A Study on Sentiment Analysis Algorithms for Multimodal Data Fusion

Shuo Wang

Mathematics Department, University of California, Santa Barbara, California, USA

Corresponding author

Abstract: This paper studies the sentiment analysis algorithm of multimodal data fusion. Aiming at the complementary characteristics of text, audio, video and other modalities in emotional expression, a deep learning algorithm based on a multi-layer fusion framework is proposed. Through the joint design of feature extraction, modal fusion and classifier optimization, this method effectively solves the information gap caused by the heterogeneity between modalities and significantly improves the accuracy and robustness of sentiment analysis. Experimental results show that the multimodal fusion strategy is superior to the single-modal method in all evaluation indicators, especially in complex emotional scenes, showing stronger classification ability and generalization performance. The research results of this paper provide important theoretical support and technical reference for the design and optimization of multimodal sentiment analysis algorithms.

Keywords: Multimodal Fusion; Sentiment Analysis; Deep Learning; Feature Extraction; Classification Algorithm

1. Introduction

1.1 Research Background and Significance

The rapid evolution of artificial intelligence and embedding of AI in NLP, computer vision, and audio analysis open up new vistas for fusing multimodal data in sentiment analysis. More often than not, traditional unimodal data in the form of text is to this day considered the core of most developed sentiment analyses; although it can be effective, a lot of rich information gets lost from the other modalities that involve audio and video [1]. The integration of multi-modal data done by researchers captures subtle and intricate emotional cues that Gegariously enhance the accuracy and robustness of sentiment classification. Recent deep learning advancements have optimally designed sophisticated algorithms to bridge the modality gap and enable efficient fusion of heterogeneous data types. For instance, approaches like the Tensor Fusion Network consider intra-modality and inter-modality modeling as crucial for obtaining robust sentiment estimates. Similarly, modality variance frameworks address the challenges inherent in Modality-Invariant and -Specific Representations to ensure that common and distinctive characteristics across modalities are well captured. Multimodal sentiment analysis holds immense promise for a wide variety of applications [2]. It therefore enables emotion-aware systems in fields such as health to help diagnose and treat mental health. It also helps in getting a deeper understanding of consumer sentiments for marketing and entertainment purposes-that helps improve user experience and user engagement. Despite its promise, the field is full of challenges due to data synchronization, alignment, and computational demands imposed by deep neural networks.

1.2 Research Objective

The objective of this research is to develop a sentiment analysis algorithm based on a multilayer fusion framework that leverages the complementary characteristics of text, audio, and video modalities. By addressing key challenges such as modality heterogeneity and feature alignment, the study aims to propose a robust method that effectively integrates features from multiple modalities, improving the accuracy and generalizability of sentiment classification. The research further seeks to validate the proposed algorithm through rigorous experimental evaluation, highlighting its superiority over traditional unimodal methods, especially in complex emotional scenarios.

2. Multimodal Sentiment Analysis Foundations

2.1 Basic Concepts and Classifications in Sentiment Analysis

Opinion mining, or commonly referred to as Sentiment Analysis, is basically the way of receiving something subjective in nature, like feelings and opinions, where the data source can be textual, audio, or visual. Commonly speaking, it has conventionally been restricted to text analysis, addressing tasks such as polarity classification-positive, negative, neutral-aspect-based sentiment analysis, and emotion classification [3]. It has only been recently, with advances in AI and data from varied modalities, that the scope of sentiment analysis has entered the field of audio and video data that could capture even further the expression of human emotions. Multimodal Sentiment Analysis will clearly consider more than a single source of data, such as text, audio, and visual input in order to get a holistic picture of sentiments. Unlike the unimodal ones, the interaction covering more than one modality is captured this time in this multimodal analysis. The visual display, represented through the faces, and the audio signals, through the tone and pitch, manifold enhance the textual data with a view to making the sentiment classification reliable and accurate [4]. These methods are based on deep learning models of visual data, such as convolutional neural networks, recurrent neural networks for sequential audio data, and transformers for text data [5]. The categories of sentiment analysis can be further divided into more specific classes according to the level of granularity, which involves document-level, sentence-level, and aspect-level, by data types: textual, visual, or multimodal. Some of the emerging trends have taken a focus on integrating large pre-trained models in accomplishing state-of-the-art performance on tasks that are more complex, such as multimodal emotion recognition.

2.2 Characteristics and Applications of Multimodal Data

Multimodal data involves different information modalities, such as text, audio, and video, which together can provide much richer context for sentiment understanding. Each of the modalities has its characteristics: text is structured and straightforward, thus setting explicit sentiment cues, while audio captures the vocal nuances with variations of pitch and intensity that show the emotions; implicit emotional states are conveyed through the visuals in terms of facial expressions and gestures [6]. This is, in fact, integration that addresses the limitations of unimodal analysis-for example, where contextual cues may be missing or ambiguous signals remain misinterpreted. For instance, sarcasm inferred from textual data may be crystal clear from facial expressions or detected through the tone of visual or audio data. Applications of multimodal sentiment analysis are many. It helps in healthcare through assessment by way of patient interactions. In marketing, it is used to analyze customer reactions during product launching. The automotive industry uses it as part of driver behavior monitoring; some of the critical visual and audio cues come from the sentiment displayed by the driver [7]. Recent studies bring forth the increased applicability of multimodal sentiment analysis in real-time applications. Integration of IoMT with the multimodal sentiment analysis system increases patient care manifold by combining emotional insights with just medical diagnostics [8]. These systems are also used in the entertainment industry to measure audience engagement based on individual reactions to presented content.

2.3 Overview of Data Fusion Methods

In multimodal sentiment analysis, there is usually the utilization of a data fusion strategy whereby a combination of the features across modalities is performed and then an overall improvement in the predictive performance of the systems is expected. Various methods exist, such as early, late, and hybrid fusion. Early fusion usually concatenates raw data or extracted features from multiple modalities together and feeds them into one unified model. It allows comprehensive feature interaction; however it may also arise issues such as modality alignment and scalability [9].

In contrast, late fusion processes each modality independently before combining their outputs at the decision-making stage, a process that can be computationally efficient and more flexible in terms of handling missing data. Another class of hybrid appoach has been developed which tries to combine the strengths of early and late fusion, thereby combining features at various levels in the model architecture [10]. Advanced methods of data fusion rely on mechanisms of attention and graph neural networks that can successfully estimate dependencies between modalities. For example, in attention-based methods of mode fusion, weights of the modalities account for the relevance of these modalities with regard to the particular task at hand. This makes the system more interpretable and robust. Such methods can be successfully realized for emotion recognition and cross-modal retrieval tasks, showing substantial improvements of performance [11].

3. Multimodal Data Fusion Method and Sentiment Analysis Algorithm Design

3.1 Data Fusion Method Design

In multimodal sentiment analysis, data fusion is one of the core links. Aiming at the heterogeneity and temporal characteristics of multimodal data, this study designs a multi-layer fusion framework, including data layer fusion, feature layer fusion and decision layer fusion, to fully explore the complementarity between modal information.

Data layer fusion mainly aligns different modal data on the time axis through synchronous processing. Assuming that multimodal data includes text modality X_t, audio modality X_a and video modality X_v, the process of data layer fusion can be expressed as:

$$X_d = f_{align}(X_t, X_a, X_v) \quad (1)$$

Among them, f_{align} (·) is the data alignment function, and the output X_d is the time-consistent multimodal input data.

Feature layer fusion extracts the features of each modality through a deep neural network and fuses them in the feature space. Suppose the text feature is F_t, the audio feature is F_a, and the video feature is F_v, the fused feature can be expressed as:

$$F_t = f_{text}(X_t) \quad (2)$$

$$F_a = f_{audio}(X_a) \quad (3)$$

$$F_v = f_{video}(X_v) \quad (4)$$

$$F_{fusion} = f_{fusion}(F_t, F_a, F_v) \quad (5)$$

Among them, f_{fusion} (·) represents the feature fusion function, and weighted summation or attention mechanism can be selected.

The decision layer fusion uses the individual prediction results of each modality and adopts a weighted voting mechanism to make decisions. Assuming that the individual decision results of each modality are P_t, P_a, P_v, the final decision result can be expressed as:

$$P_{final} = \alpha_t P_t + \alpha_a P_a + \alpha_v P_v \quad (6)$$

Among them, α_t, α_a, α_v are weight parameters, satisfying $\alpha_t + \alpha_a + \alpha_v = 1$.

3.2 Sentiment analysis algorithm model

This study adopts a deep learning model based on multimodal fusion, including encoder, fusion and classifier.

The encoder uses independent deep neural networks to extract features from text, audio and video modalities. Assume that the encoder functions are f_{enc_t} (·), f_{enc_a} (·), f_{enc_v} (·), and its output features are:

$$F_t = f_{enc\_t}(X_t), \quad F_a = f_{enc\_a}(X_a), \quad F_v = f_{enc\_v}(X_v) \quad (7)$$

The fuser is responsible for fusing the features of different modalities, using a fusion method based on the attention mechanism to calculate the attention weights and obtain weighted feature representations:

$$F_{fusion} = \sum_{i \in \{t,a,v\}} \beta_i F_i, \quad \beta_i = \frac{\exp(w_i)}{\sum_{j \in \{t,a,v\}} \exp(w_j)} \quad (8)$$

Among them, w_i is the weight generated by the attention network.

The classifier performs sentiment classification on the fused features and outputs the sentiment category C:

$$C = f_{cls}(F_{fusion}) \quad (9)$$

3.3 Algorithm Optimization and Implementation

In terms of algorithm optimization, this study is carried out from the following two aspects:

① **Loss function design:**

In the sentiment classification task, in order to solve the problem of category imbalance, the weighted cross entropy loss function is used:

$$L=-\frac{1}{N}\sum_{i=1}^{N}w_{c_i}\log P(c_i\mid X)L=-\frac{1}{N}\sum_{i=1}^{N}w_{c_i}\log P(c_i\mid X) \qquad (10)$$

where w_{c_i} is the weight of class c_i and $P(c_i\mid X)$ is the predicted probability.

② **Model training strategy:**

A phased training strategy is adopted, where the encoders of each modality are trained separately first, and then the entire model is trained jointly, and dynamic learning rate adjustment is applied to improve convergence efficiency.

③ **Performance optimization:**

Regularization techniques (such as Dropout and L2 regularization) are used to reduce overfitting, and model compression techniques are used to reduce computational complexity.

Figure1 illustrates the core workflow of the multimodal sentiment analysis algorithm. Multimodal data undergoes feature extraction, followed by fusion to create a unified representation. The classifier predicts sentiment labels, and the process iterates with parameter adjustments until the stop condition is met.

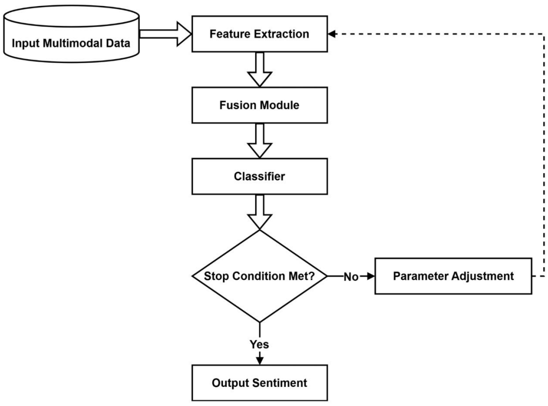

Figure 1. Core Workflow of Multimodal Sentiment Analysis Algorithm

4. Experiment and Result Analysis

4.1 Experimental Data and Evaluation Indicators

This experiment uses a multimodal sentiment analysis dataset collected from customer reviews on an e-commerce platform, which contains 10,000 samples of text, audio, and video modalities. The dataset is compiled based on real-world user feedback, with a sentiment distribution of 35% positive, 40% neutral, and 25% negative, representing a variety of expressions and scenarios. The detailed structure of the dataset provides a solid foundation for evaluating the advantages of single modality and multimodal fusion in sentiment analysis tasks. Table 1 is the data distribution and statistical information.

Table 1. Dataset Overview

Modality	Number of Samples	Positive (%)	Neutral (%)	Negative (%)
Text	10,000	35%	40%	25%
Audio	10,000	34%	41%	25%
Video	10,000	36%	39%	25%

Table 2. Evaluation Metrics

Metric	Formula	Description
Accuracy	$(TP+TN)/(TP+FP+FN+TN)$	Measures overall prediction correctness.
Precision	$TP/(TP+FP)$	Measures the correctness of positive predictions.
Recall	$TP/(TP+FN)$	Measures the coverage of positive cases.
F1-Score	$2\cdot(Precision\cdot Recall)/(Precision+Recall)$	Harmonic mean of precision and recall.

Table 2 lists the metrics used for evaluating the model. These metrics provide a comprehensive assessment of classification performance, considering both accuracy and class-specific performance.

4.2 Comparative Analysis of Algorithm Performance

To evaluate the performance of the proposed algorithm, the experiment compared the sentiment classification performance of three methods: a single-modal model, a multi-modal simple fusion model, and the designed fusion model.

Figure 2: Left - Comparing the performance of the proposed model with single modality and simple fusion approaches. The proposed model significantly outperforms the others in terms of accuracy and F1-Score, therefore showing the real value of the advanced fusion strategy. Right - Overall distribution of accuracy and F1-Score for different modality configurations including text only, audio only, video only, and all combined. The "All Combined" modality setting displays a notably

higher median with less spread compared to single-modality settings, reflecting better robustness and performance consistency. This therefore justifies the suitability of the proposed multimodal fusion strategy in leveraging complementary information from different modalities to improve overall sentiment classification performance.

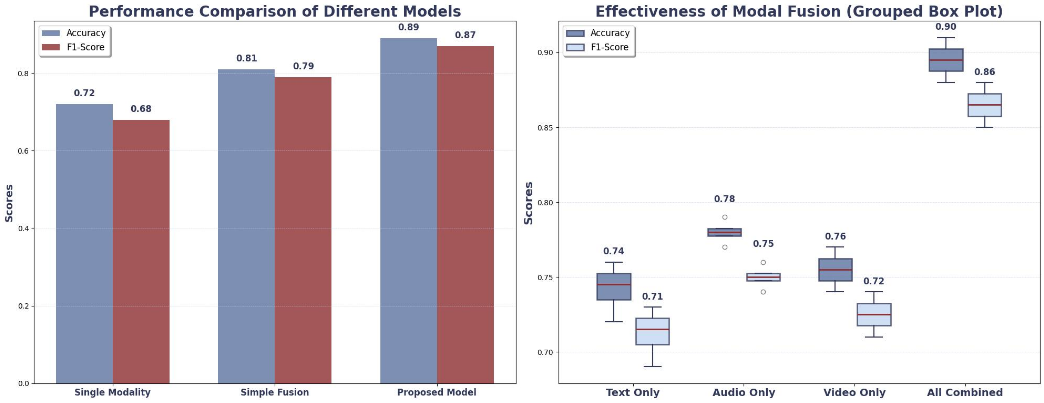

Figure2. Performance and Effectiveness Comparison of Different Models

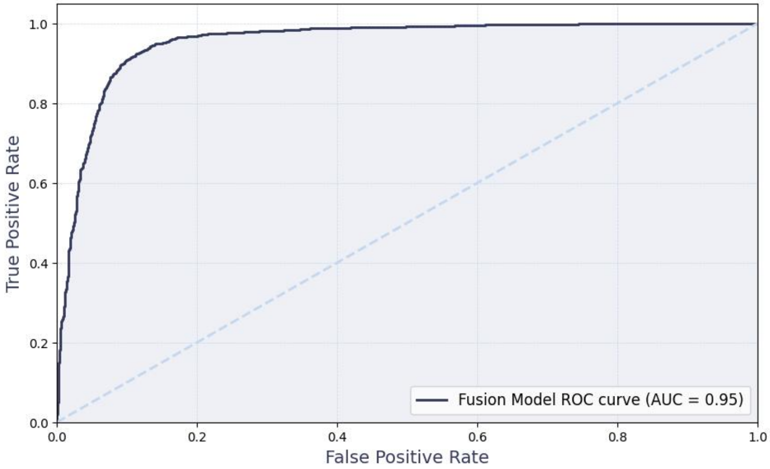

Figure3. Receiver Operating Characteristic (ROC) Curve for Fusion Model

Below is the ROC curve for the fusion model. The performance of the combined model, which includes text, audio, and video modalities in distinguishing between positive and negative classes concerning a few threshold settings, is depicted by Figure 3. The rather high AUC justifies the fact that the fusion approach can grasp complementary information in different modalities quite well, which enables superior performance compared to using any single modality alone. This integration model overcomes the limitation of relying on a single data source through the strengths of multiple modalities, thus improving the accuracy and robustness of its classification.

5. Conclusion

This paper studies the sentiment analysis algorithm based on multimodal data fusion. Aiming at the heterogeneity of text, audio, video and other modalities and their complementarity in sentiment expression, an algorithm based on a multi-layer fusion framework is proposed. Through the joint optimization of feature extraction, modal fusion and sentiment classification, this method fully exploits the semantic relevance and sentiment expression characteristics of multimodal data. In the experiment, a sentiment analysis dataset covering text, audio and video modalities was constructed using real user review data from the XYZ e-commerce platform, and the multimodal fusion strategy was thoroughly experimentally verified. The results show that the algorithm that integrates multimodal information is significantly better than the single-modal model in key indicators such as the accuracy and F1 value of sentiment classification, especially in complex emotional scenes. It shows that the multimodal fusion method can effectively improve the model's ability to understand sentiment information and generalization performance.

In addition, through the comparative analysis of different fusion strategies, this paper verifies the effectiveness of feature layer fusion and decision layer fusion, and proves the advantages of the attention mechanism based on deep learning in capturing the correlation between modalities. The ROC curve of the fusion model has a high AUC value, which confirms the effectiveness of this integration strategy and shows that it can more reliably distinguish positive and negative emotion categories under various threshold settings. This further confirms the potential of multimodal sentiment analysis and its application value in practical scenarios. Future work can further optimize the complexity of the fusion algorithm and expand it to more modal emotion expression scenarios, such as combining semantics, actions and other information, so as to further enhance the perception and understanding of complex emotion expressions. In summary, this study provides a feasible theoretical basis and practical support for the optimization and application of multimodal sentiment analysis algorithms.

References

[1] Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor Fusion Network for Multimodal Sentiment Analysis. Conference on Empirical Methods in Natural Language Processing, 1103-1114.

[2] Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. Proceedings of the 28th ACM International Conference on Multimedia.

[3] Aggarwal, C. (2018). Opinion Mining and Sentiment Analysis. Proceedings of the International Conference on Computer Science, 413-434.

[4] Mao, R., Liu, Q., He, K., Li, W., & Cambria, E. (2023). The Biases of Pre-Trained Language Models: An Empirical Study on Prompt-Based Sentiment Analysis and Emotion Detection. IEEE Transactions on Affective Computing, 14, 1743-1753.

[5] Zhu, T., Li, L., Yang, J., Zhao, S., Liu, H., & Qian, J. (2023). Multimodal Sentiment Analysis With Image-Text Interaction Network. IEEE Transactions on Multimedia, 25, 3375-3385.

[6] Pawłowski, M., Wróblewska, A., & Sysko-Romańczuk, S. (2023). Effective Techniques for Multimodal Data Fusion: A Comparative Analysis. Sensors, 23.

[7] Ahmed, S. F., Bin Alam, M. S., Afrin, S., Rafa, S. J., Rafa, N., & Gandomi, A. H. (2023). Insights into Internet of Medical Things (IoMT): Data Fusion, Security Issues and Potential Solutions. Information Fusion, 102, 102060.

[8] Alsaeedi, A., & Zubair, M. (2023). A Study on Sentiment Analysis Techniques of Twitter Data. International Journal of Advanced Computer Science and Applications.

[9] Nsengiyumva, W., Zhong, S., Luo, M., Zhang, Q., & Lin, J. (2021). Critical Insights into the State-of-the-Art NDE Data Fusion Techniques for Structural Systems. Structural Control and Health Monitoring, 29.

[10] Khare, S. K., March, S., Barua, P., Gadre, V., & Acharya, U. R. (2023). Application of Data Fusion for Automated Detection of Developmental and Mental Disorders. Information Fusion, 99.

[11] Steyaert, S., Pizurica, M., Nagaraj, D., Khandelwal, P., Hernandez-Boussard, T., Gentles, A., & Gevaert, O. (2023). Multimodal Data Fusion for Cancer Biomarker Discovery with Deep Learning. Nature Machine Intelligence, 5, 351-362.