has a negative attitude. However, due to the characteristics of the Internet itself, such as openness and randomness, video comments usually contain a large number of non-standard linguistic phenomena, such as misspelled words, abbreviations, Internet slang, etc., so a large number of different sentiment analysis methods have emerged.

A. Dictionary-based Approach

The sentiment lexicon plays an important role in textual sentiment analysis by matching words with predefined

sentiment categories (e.g., positive, negative) [11], as shown in Fig. 1.

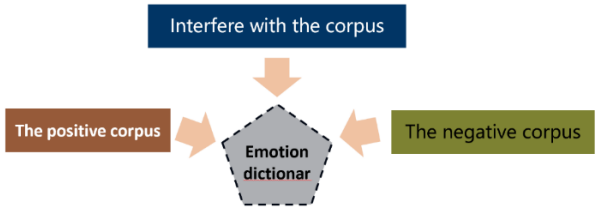

Fig. 1. Dictionary-based approach to text sentiment analysis

B. Machine Learning Methods

Support vector machines (SVMs), plain bayes, and logistic regression are among the frequently used algorithms in machine learning techniques, which are extensively employed in sentiment analysis [12], as shown in Fig. 2. These methods improve classification performance through feature selection and extraction techniques such as bag-of-words models and n-gram features. However, these methods may face computational inefficiency when dealing with large-scale data and require a large amount of labelled data for model training.

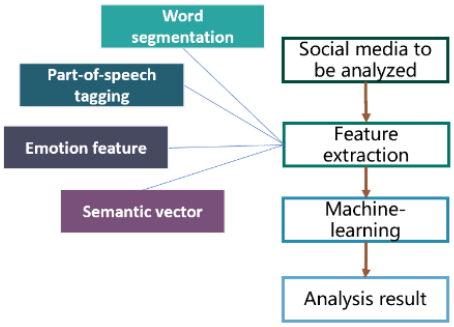

Fig. 2. Sentiment analysis of text based on machine learning approach analysis.

C. Deep Learning Methods

Deep learning methods have made significant progress in sentiment analysis, mainly through models such as Convolutional Neural Networks (CNNs), Long Short-Term Memory Networks (LSTMs) and BERT [13]. These models are able to automatically learn complex features in text without the need to manually design features, which improves the accuracy of sentiment classification, as shown in Fig. 3.

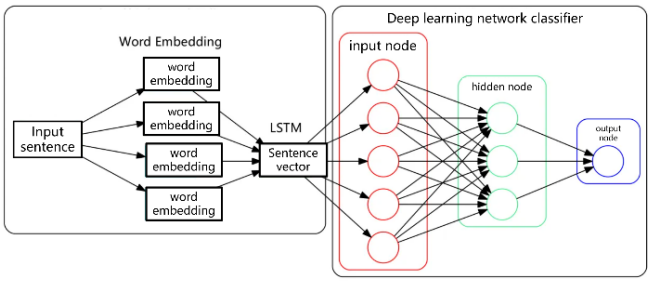

Fig. 3. Sentiment analysis of text based on deep learning approach.

D. Fine-Grained Sentiment Analysis

Fine-grained sentiment analysis focuses on identifying and analysing specific attributes in reviews and their emotional tendencies [14], which can provide more accurate sentiment analysis results. This approach classifies the polarity of comments by identifying object attributes and their corresponding sentiment words, as shown in Fig. 4.

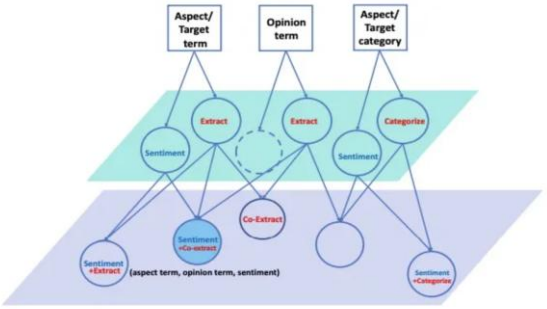

Fig. 4. Sentiment analysis of text based on fine-grained sentiment analysis.

E. Multimodal Sentiment Analysis

Multimodal sentiment analysis integrates many forms of information, including text, audio, and video, to facilitate sentiment recognition [15], thereby providing a more thorough comprehension of the user's emotional disposition. This method enhances the precision of sentiment analysis via modal fusion techniques, including feature-level fusion and decision-level fusion, as seen in Fig. 5.

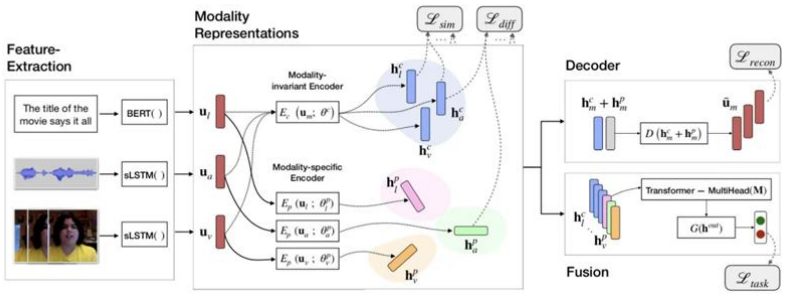

Fig. 5. Sentiment analysis of text based on multimodal information.

III. Multimodal Information Sentiment Analysis

The specific operation of the multimodal sentiment analysis task for video comment text is shown in Fig. 6. Firstly, the information of different modalities is fed into the feature extraction model, then the extracted features are fused into the feature fusion according to a certain fusion method, and finally the fused features are fed into the classifier to perform sentiment analysis on the multimodal data [16].

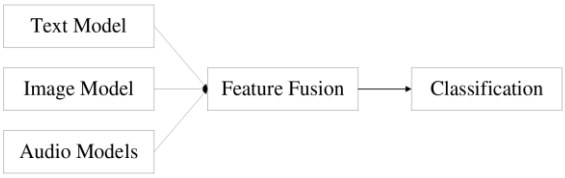

Fig. 6. Multimodal sentiment analysis architecture.

A. Feature Extraction in Multimodal Sentiment Analysis

Extraction of representative features from all modal data utilized in subsequent fusion tasks is referred to as feature extraction in multimodal sentiment analysis. Multimodal sentiment analysis tasks in machine learning, deep learning, artificial intelligence, natural language processing, image processing, and audio processing depend on feature extraction, and the accuracy and speed of the model operation can be directly impacted by the quality of feature extraction [17].

The objective of feature extraction is to derive more representative and interpretable features from the source data to enhance its description and differentiation. Feature extraction typically encompasses the subsequent steps: 1) Data preprocessing; 2) Feature selection; 3) Feature extraction; and 4) Dimensionality reduction of features.

B. Feature Fusion in Multimodal Sentiment Analysis

In the process of multimodal sentiment analysis, the feature extraction stage is not significantly different from the unimodal feature extraction method. However, the core difference between multimodal sentiment analysis and unimodal analysis lies in how to effectively fuse the information from different modalities to derive accurate sentiment polarity.

The methods of feature fusion mainly include feature-level fusion (Feature-level fusion) and decision-level fusion (Decision-level fusion), as shown in Fig. 7 and Fig. 8.

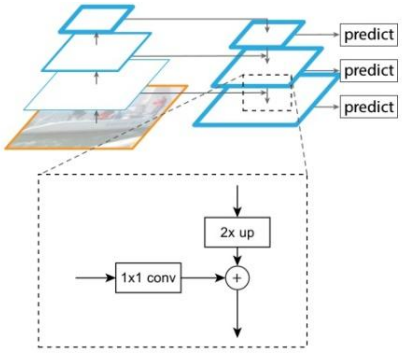

Fig. 7. Feature level fusion.

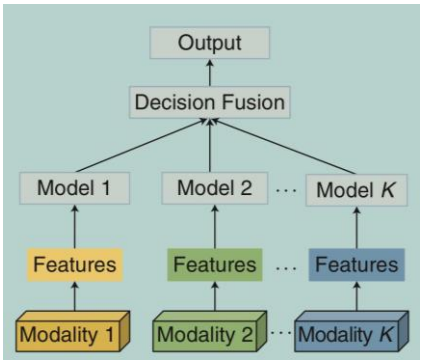

Fig. 8. Decision-level fusion.

Feature-level fusion (FLF) refers to the fusion of multiple features from different feature extraction methods or feature representations to generate a more representative and enriched feature vector. Feature-level fusion is one of the most commonly used methods for multimodal data fusion, and by fusing features from different modalities, a more comprehensive, accurate and robust feature representation can be extracted, thus improving the performance of the model.

Decision-level fusion (DLF) is a fusion strategy in which the features of each modality are first analysed individually, and the results of their respective analyses are subsequently integrated into a single decision vector to arrive at a final decision. The significant advantage of this fusion approach is that when data is missing for one modality, it is still possible to rely on information from other modalities for decision making.

IV. Sentiment Analysis based on the M-S Multimodal Sentiment Model Research

A. M-S Model Multimodal Fusion Sentiment Analysis Model

In this paper, a novel fusion model M-S model Multimodal Fusion Sentiment Analysis Model is proposed to perform the task of video comment text sentiment analysis as shown in Fig. 9. This model firstly uses a one-dimensional convolutional neural network to extract the text information, audio information, and image information, and the three modal features Feature-L, Feature-A, and Feature-V (text feature, audio feature, and image feature) obtained, and then the obtained features are sent to the Transformer-layer to be processed, and then the features containing cross-modal The features containing cross-modal attention information (Cross-model attention Feature-X→Feature-Y), and the features of different modalities are sent to the main and sub dual channels for processing according to their importance, and finally, the processed features of the two channels are sent to the Fully Connected Layer (FC-layer) for processing, and are outputted to the Softmax classifier for classification.

This paper aims to propose a lightweight model with low time complexity that is easy to implement. Therefore, the most fundamental one-dimensional convolutional neural network is selected for feature extraction, while the Transformer model (Fig. 10) is utilized for feature processing to investigate the relationships and influences of various modalities. By fine-tuning critical parameters in the Transformer, it significantly contributes to both cross-modal attention and multi-modal

feature fusion. By modifying the critical parameters in the Transformer, it significantly contributes to both cross-modal attention and multi-modal feature fusion. The primary and secondary dual-channel fusion approach introduced in this study enables the model to achieve high accuracy and stability.

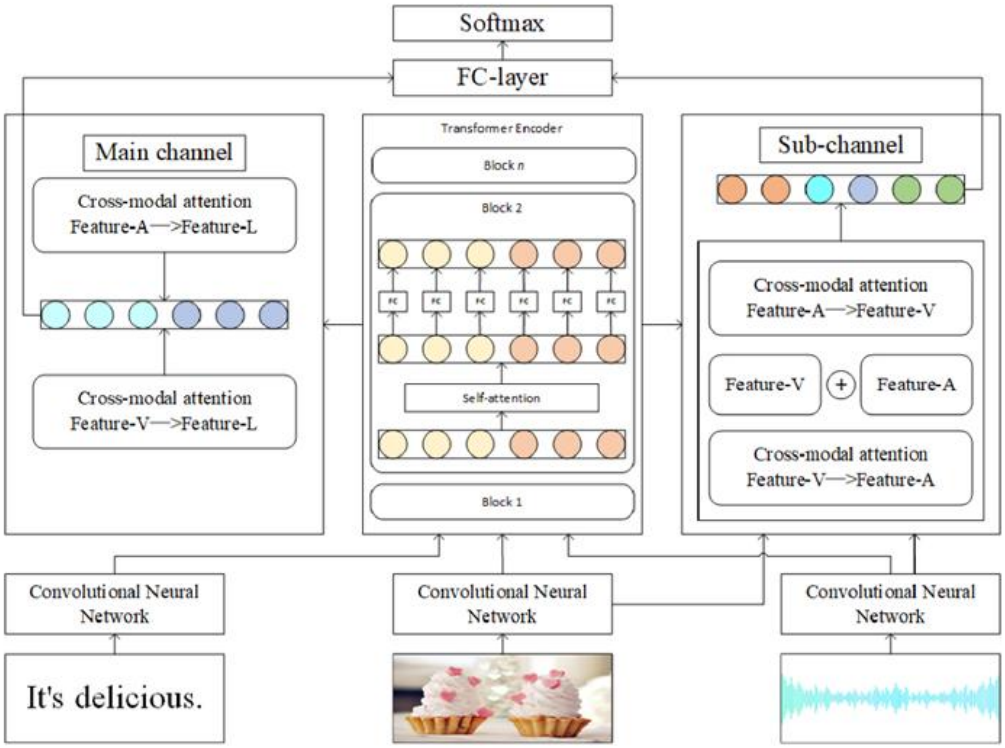

Fig. 9. M-S model.

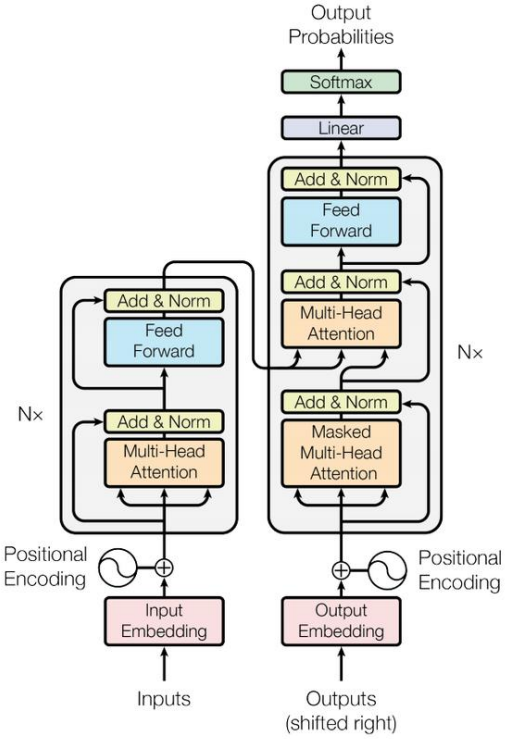

Fig. 10. Transformer model.

Multimodal information feature extraction: In this paper, we use 1D Convolutional Neural Network (1D-CNN) for text information, image information, and audio information, which is a variant of convolutional neural network, and compared with traditional fully-connected neural network, 1D-CNN can better deal with localised Relationships. Considering that multimodal datasets are often video-based, resulting in three modes of data containing time series, 1D convolutional networks can effectively extract the features of each mode in the dataset, as shown in Fig. 11.

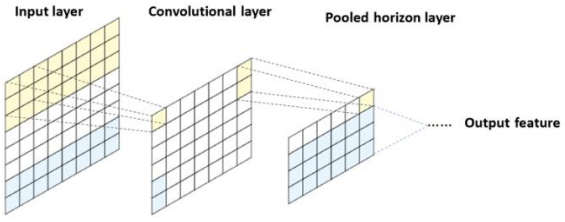

Fig. 11. One-dimensional convolutional neural network.

This paper employs a Transformer-based feature fusion model for multimodal sentiment analysis, emphasizing the significance of fused features in influencing outcomes. The objective is to preserve or enhance the maximum information prior to fusion, utilizing the Transformer encoding and decoding techniques for feature integration. The fusion method ensures

that the extracted features optimally preserve the characteristics prior to fusion, after which the three modal features are combined and encoded to provide a segment of the fused features post-fusion. Subsequently, the fully connected method is employed, incorporating a Dropout layer to mitigate overfitting. The FC-layer consists of two Dropout layers, two Linear layers, and one ReLU activation layer, after which the resultant features are input into a Softmax classifier to derive their sentiment class labels.

B. Multimodal Information Feature Extraction

A one-dimensional convolutional neural network is a specialized neural network designed for sequence data processing, comprising an input layer, a convolutional layer, and a pooling layer, as seen in Fig. 11. The input layer incorporates information from three distinct modalities, while the convolutional layer is trainable to derive an optimal set of convolutional kernels that minimize the loss function, hence facilitating automatic feature extraction.

$A = [a_1, a_2, L, a_s]^T$ is passed as model input to the input layer, where $A \in R^{s \times d}$ is the time series, s is the length of the time series, and d is the number of eigenvalues. a_i denotes the feature vector at the time of i, with the dimension of d. The sequence data is mapped into convolutional layer by one dimensional convolution operation:

$$f_r(z) = \max(z, 0) \tag{1}$$

$$y_c^{\ j} = f_r(A \otimes W_c^{\ J} + b) \tag{2}$$

Where: $\otimes$ denotes one-dimensional convolution operation; $y_c^{\ j}$ denotes the j th feature mapping generated by the convolution kernel $w_c^{\ j}$, $j \in [1, n_c]$, n_c denotes the number of convolution kernels, and the convolution kernel $W_c^{\ j} \in R^{m \times d}$ is a weight matrix, where m is the size of the convolution kernel, and m denotes the width of the local time window for extracting the time series features for the time series; b is the bias; and $f_r(z)$ is the activation function ($Relu$ activation function), which is used to non-linearise the data after the convolution operation.

The pooling procedure is employed to extract the most pertinent information on the sequence of features in the convolutional layer, hence creating the pooling layer. This article employs max-pooling to aggregate the features. Upon the final application of the pooling procedure, global max-pooling is employed to extract the most pertinent global temporal information, resulting in a sequence length of 1, and the features of the three modalities are extracted from the information $y_P^{\ j}{}_{,last}$

.

$$y_P^{\ j}(k) = \max(y_c^{\ j}(2k-1), y_c^{\ j}(2k)) \tag{3}$$

$$y_P^{\ j}{}_{,last} = \max(y_c^{\ j}) \tag{4}$$

V. Experiments and Analysis of Results

A. Baseline Modelling

1) MCTN model: The method introduces an innovative strategy to learn robust transitions between joint representations modalities through a translation process [18]. Specifically, the translation process from source to target modality not only provides a new way to learn joint representations, but also requires only the modality of the source modality as input. In order to further strengthen the translation effect of modalities, the method utilises cyclic consistency loss to ensure that the joint representation retains the maximum information of all modalities, as shown in Fig. 12.

2) LMF model: While previous studies have explored the tensor expressivity of multimodal representations, these methods often suffer from a dramatic increase in dimensionality and computational complexity due to the transformation of inputs into tensors [19]. To address this issue, this method proposes a low-rank multimodal fusion strategy that utilises a low-rank tensor for multimodal fusion, thus significantly improving the efficiency, as shown in Fig. 13.

3) TFN model: The model redefines the problem of multimodal sentiment analysis as a problem of modelling intra- and inter-modal dynamics [20]. To this end, it introduces a novel model, the tensor fusion network, which is capable of learning both dynamics end-to-end. The method is optimised especially for the instability of spoken language in online videos and the accompanying gestures and speech, as shown in Fig. 14.

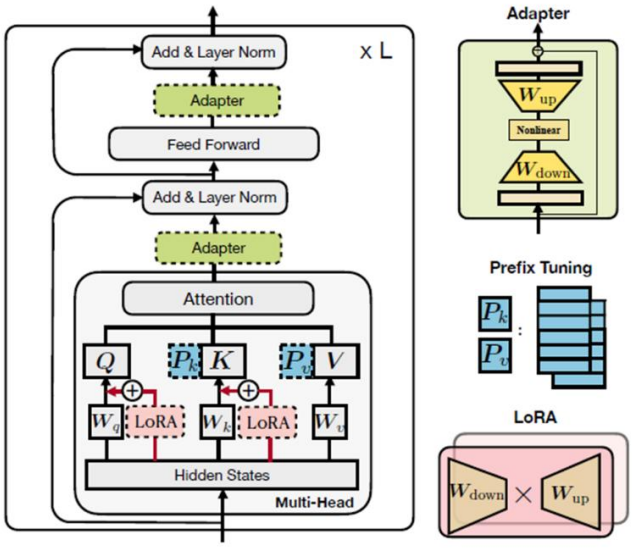

Fig. 12. MCTN model.

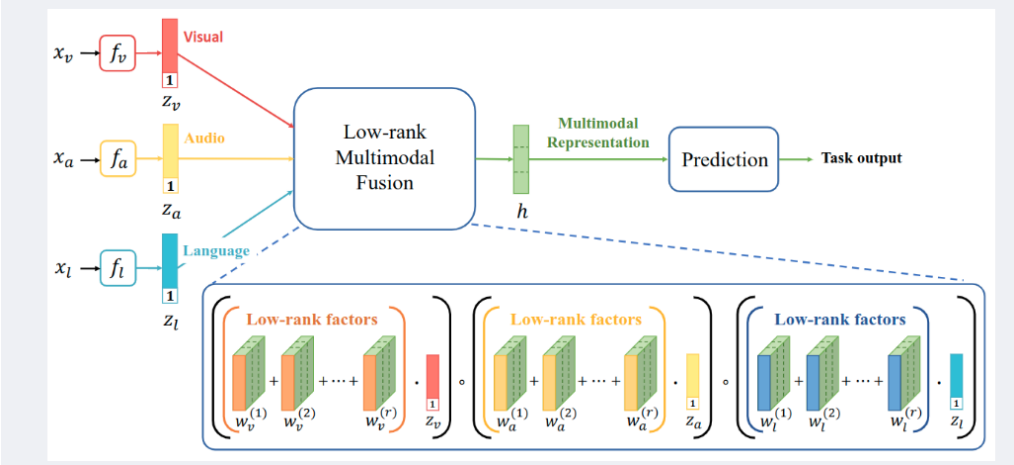

Fig. 13. LMF model.

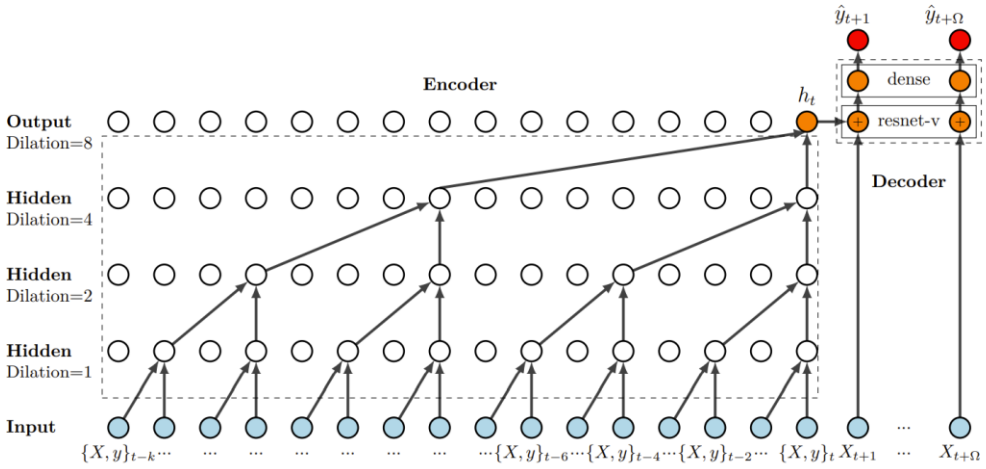

Fig. 14. TFN model.

4) MFN model: The proposed approach to address the multi-view sequence learning problem is the Memory Fusion Network (MFN) [21]. This neural architecture explicitly accounts for intra-view and inter-view interactions, modeling them sequentially across the temporal dimension. The MFN comprises an LSTM system designed to learn view-specific interactions in isolation while identifying cross-view interactions via a mechanism known as Delta-memory Attention Network (DMAN). This process culminates in a multi-view gated memory for temporal summarisation, as illustrated in Fig. 15.

5) EF-LSTM model: The EF-LSTM model does this by connecting multimodal input data and processing it using an LSTM network (Hochreiter and Schmidhuber 1997). This design allows the model to process information from different modalities simultaneously, thus improving the performance of sentiment analysis.

6) MulT model: The MulT model effectively tackles the challenges of aligning information across different modalities and managing long-term dependencies within the same modality in an end-to-end framework, eliminating the necessity for explicit data alignment [22]. The model's foundation is its directed two-by-two cross-modal attention mechanism, which facilitates the examination of interactions across various time steps in a multimodal sequence, potentially allowing for alignment between modalities, as illustrated in Fig. 16.

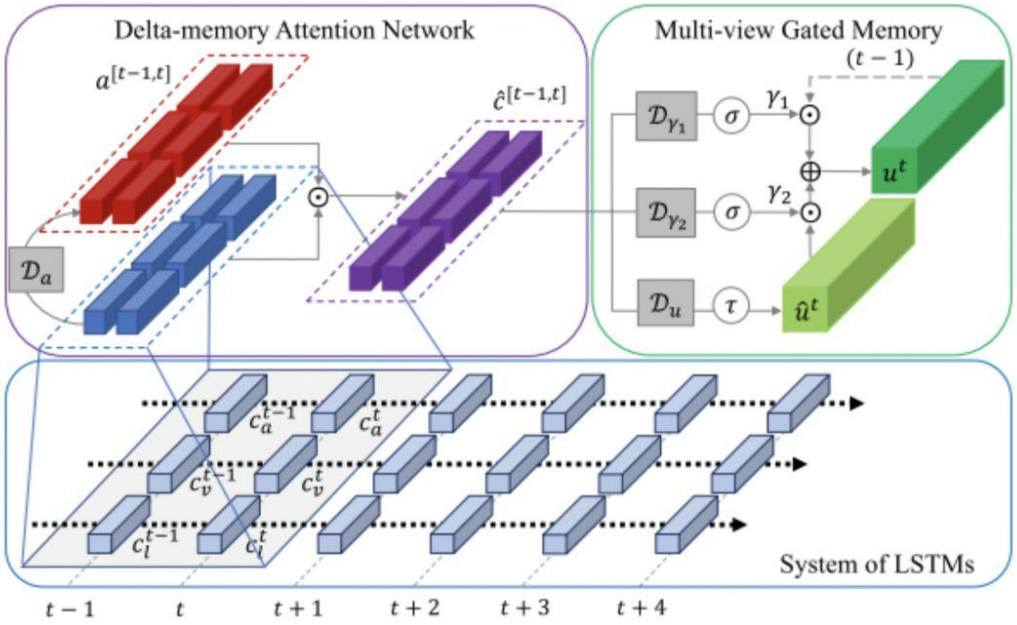

Fig. 15. Schematic diagram of the MFN model.

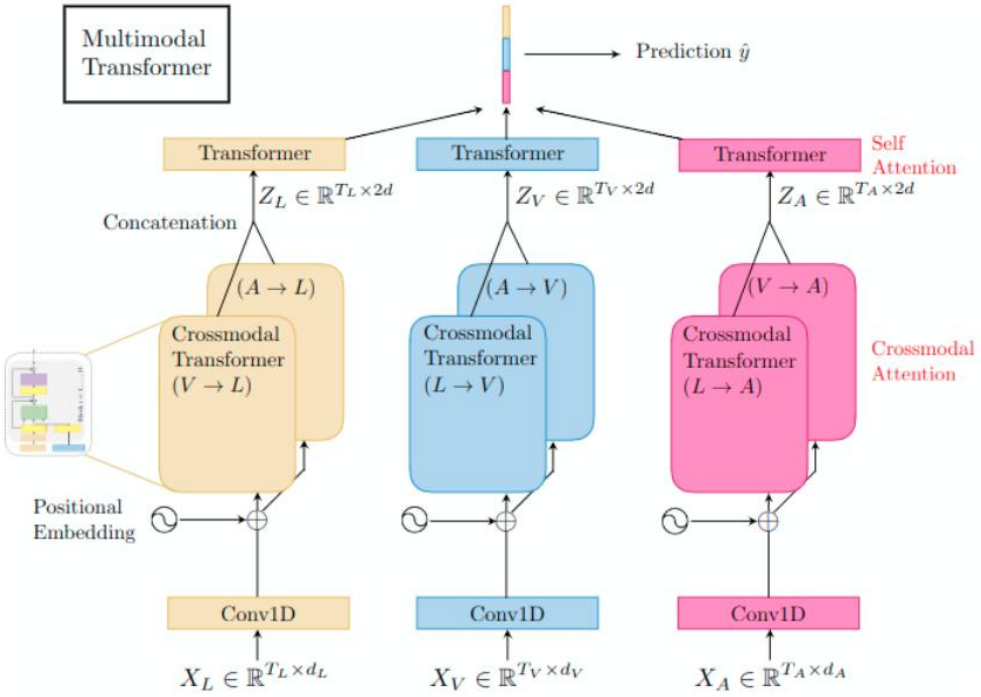

Fig. 16. Schematic diagram of the MulT model.

The dataset used in this paper is the CMU-MOSEI multimodal dataset, which is the first, largest and highly regarded multimodal emotion and sentiment recognition dataset.

B. Experimental Setup

In this paper, three experiments are set up, namely the comparison experiment, the principal mode selection experiment and the ablation experiment. Source of experimental data were obtained from the text of the Communist Youth League video comments.

1) Comparison test: This experiment is intended to prove the applicability of the method used in the M-S model and the accuracy of the model, this paper uses the CMU-MOSEI dataset to train on each of the six baseline models 10 times respectively (the MulT model is the most effective of the baseline models,

so 50 experiments were carried out), and in order to prove the stability of the M-S model proposed in this paper, the was trained 50 times and the average results obtained.

2) Main modality selection experiment: Text features, image features and audio features are selected to be sent to the main channel for processing, and the features of the remaining modalities are sent to the subchannel for processing, and the results are obtained through the M-S model to determine the main modality to be sent to the main channel for processing.

3) Ablation experiment: Utilizing solely unimodal information (input from only one modality—text, image, or audio—while excluding data from the other modalities), the experimental findings indicate that information from all modalities in multimodal sentiment analysis positively influences the sentiment analysis task.

C. Experimental Analyses

Fig. 17 illustrates that the M-S model introduced in this paper yields superior outcomes across four metrics: sentiment 2 classification, sentiment five classification, sentiment 7 classification, and F1 value. This demonstrates that the proposed primary and secondary dual channels not only attain high accuracy but also exhibit versatile applicability in multimodal sentiment analysis tasks. Furthermore, in comparison to sentiment 2 classification, the model shows greater improvement in multicategory sentiment classifications (sentiment 5 and sentiment 7), suggesting that the proposed method is more adept at precise sentiment analysis. The Classification, Sentiment 7 Classification demonstrates greater improvement, suggesting that the method provided herein is more suitable for precise sentiment analysis.

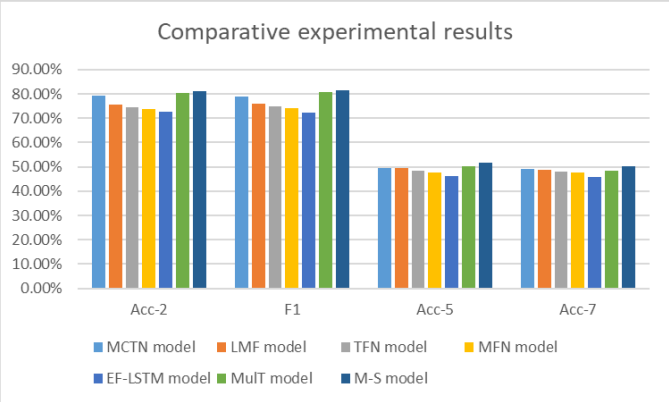

Fig. 17. Comparative experimental results.

Fig. 18 indicates that the results are markedly superior when text features are utilized as the primary modal compared to when image and audio features are employed. This demonstrates that the paper prioritizes text information as the main modality, directing the text features extracted by the one-dimensional convolutional neural network to the primary channel for processing, while the extracted image and audio features are allocated to the subchannel. This approach yields optimal results and substantiates the beneficial impact of the proposed main and subchannel methods on the multimodal sentiment analysis task. The vice-channel strategies presented in this paper positively influence the multimodal sentiment analysis challenge.

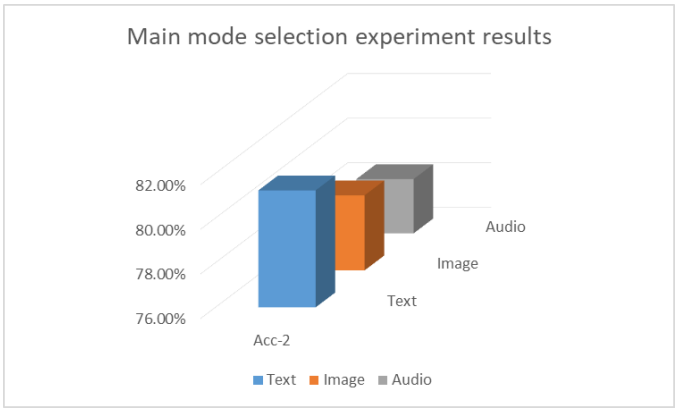

Fig. 18. Main mode selection experimental results.

In Fig. 19, the fusion method proposed in this paper in the CMU-MOSEI dataset when experiments are conducted using multimodal data, the effect is 6.94% more accurate with single text modal sentiment analysis, 21.11% more accurate with single image modal sentiment analysis, and even more accurate with 37.6% more accurate than the single audio modal, which demonstrates that each modal information in this method makes a positive effect on the present sentiment analysis task.

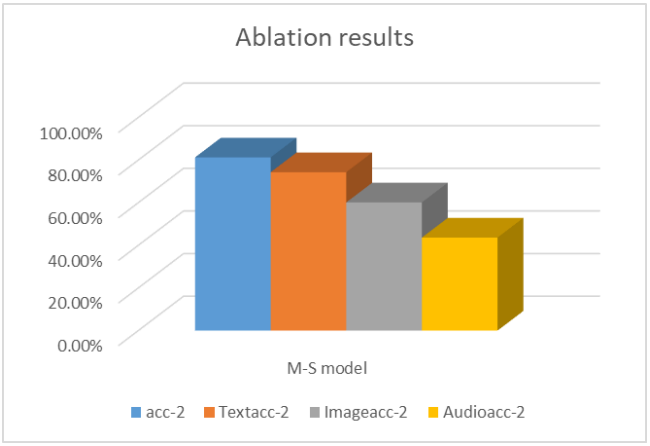

Fig. 19. Ablation results.

VI. Conclusion

The experimental results indicate that the M-S model introduced in this paper enhances the sentiment analysis of multimodal information in video review texts compared to the baseline model. The accuracy of the proposed model has been substantiated through various datasets and multiple experimental sets. Furthermore, the ablation experiments demonstrate that the textual, audio, and visual information utilized in this model contribute positively to the multimodal sentiment analysis task.

The M-S model proposed in this paper mainly solves the problems of the multimodal sentiment analysis task in which the time complexity of the model is too large for practical application, the interrelationships and mutual influences between multimodal information are not considered, and the weights of the multimodal features cannot be accurately defined by numerical values, but there are still a lot of challenges that need to be solved by researchers in this task. This paper

summarises some of the future work on the multimodal sentiment analysis task:

- Multimodal datasets are scarce, unsupervised or semi-supervised methods can be considered to train the model and get better results with fewer datasets;
- Use the topology of the hypergraph to establish the relationship between the multimodal data and obtain the feature tensor between the multimodalities;
- Majority of existing methodologies predominantly focus on textual, visual, and auditory data, while pose-oriented and ECG-based sentiment analysis remains exceedingly limited; future efforts should enhance collaboration with other disciplines to develop more comprehensive multimodal datasets;
- Most existing methodologies do sentiment analysis on aggregate data; future approaches may explore sentiment analysis of specific entities within the data or at the aspect level.

ACKNOWLEDGMENT

This work was supported by High-level Talent Project of Huzhou Vocational & Technical College (2024ZS03) and Jinghu Talent Training Project of Huzhou Vocational & Technical College.

REFERENCES

[1] Sayrol E .Development of a platform offering video copyright protection and security against illegal distribution[C]//2005:76-83.DOI:10.1117/12.591742.

[2] Zongyue W , Sujuan Q .A sentiment analysis method of Chinese specialised field short commentary[C]//IEEE International Conference on Computer & ; Communications.IEEE, 2017:2528-2531.DOI:10.1109/CompComm.2017.8322991.

[3] Simm W , Ferrario M A , Piao S S ,et al. Classification of Short Text Comments by Sentiment and Actionability for VoiceYourView[J].IEEE, 2010.DOI. 10.1109/SocialCom.2010.87.

[4] Simm W , Ferrario M A , Piao S S ,et al. Classification of Short Text Comments by Sentiment and Actionability for VoiceYourView[J].IEEE, 2010.DOI. 10.1109/SocialCom.2010.87.

[5] Yu S .A Bullet Screen Sentiment Analysis Method That Integrates the Sentiment Lexicon with RoBERTa-CNN[J].Electronics, 2024, 13.DOI:10.3390/ electronics13203984.

[6] Shi P , Shi M .Emotion analysis method based on domain dictionary and machine learning[J]. 2021.

[7] Hamouda A , Marei M , Rohaim M .Building Machine Learning Based Senti-word Lexicon for Sentiment Analysis[J]. Technology, 2011, 2(4):199-203.DOI:10.4304/jait.2.4.199-203.

[8] Liang J , Chai Y , Yuan H ,et al. Deep learning for Chinese microblog sentiment analysis[J].

[9] Poria S , Majumder N , Hazarika D ,et al. Multimodal Sentiment Analysis[C]//IEEE Educational Activities Department, PUB766, Piscataway, NJ, USA. IEEE Educational Activities Department, PUB766, Piscataway, NJ, USA, 2018.DOI:10.1007/978-3-319-95020-4_7.

[10] Zhao L , Pan Z .Cross-Modal Semantic Fusion Video Emotion Analysis Based on Attention Mechanism[J].2023 8th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), 2023:381-386.DOI:10.1109/ICCCBDA56900.2023.10154781.

[11] Feng X , Ju F , Hou H ,et al. Sentence Level Fine-grained Emotion Computation Based on Dependency Syntax Improvement Dictionary 1[J]. 2022.

[12] Shetty P , Kini S , Fernandes R .A Comprehensive Analysis of 'Machine Learning and Deep Learning' Methods for Sentiment Analysis in Twitter[J].SN Computer Science, 2024, 5(7):1-13.DOI:10.1007/s42979-024-03216-2.

[13] Oumaima B , Amine B , Mostafa B .Deep Learning or Traditional Methods for Sentiment Analysis: a Review[C]//The Proceedings of the International Conference on Smart City Applications.Springer, Cham, 2024.DOI:10.1007/978-3-031-53824-7_3.

[14] Gonda R , Park J .Fine-Grained Sentiment Analysis of Covid-19 Quarantine Hotels through Text Mining[J]. Conference on Industrial Engineering and Operations Management, 2023.DOI:10.46254/an13.20230207.

[15] Yujie W , Yuzhong C , Chen J D .A knowledge-augmented heterogeneous graph convolutional network for aspect-level multimodal sentiment analysis[J] . .Computer speech & language, 2024, 85(Apr.):101587.1-101587.19.DOI:10.1016/j.csl.2023.101587.

[16] Zhang H , Wang Y , Yin G ,et al. Learning Language-guided Adaptive Hyper-modality Representation for Multimodal Sentiment Analysis[J]. 2023.

[17] Cheng H , Yang Z , Zhang X ,et al.Multimodal Sentiment Analysis Based on Attentional Temporal Convolutional Network and Multi-Layer Feature Fusion[J].IEEE transactions on affective computing, 2023(4):14.DOI:10.1109/TAFFC.2023.3265653.

[18] Pham H, Liang P P, Manzini T, et al. Found in translation: learning robust joint representations by cyclic translations between modalities[J]. arXiv:1812.07809(2019).

[19] Liu Z, Shen Y, Lakshminarasimhan V B, et al. Efficient low-rank multimodal fusion with modality-specific factors[J]. arXiv:1806.00064 (2019).

[20] Zadeh A, Chen M, Poria S, et al. Tensor fusion network for multimodal sentiment analysis[J]. arXiv:1707.0725 (2017).

[21] Zadeh A, Liang P P, Mazumder N, et al. Memory fusion network for multi-view sequential learning[J]. arXiv: 1802.00927 (2018).

[22] Tsai Y H, Bai S J, Liang P P, et al. Multimodal transformer for unaligned multimodal language sequences[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. florence: 2019. 6558-6569.