The Research and Preprocessing Results of The End-to-End Multimodal Sentiment Analysis System

Ziyi Yuan *

School of Computing, Zhuhai College of Science and Technology, Zhuhai, China

* Corresponding Author Email: zeapyphh@krae.edu.kg

Abstract. With the rapid development of artificial intelligence technology, the application of sentiment analysis in areas such as social media monitoring, customer service, and intelligent assistants is gradually increasing. Traditional sentiment analysis usually relies on unimodal analysis, which has certain limitations. Therefore, end-to-end multimodal sentiment analysis systems have emerged. This paper combines several typical multimodal sentiment analysis systems, with the Tension-Tension Experimental Engineering (TTEE) model for efficient computation and reasoning, and incorporates weighted vector fusion and cross-modal attention mechanisms to improve the accuracy of sentiment analysis. The system uses Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to extract high-level features from text, speech, and images, enabling deep fusion of multimodal data. Additionally, the system adopts the Multi-Task Multi-View Network (MTMVN) architecture for multi-task and multi-perspective learning, thereby enhancing the robustness and accuracy of the sentiment analysis model. Compared to traditional unimodal analysis methods, multimodal sentiment analysis systems demonstrate higher accuracy and stability across multiple sentiment analysis tasks.

Keywords: Summary of Preprocessing, End-to-End Systems, Practical Applications.

1. Introduction

Sentiment analysis is an important research direction in the field of Natural Language Processing (NLP), aiming to identify and assess the emotional states expressed in human language, speech, facial expressions, and other forms of communication. With the development of social media, e-commerce, customer service, and other areas, the demand for sentiment analysis applications is increasing, requiring efficient and accurate recognition of emotional information. Traditional sentiment analysis techniques typically rely on a single modality for emotion recognition, such as text, speech, or image sentiment analysis. Each modality has its unique way of expressing emotion, but also comes with limitations. Therefore, unimodal sentiment analysis cannot fully capture the diversity of emotions.

End-to-end multimodal sentiment analysis systems address the shortcomings of unimodal analysis by integrating data from multiple modalities such as text, speech, and images. These systems can combine information from different sources, comprehensively extract emotional features, and enhance the accuracy and robustness of the analysis. With the application of deep learning technologies, particularly Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and self-attention mechanisms such as Transformers, end-to-end multimodal sentiment analysis systems have seen significant improvements. By combining information from different modalities, these systems can capture emotional expressions comprehensively and accurately, reducing errors introduced by single modalities. This fusion enhances the system's robustness, as emotional expressions in different modalities may be affected by environmental noise or changes in facial expressions. Through multimodal fusion, the system can fully leverage the advantages of each modality, compensating for the shortcomings of any single modality and improving sentiment recognition accuracy.

Multimodal sentiment analysis has also driven the development of cross-domain applications. As the technology matures, more industries are beginning to benefit from multimodal sentiment analysis. These applications, by achieving more accurate sentiment recognition, can enhance human-computer interaction experiences and better understand user needs. Additionally, this technology has promoted

the advancement of multimodal learning and fusion research, becoming a hot topic in the field of artificial intelligence. Multimodal sentiment analysis technology has not only driven the interdisciplinary development of deep learning, natural language processing, and computer vision but also provided data support for personalized services. Current research primarily focuses on multimodal learning, text sentiment analysis, and multi-task learning, proposing various novel network structures and algorithms to address the limitations of traditional methods and further improve the performance and efficiency of multimodal data processing. Optimization of model fusion and feature extraction is also an important research direction.

This paper mainly introduces five typical systems and the experimental results of preprocessing, and through comparison, identifies their shortcomings. Finally, it provides a brief overview of current research on the application of end-to-end multimodal sentiment analysis systems in daily life. In recent years, end-to-end multimodal sentiment recognition systems have become a research hotspot in the field of sentiment computing, using deep learning technologies (such as CNN, RNN, Transformers, etc.) to integrate text, speech, and visual data, thereby improving sentiment recognition accuracy. Research is mainly focused on modality fusion methods, multidimensional modeling of emotions, and challenges related to datasets and annotations. Moreover, application scenarios cover fields such as intelligent customer service and emotional health monitoring. However, challenges still remain, such as modality information fusion, cross-cultural adaptation, system robustness, and interpretability. Future research will focus on addressing these challenges and promoting the widespread application of multimodal sentiment recognition technology.

2. Introduction to Typical Systems

2.1. CMMT Model

The advantages of the Convolutional Multi-Modal Transformer (CMMT) model in tasks are demonstrated through multiple experimental results, especially in its performance across different Twitter datasets. The CMMT model shows excellent precision, recall, and F1 score on all three Twitter datasets. For example, on the Twitter-2015 dataset, the CMMT model achieves an F1 score of 66.5, significantly outperforming other comparison methods. The CMMT model demonstrates higher precision and F1 scores across multiple benchmark methods such as TextSPAN, D-GCN, and RoBERTa. Specifically, compared to traditional unimodal methods, CMMT improves the overall performance in multimodal tasks by combining visual and textual information, allowing it to better capture the interaction between text and images. Moreover, ablation studies on different components of the CMMT model further validate the effectiveness of its design, with visual information and the TASE module being crucial for enhancing model performance. CMMT, through multimodal fusion (the combination of visual and textual information), effectively improves sentiment classification tasks, especially in scenarios involving image-text interaction, outperforming traditional models that rely solely on text. Its model design better captures cross-modal relational information [1].

Therefore, the CMMT model demonstrates its advantages in multimodal tasks in experimental results, particularly in improving precision, F1 scores, and handling cross-modal information, showing significant improvements compared to traditional methods.

2.2. TTEE Model

The TTEE model establishes a twin-tower structure, which encodes the context and the given aspect-sentiment pairs separately. This structure significantly reduces redundant computations and improves computational efficiency. By doing so, TTEE better captures the dependency between context and aspect-sentiment, making sentiment analysis more accurate and avoiding the issue of mixed context and sentiment labels found in traditional methods. On the other hand, the TTEE model significantly improves computational efficiency during both the training and inference stages, especially on large-scale datasets. For example, TTEE can be up to 100 times faster than benchmark methods, allowing it to train and infer efficiently on large amounts of data without being limited by

sample size or the number of aspect categories. This provides a significant advantage in practical applications, especially when handling sentiment analysis tasks involving large volumes of reviews or large-scale data. Moreover, the TTEE model does not require additional model architectures when dealing with implicit target entities and their related aspect-sentiments. This design simplifies the model structure and avoids the added complexity traditionally required to handle implicit targets. Through effective context decoding, TTEE can more accurately extract the target entities and their associated aspects and sentiments. When compared to generative models, TTEE is significantly more efficient during inference. Generative models are typically autoregressive, meaning the generation process requires multiple steps, whereas TTEE obtains results with a single encoding during the inference phase, greatly improving speed and efficiency. Additionally, the results of generative models are often uncontrollable, and it is difficult to ensure that the generated target entities fully align with the context, whereas TTEE ensures that the extracted targets are perfectly consistent with the context, avoiding the common error issues found in generative models. TTEE can fully leverage the correlations in multi-task learning, especially during the training phase of multiple Aspect-based Sentiment Analysis (ABSA) sub-tasks. Through effective joint training, TTEE can improve the accuracy and robustness of sentiment classification tasks, demonstrating better performance than traditional methods [2].

2.3. Weighted Vector Fusion of Multimodal Features

Weighted vector fusion of multimodal features is a common technique used to effectively combine features from different modalities (such as text, audio, vision, etc.) to enhance the overall performance of a task. In tasks like multimodal sentiment analysis and emotion recognition, weighted vector fusion is typically used to integrate feature information from different modalities. Each modality employs independent networks or methods for feature extraction. For text, text features are obtained through text encoders (such as BERT, LSTM). For audio, audio features are extracted using convolutional neural networks (CNN) or other sequential models (such as LSTM, GRU). For vision, image features are extracted using deep learning models like CNN.

Feature weighting: For the features of each modality, a weight can be assigned, typically reflecting the importance of each modality in the specific task. For example, in sentiment recognition, the text modality may be more important than the visual or audio modalities. The weighting can be dynamically assigned to each modality through a learning module (such as an attention mechanism), or it can be manually set based on task requirements or prior knowledge. Finally, the feature vectors from different modalities are weighted and summed according to their corresponding weights to extract the final features.

$$ffusion = w1 \cdot ftext + w2 \cdot faudio + w3 \cdot fvisual \quad (1)$$

In this, $w1, w2, w3$ represents the weight of the modality, and $f(text, audio, visual)$ represents the feature vector of each modality. After concatenating the features from different modalities, a weighted matrix is applied for the weighting operation. This approach can capture more interactive information between modalities. In multimodal fusion, attention mechanisms are typically used to compute the weights between modalities. By learning the importance of different modalities, the attention mechanism can dynamically adjust the contribution of each modality.

Weighted vector fusion, by performing weighted summation or concatenation of multimodal features and integrating techniques such as attention mechanisms, enables the model to automatically learn the contributions of different modalities. Ultimately, the fused feature vector can be used for more accurate predictions or classifications, demonstrating powerful capabilities, especially in multimodal tasks like sentiment analysis and emotion recognition [3].

2.4. Multitask Multiview Neural Network Architecture

The MTMVN (Multitask Multiview Neural Network) architecture improves the performance of aspect-based sentiment analysis (ABSA) by combining multitask learning and multiview learning.

First, MTMVN uses the multitask learning strategy by simultaneously training two sub-tasks (AE task: Aspect Extraction, ASP task: Aspect Sentiment Polarity) to enhance overall performance. The AE task is responsible for extracting aspect information, while the ASP task is responsible for determining sentiment polarity. By sharing feature representations, the model can learn more useful information from both tasks, thereby enhancing its understanding of sentiment and aspect recognition capabilities.

Secondly, MTMVN designs three network branches, each handling a different task (AE task, ASP task, and complete ABSA task). Each branch can learn from different "views," and this multiview approach allows the model to extract useful features from various information sources, thus improving the understanding of complex sentiments and aspects.

MTMVN is based on Convolutional Neural Networks (CNN) and Gated Recurrent Units (GRU), which allows the model to capture local features in text (handled by CNN) and long-range contextual information (handled by GRU). This combination strengthens the model's performance in sentiment analysis, especially when dealing with long texts and complex sentiments.

MTMVN incorporates an effective information interaction mechanism that allows different tasks to share useful information. The relationship between the AE and ASP tasks is fully utilized through shared representations and network structures. Through the carefully designed interaction mechanism, the two sub-tasks can mutually promote each other, avoiding the feature redundancy or information conflict that may arise from independent training.

Finally, traditional unified label methods may lead to confusion in the learned representations, whereas MTMVN adopts a strategy of separately representing the AE and ASP tasks, allowing each sub-task to independently learn its feature space. This design avoids the label confusion problem and ensures efficient training and accurate prediction for each sub-task [4].

2.5. FV2ES System

The Fully End-to-End Multimodal System system (FV2ES) has the main advantage of efficient multimodal sentiment recognition, especially in video sentiment analysis tasks. FV2ES adopts a fully end-to-end architecture, eliminating the need for complex manual feature extraction or preprocessing steps. All inputs, such as video, audio, and text, are directly fed into the system and processed through a unified multimodal network. This design makes the system simpler, reducing human intervention in traditional methods while effectively capturing complex relationships between modalities.

The FV2ES system can simultaneously handle video, audio, and text information, and integrates different modality information through an advanced fusion mechanism. This multimodal information fusion helps improve sentiment recognition accuracy, as sentiment is not solely dependent on a single modality (such as voice or facial expressions), but rather is a combined result of various information sources.

FV2ES employs a model based on self-attention mechanisms and Convolutional Neural Networks (CNNs), which can effectively capture spatiotemporal dependencies in video and audio. The self-attention mechanism helps the model focus on the parts of the input data that are relevant to sentiment, thus improving the accuracy of sentiment analysis.

By optimizing multimodal inputs, FV2ES performs excellently on various public datasets, especially on video sentiment datasets such as IEMOCAP, where its accuracy and real-time performance have significantly improved [5]. Compared to other models, FV2ES achieves higher sentiment recognition efficiency while maintaining relatively low computational complexity.

Data optimization and augmentation: In processing video data, a clean dataset is crucial for model training. FV2ES optimizes the data preprocessing steps, removing noise and background interference from the video, ensuring that the model focuses on core sentiment information. Additionally, to handle complex backgrounds and noise, model designers can use data augmentation methods, such as adding background noise or simulating different environmental changes, to improve the model’s generalization ability.

FV2ES also adopts a multitask learning strategy, combining sentiment recognition with other related tasks, such as facial expression recognition and emotion classification. Through multitask learning, the model can share information and learn more sentiment-related features from different tasks, thereby improving overall performance in sentiment recognition.

Since video sentiment recognition tasks often require significant computational resources, FV2ES can optimize processing speed using parallel computing and streaming processing technologies, reducing video analysis delays. This is particularly important for sentiment analysis in short videos on social networks, as it enables timely feedback on user emotions, enhancing interactivity and user experience [6].

3. The Result of Preprocessing

Text Processing (Implemented in C++), The key part is using the cppjieba library for Chinese word segmentation, removing stop words to improve efficiency. TF-IDF is used to generate the word weight vector. Then, sentiment scores are calculated using a sentiment lexicon, followed by pre-extracted BERT features in Python. These features are directly loaded in C++ to avoid redundant calculations. These three methods are used for feature extraction. Afterward, TF-IDF, sentiment scores, and BERT features are fused, normalized, and input into the classifier for feature fusion. The stop word list can be dynamically adjusted to improve the accuracy of sentiment scores. Multi-threading supports real-time or batch mode switching.

Audio Processing (Implemented in C++), The main preprocessing tasks are noise reduction and frame-based MFCC to extract spectral envelope features, followed by Chroma (12 dimensions) for pitch class features. Finally, the Zero Crossing Rate (ZCR) is used as a frequency indicator. These three methods are used for feature extraction, but OpenSMILE is required for command-line or C++ interface feature extraction, and Librosa is used to assist in verifying and supplementing features.

Image Processing (Implemented in C++), OpenCV is used to extract keyframes from the video, and the Haar Cascade Classifier is used to locate faces. Facial expression features are extracted using the 68 facial landmarks from the Dlib library, calculating distances, angles, and other attributes of the landmarks. As shown in Table 1, the summary of the preprocessing steps is as follows:

Table 1. Preprocessing Summary

modality	key features	tools/libraries	output
Text	TF-IDF、sentiment scores、BERT	cppjieba、Python-BERT	normalized text feature vector
Audio	MFCC、Chroma、ZCR	OpenSMILE、Librosa	37-dimensional audio feature vector
Image	68 facial landmarks、geometric features	OpenCV、Dlib	emotion feature vector

4. Analysis of the Limitations of the End-to-end System

Although end-to-end systems can effectively integrate information from different modalities, there are still challenges in handling the complex interactions and mutual influences between modalities. Meanwhile, assigning appropriate weights during the multimodal fusion process is a difficult task. Traditional weighted fusion methods may not fully capture the importance of each modality under different circumstances, leading to imprecise information fusion. Regarding computational efficiency and resource consumption, although the design of end-to-end systems is simple, the computational efficiency of the model may be limited in large-scale datasets and real-time sentiment analysis frameworks, especially when hardware resources are limited. The TTEE model improves computational efficiency, but it may still encounter bottlenecks in more complex tasks. For tasks involving target entities or complex emotional analysis, end-to-end systems often rely on general models, which may struggle to capture all potential information. The CMMT model's performance

degrades when the quality of visual information is poor, indicating that it is sensitive to input data of varying quality. In a multi-task learning framework, while information sharing between tasks can improve efficiency, it may also lead to interference between tasks. The MTMVN model's multi-task learning may cause information conflicts during training due to overly tight associations between tasks, increasing the model's complexity. Although end-to-end systems improve task accuracy through joint training, different task requirements may vary, which may require the system to be adjusted when handling unseen data or tasks. Additionally, the scalability of end-to-end models may be limited. The TTEE model is more efficient than generative models during the inference phase, but the results of generative models are often uncontrollable, and the generated targets may not align with the context, which remains an issue in the application of end-to-end systems [7].

5. Applied Research

5.1. Intelligent Customer Service System

With the development of artificial intelligence and natural language processing technologies, intelligent customer service has become an important tool for interaction between businesses and customers. However, traditional text-based customer service systems often fail to accurately capture the emotional state of customers, leading to delayed responses to customer dissatisfaction. End-to-end multimodal sentiment analysis systems can accurately identify customers' emotional states (such as anger, anxiety, satisfaction, etc.) by analyzing their voice, text, and facial expressions, thereby providing emotional feedback to the customer service system. If the system detects significant emotional fluctuations, it can automatically trigger emotional intervention strategies to optimize the customer experience. At the same time, the technology enhances the sentiment recognition capabilities of the intelligent customer service system through multimodal sentiment analysis. Specifically, the system uses deep learning techniques, combining semantic understanding and context analysis, and employs Bidirectional Long Short-Term Memory Network (BiLSTM) to encode sentences and capture long-term dependencies in the text. In addition, the system integrates a dynamic semantic feature extraction module and Gated Recurrent Unit (GRU) to further improve the accuracy of sentiment analysis. These technologies ensure that, based on multimodal information, the system can accurately understand the user's emotions and respond accordingly [8].

5.2. Autonomous Driving and Intelligent Transportation

The development of autonomous driving technology requires vehicles to make fast and accurate decisions in complex environments. In some cases, the emotional state of the driver directly affects their reaction speed and judgment. Therefore, understanding the emotional changes of the driver, especially in emergency situations, is crucial. By collecting the driver's facial expressions, voice, and physiological signals (such as heart rate), combined with data from the driving environment, an end-to-end multimodal sentiment analysis system can assess the driver's emotional state, which in turn influences the vehicle's decision-making. If significant emotional fluctuations are detected in the driver, intervention measures can be taken. Additionally, by observing the driver's facial expressions, the system can detect signs of drowsy driving. If detected, the system can remind the driver to rest or switch drivers through voice prompts. Finally, if excessive emotional fluctuations are detected, an autonomous driving strategy can be employed to ensure safety. The next five years are divided into two phases: The first phase, from 2018 to 2019, focused more on the vehicle design itself, with attention given to vehicle design and configuration; in addition, there was also interest in the overall development of the industry. The second phase, from 2020 to 2022, saw the public's attention shift not only to design and industry development but also to the application of technology and technical testing, which indirectly indicates that autonomous driving technology in China is flourishing. With the continuous development of autonomous driving technology, we can foresee that urban traffic will undergo significant changes in the future [9].

5.3. Education and Online Learning Platforms

Online learning platforms are gaining increasing attention, but students' emotional changes (such as anxiety, interest, fatigue, etc.) have a significant impact on learning outcomes. Effectively recognizing students' emotional states and providing personalized learning support is an important challenge faced by current educational technologies. By analyzing changes in students' facial expressions, video, and voice, for example, if a student appears mentally fatigued, appropriate rewards can be given to help motivate the student. Learning strategies can also be automatically adjusted based on the student's emotional state to facilitate efficient learning.

The end-to-end multimodal sentiment analysis system has enormous potential for application across various industries. By combining multiple modalities (such as text, speech, images, etc.) to comprehensively analyze emotional information, it can significantly improve the accuracy of sentiment recognition and the adaptability of the system. In the future, these technologies are expected to be widely applied in more fields, such as smart healthcare, education, and intelligent transportation, not only providing more personalized and accurate services for users but also offering technological support for industry innovation and development [10].

6. Conclusion

The application and potential of the FV2ES system in multimodal sentiment analysis are the focus of this article, which highlights its end-to-end architecture, multimodal fusion, efficient computational models, and optimized performance. FV2ES effectively captures the spatiotemporal dependencies between multimodal data by combining video, audio, and text using self-attention mechanisms and convolutional neural networks. This significantly improves the accuracy and efficiency of sentiment recognition. By optimizing data processing and employing a multi-task learning strategy, FV2ES demonstrates excellent performance on various public datasets, particularly surpassing traditional models in terms of real-time performance and computational efficiency.

Although FV2ES exhibits powerful multimodal fusion capabilities, challenges remain in handling complex interactions between modalities and weight distribution. Other models, such as CMMT, TTEE, and MTMVN, also have their own strengths and weaknesses in multimodal sentiment analysis. CMMT's performance deteriorates with poor visual information quality, TTEE has limitations in expressing complex emotions, and MTMVN's multi-task learning can lead to task interference and increased model complexity.

End-to-end multimodal sentiment analysis systems have broad application prospects, especially in areas like intelligent customer service, autonomous driving, education, and online learning platforms. By combining various input methods such as speech, facial expressions, and text, these systems can more accurately recognize emotional states, providing users with personalized and intelligent services. In the future, with further technological optimization, these systems are expected to play a greater role in multiple industries, driving the development of intelligent services and enhancing user experiences.

References

[1] Yang L, Na J C, Yu J F. Cross-Modal Multitask Transformer for End-to-End Multimodal Aspect-Based Sentiment Analysis. Information Processing & Management. 2022, Volume 59, Issue 5, September 2022, 103038.

[2] Li Z L, Song Y Q, Lu X L, Liu M. Twin Towers End to End model for aspect-based sentiment analysis. Expert Systems with Applications. 2024, Volume 249, Part C, 1 September 2024, 123713.

[3] Tzirakis P, Chen J X, Zafeiriou S, Schuller B. End-to-end multimodal affect recognition in real-world environments. Information Fusion. 2024, Volume 249, Part C, 1 September 2024, 123713

[4] Bie Y, Yang Y. A Multitask Multiview Neural Network for End-to-End Aspect-Based Sentiment Analysis. Big Data Mining and Analytics. 2021, ISSN 2096-0654 05/06 pp195–207, Volume 4, Number 3, September 2021.

[5] Wei Q, Huang X, Zhang Y. FV2ES: A fully end2end multimodal system for fast yet effective video emotion recognition inference. IEEE Transactions on Broadcasting, 2023, 69(1): 10-20.

[6] Yin M, Qiao S, Chen W, et al. Social network rumor detection method based on group emotion stabilization. Journal of Software, 2020, 1-24.

[7] Wang D, He Y, Liang X, et al. Tmfn: A target-oriented multi-grained fusion network for end-to-end aspect-based multimodal sentiment analysis//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024, 16187-16197.

[8] Wei H Z. Technical research on real-time multimodal interaction architecture based on AI big language model for dynamic adjustment of semantic features in intelligent customer service robots. COMBIN. MATH. COMBIN. COMPUT. 2025, 127a (2025) 3867--3886

[9] Chen C, Zhang J. Public Attitudes towards Autonomous Driving: Sentiment and Topic Analysis Based on Weibo Data. The Chinese Library Classification (CLC). 2023.

[10] Tang Q W, Zhang H, Wu Y A. Construction of a Multimodal Learning Analysis Model for Smart Classrooms. Journal of Teacher Education. 2024.