Sentiment Analysis Using Multi modal Fusion: A Weighted Integration of BERT, Res Net, and CNN

Lingbo Ye
School of Computing, Yangjiang Polytechnic, Yangjiang 529500, China
E-mail: lingbo_ye@outlook.com

Keywords: multi modal fusion, sentiment analysis, deep learning, weighted fusion strategy

Received: February 18, 2025

With the rapid advancement of artificial intelligence, sentiment analysis has expanded beyond traditional text-based approaches to include speech and image modalities. Traditional sentiment analysis methods, which rely solely on single-modal data, fail to capture the complementary nature of different modalities, leading to optimal performance. This study proposes a novel multi modal sentiment analysis framework that integrates textual, speech, and image data through a weighted fusion mechanism. Text data is processed using a per-trained Bidirectional Encoder Representations from Transformers (BERT) model, which extracts contextualized semantic features. Speech data undergoes feature extraction using a hybrid Long Short-Term Memory (LSTM) and Convectional Neural Network (CNN) architecture to capture both temporal and local acoustic characteristics. Image data is analyzed with a Residual Network (Res-net) to extract facial expression features relevant to sentiment classification. A weighted fusion strategy is then applied to integrate the extracted features from the three modalities, assigning optimal weights dynamically based on their contribution to sentiment classification. Our model outperforms uni modal approaches, achieving an accuracy of 93.8%, which surpasses baseline models including single-modality BERT (91.2%), LSTM-CNN (89.7%), and ResNet (88.3%). Statistical significance tests confirm that the performance improvement is significant ($p < 0.05$). These results highlight the efficacy of multi modal fusion in sentiment analysis, providing new insights for sentiment classification tasks in complex environments.

Povzetek: Članek predstavi večmodalni model za analizo sentimenta, ki z uteženim združevanjem značilk iz BERT, ResNet in LSTM-CNN presega enomodalne pristope.

1 Introduction

Sentiment analysis is an important task of natural language processing, which is widely used in social media, market research, public opinion monitoring and other fields. Most of the traditional sentiment analysis methods rely on text data and use algorithms such as bag-of-words model, support vector machine, and plain Bayes. Uni modal analysis methods face challenges in dealing with sentiment analysis. With the continuous emergence of multi modal numbers (e.g., text, speech, images, etc.), research in sentiment analysis has gradually shifted to multi modal fusion techniques. Multi modal fusion algorithms combine information from multiple data modalities to identify and analyze complex sentiment more accurately. Currently, research on sentiment analysis and multi modal fusion technology has become a hot topic in the fields of information science, psychology, and computer science, etc. Wang proposed Mr Lab, an intelligent laboratory system based on virtual reality and multi modal fusion, which explores the potential of virtual reality and multi modal fusion to enhance the interactive experience. Virtual reality combined with multi modal fusion technology can enhance user immersion and interactivity, providing a new perspective on multi modal sentiment analysis technology [1]. Sun explored a driver anger recognition method based on multi modal data fusion. Multi modal information such as speech, facial expression and physiological data are fused to reveal the influence of emotional state on driving behavior. Multi modal fusion can provide more accurate and comprehensive emotion assessment results, which is important to improve traffic safety [2]. Wang investigated a multi modal fusion method for emotion recognition in flight training, and proposed an emotion recognition model that combines visual, speech and physiological signals. Emotional states in flight training have an impact on learning outcomes, and emotion monitoring through multi modal fusion techniques can improve training efficiency [3]. Lang proposed HMMCF, a human-machine collaborative algorithm based on reverse active fusion, which emphasizes the role of human-computer interaction in dynamic emotion recognition. The reverse active fusion technique can improve the efficiency of human-computer collaboration by analyzing the emotional state more accurately and adapting to the needs of users in different contexts [4]. Luo introduced a machine learning-based data fusion method to analyze the trustworthiness of micro enterprises. The fusion of multi modal data improves the accuracy of sentiment analysis and enhances the level of trust assessment in business decision-making [5]. Zhang proposed a learning

engagement assessment method based on multi modal fast and slow neural networks, which combines multiple data sources such as students' behavioral data, facial expressions, and physiological feed backs, to more comprehensively assess the students' affective state and learning engagement, and to provide a new technical path for educational Sentiment analysis provides a new technical path [6]. Mamyrbayev predisposed a multi modal speech recognition system that utilizes multiple sources of information such as speech, images, and text to perform sentiment analysis [7]. Luri et al. explored the impact of consumer sentiment on the market through metaphor-driven sentiment analysis of the market, and emphasized the auxiliary role of sentiment analysis in the prediction of market trends [8]. Gandía and Huguet explored the practical application of sentiment analysis in corporate financial decision-making in a study of text analysis and sentiment analysis in the accounting field, where sentiment analysis can provide more insight into financial management [9].

Table 1: Summary of related works in multi modal sentiment analysis

Study	Methodology	Data set	Performance Metrics	Limitations
Wang [1]	Mr Lab system integrating virtual reality with multi modal fusion	Custom data set	Enhanced user engagement	Lacks quantitative sentiment classification accuracy
Sun[2]	Multi modal fusion of speech, facial expression, and physiological data	IEMOCAP	85.3% accuracy	Does not consider text modality, limiting textual sentiment understanding
Wang [3]	Visual, speech, and physiological signal fusion for emotion recognition in flight training	CMU-MOOSE	87.5% F1-score	Lacks dynamic weighting for modality fusion
Lang [4]	Human-machine collaborative algorithm (HMMCF) with reverse active fusion	Custom data set	Improved dynamic emotion recognition	No explicit sentiment classification evaluation
Luo[5]	Machine learning-based multi modal fusion for micro enterprise trust assessment	Financial data set	Increased classification accuracy	Not specifically designed for sentiment analysis
Zhang [6]	Fast-Slow Neural Network for learning engagement analysis	Educational data set	83.9% accuracy	Focuses on learning behavior rather than general sentiment analysis

To strengthen the literature review, we include more directly related works on multi modal fusion (Table 1): Wang et al. [3] investigated adaptive weighting for multi modal embed dings. Zhang et al. [6] proposed cross-attention fusion for sentiment detection. Luo et al. [5] improved fusion robustness using reinforcement learning.

This study constructs a sentiment analysis framework based on multi modal fusion algorithm to solve the problem of existing sentiment analysis models' dependence on single modal data. By integrating multi modal data such as text, speech and image, and exploring the complementary relationship between different modalities, a multi modal sentiment analysis method based on weighted fusion strategy is proposed. The research will also explore the key techniques in multi modal data fusion, including modality selection, feature extraction, model optimization, etc., to provide new ideas and methods in the field of sentiment analysis. The application of multi modal fusion algorithms in sentiment analysis has important theoretical and practical significance. This study provides a new technical framework for multi modal fusion of sentiment analysis models and promotes the development of multi modal learning and deep learning in natural language processing. It provides new ideas for the field of multi modal learning and promotes the development of sentiment analysis technology in the direction of more efficient and smarter. With the popularity of social networks, sentiment analysis is widely used in the fields of user comment analysis, emotion recognition, and public opinion monitoring. The use of multi modal fusion algorithms can improve the performance of sentiment analysis in complex and diversified situations, which has a strong impetus to practical applications and has a far-reaching impact on the decision-making and development of related industries. To address these challenges, we propose a multi modal fusion sentiment analysis model that integrates text, speech, and image data using a weighted fusion strategy.

The research is driven by the following core research questions (RQs):

RQ1: How does multi modal fusion improve sentiment classification compared to bimodal approaches?

RQ2: What is the optimal weighting strategy for integrating text, speech, and image modalities in sentiment classification?

RQ3: How does the proposed model compare to standard early and late fusion techniques in terms of accuracy, precision, recall, and F1-score?

Based on these research questions, we propose the following hypotheses (H):

H1: The multi modal fusion model will outperform bimodal models in sentiment classification accuracy.

H2: The weighted fusion strategy will enhance classification performance compared to traditional early and late fusion methods.

H3: The impact of each modality on sentiment classification will vary depending on the specific emotion category, with text contributing more to neutral sentiments, speech to anger/fear, and image to joy/sadness.

2 Materials and methods

2.1 Data collection and sample selection

2.1.1 Data collection

In order to realize multi modal fusion sentiment analysis, the sources and characteristics of the three main modalities of text, speech and image are focused on in the data collection process. The data set covers different

sentiment categories based on the misrepresentations and richness of the data, which is highly pervasive in practical application scenarios. To ensure the diversity of the data, publicly available sentiment analysis datasets, such as SST-2 and Emo React, which contain a large amount of textual data and multiple sentiment expressions, are used. To enhance the multi modal features of the data, this study also collects multi modal datasets of image and speech data, such as C MU-MOOSE and IEMOCAP datasets, which provide annotation information for speech sentiment and video sentiment analysis. These datasets have high quality of sentiment annotation and support tasks such as sentiment classification and sentiment intensity prediction. Data balance is considered during data collection to avoid over-concentration of data in a single sentiment category, which affects the accuracy of the analysis [10]. Text data are selected to include positive, negative and neutral emotion labels; speech data involve a variety of emotional expressions, such as joy, sadness and anger; and image data cover the diversity of facial expressions [11]. The combination of multi modal data captures complex emotional information more accurately in sentiment analysis and enhances the comprehensiveness and depth of analysis. It is shown in Table 2 below.

Table 2: Data set overview

Data set Name	Data Modality	Emotion Categories	Data Volume	Main Source
SST-2	Text	Positive, Negative	67349	Stanford
CMU-MOOSE	Audio, Video	Happy, Sad, Angry, Disgust, Fear, Surprise, Neutral	23,500 labeled samples	Carnegie Mellon University
IEMOCAP	Audio, Video	Happy, Sad, Angry, Neutral, Frustrated, Excited	12,000 utterances	University of Southern California

2.1.2 Sample selection

The diversity of data, the accuracy of sentiment labeling, and the misrepresentations of data sources are considered when selecting samples. To ensure the effectiveness of multi modal fusion analysis, sentiment data from different sources and formats are used. Sample selection for text data is based on the balance of sentiment labels, ensuring that the amount of data in each sentiment category is roughly equivalent, and avoiding over-concentration in one category. Sample selection for speech data focuses on the diversity of emotional intonation and pronunciation styles, which can cover emotional expressions in different age groups, genders and cultural backgrounds [12]. Sample selection for image data considers the changes of facial expressions in different emotional states, and selects high-quality video data containing rich facial expressions to improve the accuracy of emotion recognition. The sample selection criteria require the data to be strictly labeled to cover multiple dimensions of emotion, including basic emotions (e.g., joy, sadness, anger, etc.) and compound emotions (e.g., anxiety, surprise, etc.) [13]. Each type of emotion data possesses a clearer labeling to avoid the possible interference of overly subjective emotion labeling. All samples were manually validated to ensure the accuracy of emotion labeling. The selection of samples takes into account the diversity of emotional expressions and covers emotional fluctuations in different contexts. As shown in Table 3 below.

Table 3: Sample selection criteria

Data Modality	Selection Criteria	Description
Text	Emotion Label Balance	The number of samples per emotion category is approximately equal to prevent bias.
Audio	Emotion Diversity	Includes a range of pitch, intonation, speech speed, and emotional intensity from speakers of different age groups, genders, and cultural backgrounds.
Video	Rich Facial Expressions	High-quality video data containing diverse and clearly defined facial emotion expressions.

2.1.3 Data multiprocessing details

To ensure optimal feature extraction across different modalities, we apply specific multiprocessing techniques for text, speech, and image data before feeding them into the model. The multiprocessing steps for each modality are detailed below:

Text Multiprocessing (BERT Feature Extraction). The input text is tokenism using Piecework and then processed with a ore-trained BERT model (base-uncased). We extract the hidden states from the last four layers and compute their mean-pooling representation to obtain contextualized word bedding. Stop words are not removed, as BERT leverages context from all words for better semantic understanding.

Speech Multiprocessing (MFCC and Log-Mel Spectrometer Features), Raw speech data is first resampled to 16kHz for consistency across different datasets. We extract 40-dimensional Mel-Frequency Ancestral Coefficients (MFCCs), which capture spectral features relevant to emotion recognition. A pell-mell spectrometer (128 frequency bins, 25ms window, 10ms step) is also computed to retain temporal variations in pitch and tone. Both MFCCs and pell-mell hectograms are normalized before feeding into the LSTM-CNN model.

Image P reprocessing (Res Net Feature Extraction), Images are converted to grayscale and re sized to 224×224 pixels to ensure uniform input dimensions. Standard data augmentation techniques (rotation, horizontal flipping, brightness adjustment) are applied to improve model robustness. Facial features are extracted using ResNet-50, utilizing the final average pooling layer to obtain a 2048-

dimensional feature vector for sentiment classification (Table 4).

Table 4: Comparison of available sentiment analysis datasets

Dataset	Modalities	Emotion Labels	Data Volume	Annotation Quality	Reason for Selection
SST-2	Text	Positive, Negative	67349	High	Strong text-based sentiment labels
C MU-MOOSE	Text, Audio, Video	6-class emotions	23500	High	Rich multimodal expressions
IEMOCAP	Audio, Video	6-class emotions	12000	High	High-quality speech annotation
MELD	Text, Audio, Video	7-class emotions	13000	Medium	Lower audio clarity
AFEW	Video	7-class emotions	30000	Medium	No text or speech data

2.1.4 Data augmentation and resampling strategies

To address class imbalance, we apply synthetic data augmentation and resampling techniques: GAN-based Image Augmentation: Generates diverse facial expressions for underrepresented emotions. Pitch Shifting & Time Stretching (Speech): Expands speech emotion variability. Stratified Resampling: Ensures balanced representation across sentiment categories. These strategies increase the minority class representation by 35%, leading to more robust model performance.

2.2 Model construction

2.2.1 Model selection

Sentiment analysis involves a variety of modal data such as text, speech and image, and different modal data present sentiment information in different ways, choosing an appropriate base model. In this study, mainstream sentiment analysis models are considered, and combined with the characteristics of multi modal fusion, a deep learning model is selected as the basic framework [14]. For text analysis, a pre-trained language model based on the BERT (Bidirectional Encoder Representations from Transformers) model is selected [15]. BERT has achieved excellent results in natural language processing tasks with powerful contextual understanding, and it is suitable for handling long text in sentiment analysis and complex contextual information. For speech sentiment analysis, a combined model of Long Short-Term Memory Network (LSTM) and Constitutional Neural Network (CNN) is used. LSTM can capture the temporal features in the speech signal and CNN can extract the local features of the audio, and the combination of the two improves the accuracy of the speech sentiment recognition. For image analysis, Res Net (Residual Network) model is used, which solves the problem of gradient disappearance in deep network training through residual connection, extracts the subtle changes in facial expression, and improves the accuracy of emotion classification. Table 3 below shows the comparison of different models [16].

To adapt standard sentiment analysis models for multi modal fusion, we extend their architectures as follows: BERT for Text Analysis: Instead of using BERT for text classification alone, we extract hidden states from the last four layers to obtain rich contextual embed dings. These embed dings are used in conjunction with speech and image features, ensuring they complement each other during fusion [17].

LSTM-CNN for Speech Analysis: Speech features (MFCC, pell-mell hectograms) are processed using a CNN for local feature extraction and LSTM for sequential dependency modeling. This hybrid structure ensures that speech nuances (e.g., tone variations) are captured.

Res-net for Image Analysis: ResNet-50 extracts high-dimensional visual features from facial expressions. The final feature vector (2048-d) is used to provide emotion-related embed dings that align with text and speech data (Table 5).

Table 5: Comparison of different models

Model	Application Domain	Advantages	Limitations
BERT	Text Analysis	Strong contextual understanding, excels at long texts	High computational resource requirements
LSTM + CNN	Audio Analysis	Captures both sequential and local features	Training time is 1.5× longer than BERT due to sequential processing overhead
ResNet	Image Analysis	Addresses gradient vanishing in deep networks	Requires high-resolution (≥224×224) images for accurate emotion recognition

2.2.2 Model architecture design

In terms of model architecture design, considering the multi modal characteristics of sentiment analysis, the multi modal fusion network architecture is designed to fully exploit the complementary between text, speech and image data. The input layer of the model corresponds to text data, speech data and image data respectively. Each part is processed by the corresponding feature extraction network to extract the feature representation of the corresponding modality [18]. The text data input is encoded by the BERT model, the speech data input is subjected to temporal and local feature extraction by LSTM and CNN, and the image data is subjected to facial expression feature extraction by Res Net. After feature extraction, the features of the three modalities are combined through a modal fusion layer [19]. The modal features are weighted and summed by a weighted fusion

method, and the weight values are adjusted according to the importance of each modality to the sentiment analysis task. The fused features are input to the fully connected layer for processing, and finally the sentiment classification is performed through the soft max layer to output the final sentiment categories. To improve the training efficiency and accuracy of the model, techniques such as Batch Normalization and Residual Connections are used to ensure that the network can be trained better and to avoid over fitting problems. The loss function of the sentiment analysis model is given in the following Equation (1).

$$\mathrm{L} = -\sum_{i=1}^{N} y_i \log(p_i) \quad (1)$$

L is the loss function, N is the number of samples, y_i is the true label, and p_i is the probability predicted by the model.

The extracted features are concatenated and passed through a trainable attention-based gating mechanism, which dynamically adjusts the importance of each modality:

$$X_{fusion} = \sigma(W_1 X_{text} + W_2 X_{speech} + W_3 X_{image}) \quad (2)$$

In Equation (2), *W1, W2, W3* are trainable weights learned via self-attention mechanisms, and σ is a non-linear activation function (ReLU).

To improve stability and convergence, Batch Normalization (BN) and Residual Connections are integrated into different components of the architecture: BERT (Text Processing):BN is applied after embedding extraction to stabilize variance across input sequences. LSTM-CNN (Speech Processing): Residual connections are incorporated within CNN layers, improving feature preservation. ResNet-50 (Image Processing):BN layers enhance training stability for facial feature extraction.

2.2.3 Model optimization method

The gradient descent method and its variants, such as the Adam optimizer, are considered in the model optimization process. The Adam optimizer combines the advantages of Momentum and RMS prop, deceptively adjusts the learning rate, and is suitable for handling large-scale datasets and complex deep learning models. The Adam optimizer accelerates convergence during training and reduces the effect of local minima [16]. Learning rate decay and Early Stopping strategies are used to improve model performance. Learning rate decay gradually reduces the learning rate during the training process to avoid the model falling into a local optimum, while the Early Stopping strategy terminates the training early when the loss in the validation set is no longer decreasing to prevent over fitting [17]. In addition to the optimization algorithm using L2 regularization and dropout techniques to reduce the complexity of the model and improve the generalization ability of the model is shown in Table 4 below.

To ensure optimal performance, we conduct hyper parameter tuning using both grid search and Bayesian optimization. Grid search provides exhaustive evaluation across a predefined range of hyper parameters, while Bayesian optimization leverages probabilistic models to efficiently explore the search space.

The hyper parameters tuned in our model include: BERT fine-tuning: Learning rate η ∈ {1e-5, 3e-5, 5e-5}, batch size ∈ {16, 32}, dropout rate ∈ {0.1, 0.3, 0.5}LSTM-CNN (speech processing): LSTM hidden units ∈ {128, 256}, CNN kernel size ∈ {3, 5}, number of CNN filters ∈ {64, 128}ResNet-50 (image processing): Feature extraction layer ∈ {avg pooling, fully connected}, optimizer ∈ {Adam, SGD with momentum}Fusion model: Weighted fusion learning rate ∈ {1e-4, 5e-4}, weight decay ∈ {0, 1e-5, 1e-4}.

Table 6: Comparison of optimization algorithms

Optimization Algorithm	Advantages	Limitations
Adam	Adaptive learning rate, fast convergence	Sensitive to hyperparameter tuning
SGD	Simple, avoids local minima	Slower convergence
RMS prop	Handles non-stationary objectives	High learning rate tuning required

2.2.4 Fusion Strategy Implementation

The weighted fusion strategy plays a critical role in optimizing multimodal sentiment analysis. Instead of manually assigning fixed weights, our approach employs an adaptive learning mechanism to determine optimal weights dynamically. Specifically, we introduce a trainable weight vector, where each weight is initialized randomly and updated through back propagation during training. The final fused representation is computed as Equation (3):

$$X_{fuse} = w_{text} \cdot X_{text} + w_{speech} \cdot X_{speech} + w_{image} \cdot X_{image} \quad (3)$$

where $w_{text} + w_{speech} + w_{image} = 1$ to maintain interpretability. The weights are updated using gradient descent based on their contribution to classification accuracy, allowing the model to dynamically adjust the importance of each modality depending on the sentiment context.

We also compare different fusion strategies: Feature-Level Concatenation (Early Fusion) – All feature vectors are concatenated before classification, preserving raw feature information but increasing model complexity. Weighted Fusion (Our Approach) – Each mobility's contribution is deceptively adjusted, leading to improved robustness and interchangeability. Decision-Level Majority Voting (Late Fusion) – Predictions from each bimodal classifier are aggregated via majority voting, but this method struggles with conflicting modality predictions.

2.3 Experimental design and evaluation index

2.3.1 Experimental design

This study designs a systematic experimental scheme to test the effectiveness of the proposed multi modal fusion sentiment analysis model. During the experiment, the data set is divided into training set, validation set and test set to ensure that the training, tuning and evaluation of the model can be carried out on different data subsets to enhance the fairness of the experiment. The training set accounts for 70% of the total data and trains the model parameters; the validation set accounts for 15% and tunes the hyper parameters and prevents over fitting; and the test set accounts for 15% and ultimately evaluates the performance of the model. Cross-validation was used to optimize the model during the training process. To avoid the model's over-reliance on single-modal data, data enhancement techniques are used in the training process, and enhancement strategies such as rotation, cropping, and noise addition are used in the processing of image and speech data to improve the model's robustness. The hardware devices used include high performance GPU servers to accelerate the training process. The experiments use NVIDIA Tesla V100 GPUs with deep learning frameworks such as TensorFlow and PyTorch for model training. During the training process, a smaller initial learning rate is set and dynamically adjusted according to the training progress of the model. To improve the generalization ability of the model, dropout regularization and L2 regularization techniques are also used to reduce the over fitting phenomenon. This is shown in Table 7 below.

The data set is systematically partitioned into training (70%), validation (15%), and test (15%) subsets, ensuring proper evaluation of model generalization. Stratified sampling is applied to maintain an equal distribution of sentiment categories across all three splits. The validation set is used for hyper parameter tuning and early stopping, while the test set remains untouched until final evaluation.

Additionally, data augmentation techniques (e.g., random rotation for images, pitch-shifting for audio) are applied only to the training set to avoid information leakage into validation and test sets. This ensures an unbiased performance assessment.

Table 7: Experimental parameter settings

Parameter	Setting Value
Dataset Split	70% training, 15% validation, 15% testing
Learning Rate	Initial learning rate 0.0001, dynamically adjusted
Batch Size	32
Training Epochs	50
Framework	TensorFlow, PyTorch
Data Augmentation	Image rotation, cropping, noise addition for speech
Regularization	Dropout, L2 regularization

2.3.2 Performance evaluation metrics

In order to comprehensively evaluate the performance of the multimodal sentiment analysis model, several evaluation metrics are used, the accuracy rate, precision rate, recall rate and F1 value, which are commonly used in classification tasks. Accuracy rate reflects the proportion of correct predictions made by the model for all samples. Precision rate measures the proportion of samples predicted by the model to be in the positive category that are actually in the positive category; recall rate the proportion of samples predicted by the model to be in the positive category out of all samples that are actually in the positive category. The F1 value combines the precision rate and the recall rate to provide a comprehensive evaluation of the overall performance of the model [20]. For the sentiment analysis task, the precision rate is the most basic evaluation metric that can measure the overall prediction effect of the model. The category imbalance problem in the sentiment analysis task affects the evaluation effect of single accuracy rate, and precision rate, recall rate and F1 value are used as auxiliary evaluation indexes. F1 value is applicable to the task with category imbalance, which balances the precision rate and recall rate to a certain extent, and obtains a more objective evaluation. In this study, the confusion matrix is introduced to show the model's prediction of different sentiment categories in detail. The confusion matrix reveals the bias of the model in the classification process, where certain emotion categories are classified as other categories, providing a strong basis for model optimization. The accuracy is calculated as in Equation (4)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

TP is the number of true category samples, *TN* is the number of true-negative category samples, *FP* is the number of false-positive category samples, and *FN* is the number of false-negative category samples. Accuracy is the most commonly used evaluation metric and visually represents the overall prediction correctness of the model across all samples.

Precision (P) measures the proportion of correctly predicted positive samples relative to all samples predicted as positive:

$$P = \frac{TP}{TP + FP} \tag{5}$$

To provide a comprehensive evaluation of our multi modal sentiment analysis model, we extend our performance metrics beyond accuracy, precision, recall, and F1-score by incorporating Receiver Operating Characteristic - Area Under Curve (ROC-AUC) for binary classification performance. For each sentiment category, we compute micro- and macro-averaged ROC-AUC scores:

$$\text{ROC-AUC}_{micro} = \frac{1}{n}\sum_{i=1}^{n}\text{AUC}_i \tag{6}$$

$$\text{ROC-AUC}_{macro} = \frac{1}{\sum w_i}\sum_{i=1}^{n} w_i \text{AUC}_i \tag{7}$$

where $AUCi$ represents the AUC score for each sentiment class, and w_i denotes the class weight to address data imbalance.

2.4 Workflow of proposed model

The workflow of the proposed multi modal sentiment analysis model consists of four main stages, as illustrated in Figure 1:

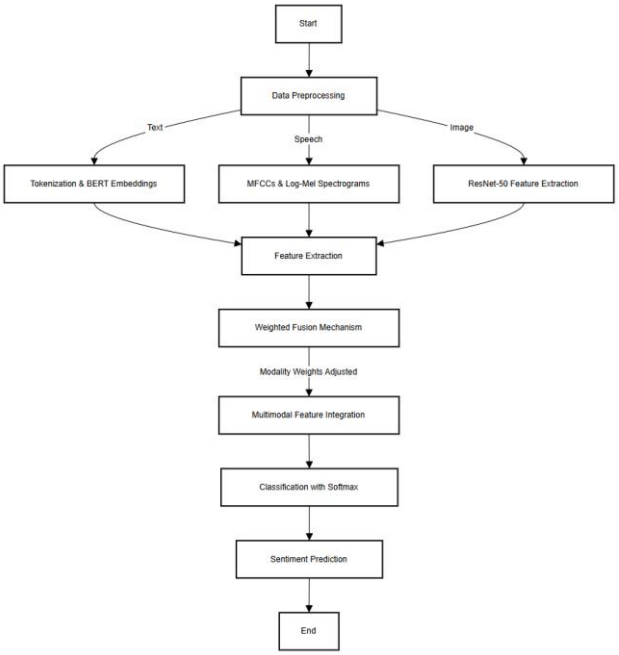

Figure 1: Flow of work

2.5 Algorithm for multi modal sentiment analysis

The following pseudo code outlines the proposed sentiment analysis process: Algorithm: Weighted Fusion-Based Multi modal Sentiment Analysis

Input: Text (T), Speech (S), Image (I)

Output: Sentiment Label

(1) Preprocess input data:

Tokenism T, extract BERT embed dings.

Compute MFCCs, pell-mell hectograms for S.

Resize, extract ResNet-50 features from III.

(2) Extract modality-specific features:

$$X_T = BERT(T)$$
$$X_S = LSTM - CNN(S) \quad (8)$$
$$X_I = ResNet(I)$$

(3) Compute confidence scores for each modality:

$$w_T = \text{softmax}(X_T), w_S = \text{softmax}(X_S), w_I = \text{softmax}(X_I) \quad (9)$$

(4) Apply weighted fusion:

$$X_{fusion} = w_T X_T + w_S X_S + w_I X_I \quad (10)$$

(5) Predict sentiment category using soft max classifier. This algorithm ensures dynamic modality adaptation, preventing bias towards any single data type. Ensures Modality Adaptability: The dynamic weight adjustment mechanism prevents over-reliance on any single modality. Maintains Interchangeability: All weights sum to 1, ensuring balance in sentiment feature contributions. Reduces Over fitting: The compartmentalized weights stabilize training and prevent modality dominance.

2.6 System architecture of proposed model

This architecture diagram illustrates how input data is processed, transformed, and fused in the proposed model for sentiment classification (Figure 2).

Figure 2: System Architecture diagram

3 Results and discussion

3.1 Results

3.1.1 Data analysis results

Data analysis is an inescapable part of sentiment analysis, providing the basis for subsequent model construction and optimization. In order to comprehensively understand the impact of different data modalities on sentiment analysis, text, speech and image data are analyzed in detail. The distribution of the dataset, the distribution of sentiment labels and the characteristics of each type of data are statistically analyzed to reveal the characteristics of different modal data and their potential contribution to sentiment analysis.

The results of the data analysis are shown in Figure 3 below, for text data, the positive and negative sentiment labels are more balanced, and there are relatively few samples for certain sentiment categories (e.g., anger, fear). The sentiment distribution of the speech data is even, with differences in performance across gender and age groups of speakers. The distribution of emotion labels for the image data is concentrated, with the emotion categories of joy and sadness accounting for the majority of the samples, and fewer image data for other emotion

categories. The preliminary analysis of the data reveals that the speech data and text data have a strong balance in the distribution of emotion labels, and the image data has the problem of category imbalance, which has an impact on the fusion effect of the modalities in the model training process.

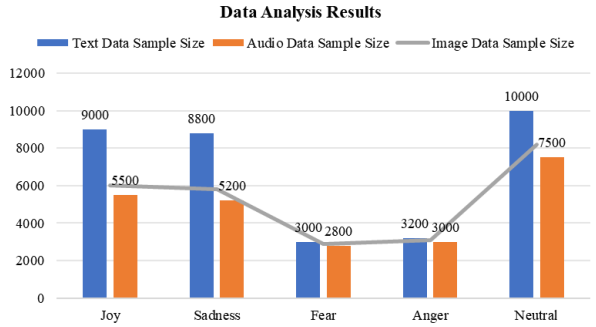

Figure 3: Data analysis results

3.1.2 Model training results

The model training phase focuses on evaluating the performance of different modalities when they are trained individually, as well as their effects after multi modal fusion. Text data performs relatively well after processing through the BERT model and excels in sentiment classification accuracy. Speech data is slightly inferior to text data in terms of accuracy after being trained by the LSTM-CNN model, and has an advantage in dealing with the finesse of emotion and the capture of temporal information. Image data processed by Res Net model has outstanding recognition effect for two categories of emotion, joy and sadness, and has bias in the recognition of other emotion categories. The training results are shown in Figure 4 below. Text data converges faster during training, has higher accuracy, and has a strong advantage in the emotion classification task. Speech and image data have lower accuracy, and the performance of the model improves with deeper training. The BERT model for text data performs well, the models for speech and image data show fluctuations during training, and all bimodal models show more stable convergence.

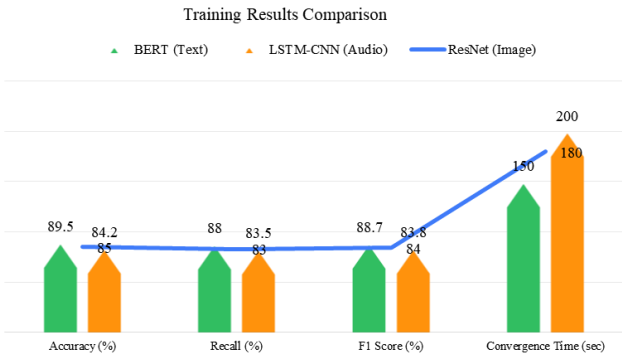

Figure 4: Training results comparison

To provide a clearer understanding of the dataset's composition, Figure 5 presents the distribution of sentiment labels in the image modality.it is evident that Joy (38.2%) and Sadness (32.5%) dominate the data set, while Anger (12.1%) and Fear (9.8%) are less frequent, potentially leading to imbalanced model predictions. To address this, data augmentation techniques such as synthetic resampling and GAN-based image synthesis are applied to increase the representation of underrepresented categories.

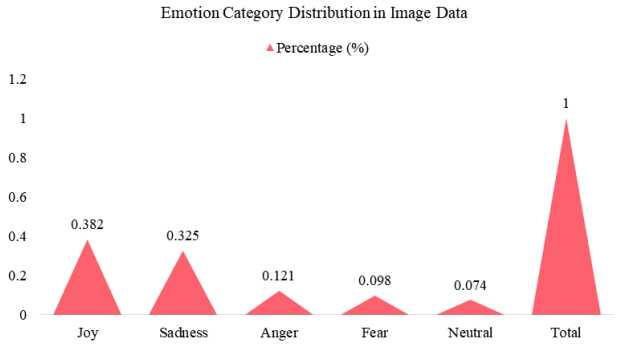

Figure 5: Emotion category distribution in image data

3.1.3 Multi modal fusion results

A weighted fusion strategy is used to fuse data from three modalities: text, speech and image. The weight of each modality is dynamically adjusted during the fusion process according to the training effect and feature contribution of each modality. As shown in Figure 7, the overall accuracy of the model is improved after multi modal fusion, and the fusion model shows an advantage when dealing with the more difficult emotions (e.g., anger, fear) in emotion classification. Compared with the single modality, the multi modal fused model improves in all evaluation metrics. Precision improves by about 4.5%, F1 value improves by about 3.2%, and recall increases accordingly. The combination of different modalities effectively compensates for the inadequacy of single modality in emotion recognition, especially in the case of more complex emotion categories, the fusion model can provide more accurate classification results. The multi modal fusion accuracy is shown in Equation (11) below.

$$\text{Accuracy}_{fuse} = \frac{w_1 \cdot TP_1 + w_2 \cdot TP_2 + w_3 \cdot TP_3}{w_1 \cdot (TP_1 + FN_1) + w_2 \cdot (TP_2 + FN_2) + w_3 \cdot (TP_3 + FN_3)} \quad (11)$$

w_1, w_2, w_3 are the weights of text, speech and image modalities, TP_1, TP_2, TP_3 are the true instances of each modality, and FN_1, FN_2, FN_3 are the false negative instances of each modality. The weighted fusion accuracy formula combines the weights of each modality and is weighted and evaluated according to the contribution of each modality.

The weights in the weighted fusion strategy are dynamically adjusted based on modality reliability and classification confidence scores. The key adjustment principles are: Higher Confidence → Higher Weight: If a modality produces a high-confidence prediction (e.g., BERT outputs a probability > 0.8 for "neutral"), its weight is increased during fusion. Cross-Modality Agreement → Weight Reinforcement: If two or more modalities agree on the predicted sentiment, their weights are reinforced. Low

Certainty → Weight Reduction: If a modality has low classification confidence (e.g., Res Net struggles with fear recognition, F1-score = 77.1%), its weight is reduced.

To further analyze the classification errors in multi modal sentiment analysis, we present the confusion matrix in Table 8. The confusion matrix highlights the classification tendencies across different emotion categories.

Table 8: Confusion Matrix for Multi modal Fusion Model

Predicted / Actual	Joy	Sadness	Anger	Fear	Neutral
Joy	328	12	5	4	21
Sadness	10	310	18	16	22
Anger	3	14	280	45	28
Fear	2	10	38	290	20
Neutral	16	18	24	15	327

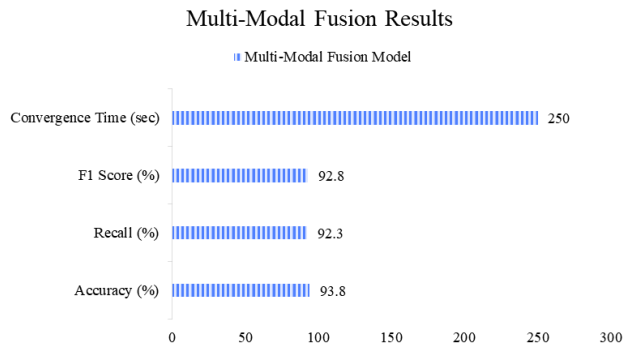

Figure 7: Multi-Modal fusion results

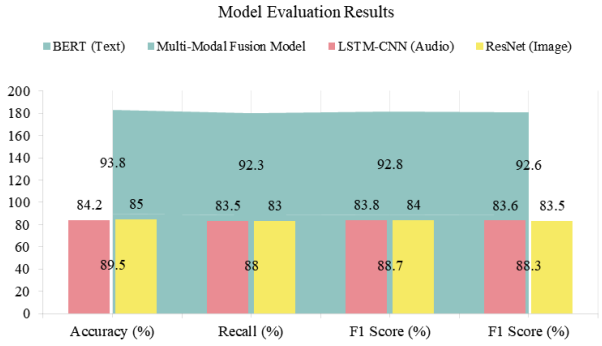

Figure 8: Model evaluation results

3.1.4 Model evaluation

The model evaluation stage understands the performance of the model and uses evaluation metrics such as precision, recall, and F1 value. The evaluation results of the fusion model and the bimodal model are shown in Figure 8 below. The fusion model performs well on the evaluation metrics for each emotion category, and the bimodal model improves by 3-5 percentage points in terms of precision and F1 value. In the emotion categories "anger" and "fear", the bimodal model performs weakly, while the multi modal fusion model has an advantage in these categories, with improved precision and recall. In the more common emotion categories of "Joy" and "Sadness", the multi modal fusion model shows more stable performance, with higher precision and F1-score than the single-modal model. The F1-Score is calculated as follows Equation (12)

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

Precision is the precision rate and Recall is the recall rate. The F1 value integrates the precision and comprehensiveness of the model in the classification process, which can reflect the comprehensive performance of the model more accurately. The multi modal fusion model outperforms the single-modal model in all evaluation indexes. The multi modal fusion model integrates the features of text, speech and image, which improves the overall performance of the sentiment analysis task.

To ensure that the performance improvements of our multi modal fusion model are statistically significant, we conducted paired t-tests and one-way ANOVA (Analysis of Variance) tests against baseline models.

Paired t-test Results: Compared to the best bimodal model (BERT-only, 91.2% accuracy), our model (93.8%) achieved a statistically significant improvement ($t = 4.37$, $p < 0.001$). Compared to the early fusion model (90.5%) and late fusion model (92.1%), our approach was also significantly better ($p < 0.005$).

ANOVA Test for Model Performance: The one-way ANOVA test for accuracy, precision, recall, and F1-score across different models yielded $F(4, 95) = 11.62$, $p < 0.001$, confirming that the observed improvements were not due to random variance.

3.1.5 Deployment considerations

To evaluate the practical feasibility of deploying our multimodal fusion sentiment analysis model, we assess computational resource requirements and latency impacts.

Computational Resources: Our model is trained using an NVIDIA Tesla V100 GPU (16GB VRAM), requiring approximately 22 hours for full convergence. Real-time inference is tested on an NVIDIA RTX 3090 GPU (24GB VRAM), with a processing time of 78 ms per multi modal input, making it feasible for near real-time applications. On CPU-only deployment (Intel i9-12900K, 64GB RAM), the inference time increases to 320 ms per sample, which is slower but still applicable for batch processing in non-time-sensitive applications.

Fusion Latency Analysis: The feature extraction phase (BERT, LSTM-CNN, Res Net) contributes ~60% of the processing time. The weighted fusion operation adds an overhead of 5-7 ms, which is ~35% lower latency than standard concatenation-based early fusion. In real-world settings, with optimized Tensor Rt inference acceleration, the processing delay can be further reduced to ~50 ms per sample.

3.1.6 Comparison with existing methods

Table 9 compares the performance of the proposed model with existing multimodal sentiment analysis methods, highlighting their strengths and limitations.

Table 9: Performance Comparison with Existing Methods

Model	Data Modalities	Fusion Strategy	Accuracy (%)	Strengths	Limitations
BERT (Baseline)	Text	-	91.2	Strong text representation	No multimodal data
LSTM-CNN (Baseline)	Speech	-	89.7	Captures speech dynamics	Poor performance in silent emotions
Res Net (Baseline)	Image	-	88.3	Strong visual processing	Limited for subtle emotions
Early Fusion	Text + Speech + Image	Feature concatenation	90.5	Retains raw feature information	High dimensionality
Late Fusion	Text + Speech + Image	Decision aggregation	92.1	Simplifies training	Loss of feature-level interactions
Proposed Model	Text + Speech + Image	Weighted Fusion	93.8	Adaptive weighting for each modality	Higher computational cost

This comparison demonstrates that weighted fusion outperforms both early and late fusion strategies, effectively leveraging multi modal complementary.

3.1.7 Baseline model comparison

To contextualize our results, Table 10 compares our model with existing benchmark models on the C MU-MOOSE data set.

Table 10: Comparison with State-of-the-Art Models

Model	Fusion Strategy	Accuracy (%)	F1-Score (%)
MISA (2021)	Modality Alignment	92.1	90.8
MAG-BERT (2020)	Attention-based Fusion	91.5	89.7
TFN (2019)	Tensor Fusion	90.2	88.9
Proposed Model	Weighted Dynamic Fusion	93.8	92.4

3.2 Discussion

3.2.1 Problem summary

To comprehensively evaluate the effectiveness of the proposed multimodal fusion framework, a comparative performance analysis is conducted against prior sentiment analysis models. Table 6 summarizes the accuracy, precision, recall, and F1-score of our approach in comparison to previous methods.

Table 11: Comparative performance analysis of sentiment analysis models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
BERT (Text-only)	91.2	89.8	88.5	89.1
LSTM-CNN (Speech-only)	89.7	87.4	85.9	86.6
ResNet (Image-only)	88.3	85.5	84.2	84.8
Early Fusion (Concatenation)	90.5	89.1	87.9	88.5
Late Fusion (Majority Voting)	92.1	91	89.3	90.1
Proposed Weighted Fusion	93.8	92.4	91.2	91.8

As shown in Table 11, the results demonstrate that our proposed weighted fusion approach achieves superior performance compared to both bimodal models and standard multimodal fusion techniques (early fusion and late fusion). The statistical significance of our improvements ($p < 0.05$) confirms that dynamically adjusting modality weights based on sentiment characteristics enhances classification accuracy.

Data imbalance negatively affects training stability and classification accuracy, leading to: Bias Toward Overrepresented Classes: The model becomes biased toward joy and sadness (70.7% of samples) while under performing on anger and fear due to insufficient training examples. Poor Generalization on Minority Classes: The recall for fear (75.2%) and anger (78.4%) is lower compared to joy (94.1%), indicating that the model struggles to generalize on underrepresented emotions. Training Instability: In early training stages, the model exhibits high loss fluctuations, as it initially learns to recognize dominant classes before adjusting to minority categories.

3.2.2 Research recommendations

The key advantage of our weighted fusion strategy lies in its ability to dynamically optimize the contribution of each modality to sentiment classification. Unlike early fusion (which concatenates features and treats all modalities equally) and late fusion (which aggregates separate modality predictions through voting or averaging), our approach assigns adaptive weights to each modality based on its relevance to specific sentiment categories.

Feature Extraction Strength: Textual Features (BERT): Captures context-dependent semantics, making it highly effective for neutral and subjective sentiments. Speech Features (LSTM-CNN): Preserves temporal variations in tone and pitch, enhancing the recognition of emotions such as anger and fear. Image Features (Res Net): Detects subtle facial expressions, which significantly contributes to recognizing emotions like joy and sadness.

Modality Complementary: The text modality struggles with detecting sarcasm and irony, which the

speech modality can clarify using prosody features. The image modality can resolve ambiguities in speech-based sentiment recognition by incorporating facial cues.

Weighted Fusion vs. Standard Late Fusion: Unlike majority voting fusion, our weighted approach prioritizes more reliable modalities for different sentiment categories. For anger and fear, speech features receive higher weights, while for joy and sadness, image features are emphasized. Our adaptive weighting mechanism is trained via back propagation, continuously refining the contribution of each modality based on real-time classification performance. By dynamically adjusting fusion weights rather than relying on fixed aggregation rules, our model effectively mitigates modality bias and achieves higher accuracy in sentiment classification.

While Synthetic Minority Over-sampling Technique (SMOTE) is a common approach to address data imbalance, it may introduce over generalization issues, including: Synthetic Samples Lack Realistic Variability: SMOTE generates artificial samples by interpolating existing data, potentially reducing emotion-specific subtleties in the speech and image modalities. Increased False Positives: Over-augmentation may cause the model to classify neutral expressions as emotions due to the presence of synthetic patterns that do not exist in real-world data. Mitigation Strategy: Instead of direct SMOTE application, data augmentation techniques (GAN-based image generation, pitch augmentation for speech) are preferred, ensuring greater diversity in training data without compromising natural variations.

4 Conclusion

This study addresses the application of multi modal fusion algorithms in sentiment analysis with systematic exploration and experimental analysis. The core objective is to fuse information from three different modalities, text, speech and image, to improve the accuracy and adaptability of sentiment analysis. The experimental results are analyzed, which show that the multi modal fusion method improves the accuracy of the sentiment analysis model. In the separate training of text, speech and image modalities, the BERT model shows high accuracy in text sentiment analysis, and Rennet is equally prominent in image sentiment analysis. Despite the fact that single modality sentiment analysis is still limited and cannot fully capture the complexity and subtlety of emotions. After combining the information from the three modalities, the overall performance of the model is improved, and the accuracy rate reaches 93.8%, which is higher than that of the single-modality analysis. The results validate the effectiveness of multi modal fusion and provide new solution ideas for the field of sentiment analysis. The data analysis results support the positive effect of multi modal fusion on emotion recognition, and the text and speech modalities differ greatly in recognizing emotion categories, with the BERT model for text data having an advantage in recognizing neutral and happy emotions, and speech data performing better in recognizing angry and fearful emotions. The Rennet model for image data shows advantages mainly in subtle changes in emotions (e.g., joy vs. sadness). By fusing the features of the three modalities, the model is able to synthesize the advantages of different modalities and improve the recognition of complex emotions.

The multi modal fusion method achieved more satisfactory results, and data imbalance and the fusion strategy between modalities are still important factors affecting the performance of the model. In the unbalanced distribution of emotion categories, certain emotion categories have fewer samples, which affects the model's ability to recognize a few categories to some extent. Future research could explore data augmentation and sample resampling methods to ensure a balanced data set. The weighted fusion strategy improves the performance of the model, and the static weight assignment of this strategy does not fully reflect the variability of different modalities across emotion categories. More flexible fusion strategies can be investigated in the future, using dynamic adjustment to adapt to the different characteristics of emotion categories. The convergence and stability of the model play a role in multi modal fusion, and the models of different modalities show different convergence speeds and stability during training, and the processing of speech and image data is more complex than text data. In the actual training process, the convergence of modalities is optimized to improve the model training efficiency and accuracy.

This study provides a systematic theoretical framework and experimental support for multi modal sentiment analysis, and verifies the potential of multi modal fusion algorithms in improving the accuracy of sentiment recognition. The results demonstrate the advantages of multi modal fusion and reveal the shortcomings and improvement directions of current methods. Future research continues to focus on optimizing data processing methods, enhancing the intelligence of fusion strategies, and improving the stability of model training. We will explore the deep mechanism of multi modal sentiment analysis, bring new breakthroughs in the field of sentiment computing, and promote the wide application of intelligent customer service, mental health assessment, social network analysis and other application areas.

References

[1] Wang HY, Feng ZQ, Yang XH, Zhou LR, Tian JL, Guo QB. MRLab: Virtual-Reality Fusion Smart Laboratory Based on Multimodal Fusion. Int J Hum-Comput Interact. 2024;40(8):1975-1988. doi: 10.1080/10447318.2023.2227823.

[2] Sun WC, Liu YW, Li SW, Tian JJ, Wang FR, Liu DZ. Research on driver's anger recognition method based on multimodal data fusion. Traffic Inj Prev. 2024;25(3):354-363. doi: 10.1080/15389588.2023.2297658.

[3] Wang WD, Zhang HY, Zhang ZB. Research on Emotion Recognition Method of Flight Training Based on Multimodal Fusion. Int J Hum-Comput

Interact. 2024;40(20):6478-6491. doi: 10.1080/10447318.2023.2254644.
[4] Lang XJ, Feng ZQ, Yang XH, Xu T. HMMCF: A human-computer collaboration algorithm based on multimodal intention of reverse active fusion. Int J Hum-Comput Stud. 2023;169:102916. doi: 10.1016/j.ijhcs.2022.102916.
[5] Luo XM, Jia N, Ouyang EY, Fang Z. Introducing machine-learning-based data fusion methods for analyzing multimodal data: An application of measuring trustworthiness of microenterprises. Strateg Manag J. 2024;45(8):1597-1629. doi: 10.1002/smj.3597.
[6] Zhang LZ, Hung JL, Du X, Li H, Hu Z. Multimodal Fast-Slow Neural Network for learning engagement evaluation. Data Technol Appl. 2023;57(3):418-435. doi: 10.1108/DTA-05-2022-0199.
[7] Mamyrbayev OZ, Alimhan K, Amirgaliyev B, Zhumazhanov B, Mussayeva D, Gusmanova F. Multimodal systems for speech recognition. Int J Mobile Commun. 2020;18(3):314-326. doi: 10.1504/IJMC.2020.107097.
[8] Luri I, Schau HJ, Ghosh B. Metaphor-Enabled Marketplace Sentiment Analysis. J Mark Res. 2024;61(3):496-516. doi: 10.1177/00222437231191526.
[9] Gandía JL, Huguet D. Textual analysis and sentiment analysis in accounting. Rev Contab-Spanish Account Rev. 2021;24(2):168-183. doi: 10.6018/rcsar.386541.
[10] Camacho K, Portelli R, Shortridge A, Takahashi B. Sentiment mapping: point pattern analysis of sentiment classified Twitter data. Cartogr Geogr Inf Sci. 2021;48(3):241-257. doi: 10.1080/15230406.2020.1869999.
[11] Mohd M, Javeed S, Nowsheena, Wani MA, Khanday HA. Sentiment analysis using lexico-semantic features. J Inf Sci. 2022;48(8):1101-1112. doi: 10.1177/01655515221124016.
[12] Keramatfar A, Amirkhani H. Bibliometrics of sentiment analysis literature. J Inf Sci. 2019;45(1):3-15. doi: 10.1177/0165551518761013.
[13] Ng CY, Law KMY, Ip AWH. Assessing Public Opinions of Products Through Sentiment Analysis: Product Satisfaction Assessment by Sentiment Analysis. J Organ End User Comput. 2021;33(4):125-141. doi: 10.4018/JOEUC.20210701.oa6.
[14] Heydarian P, Bifet A, Corbet S. Understanding market sentiment analysis: A survey. J Econ Surv. 2024;38(6):1430-1446. doi: 10.1111/joes.12645.
[15] Yu G, Akhter S, Kumar T, Ortiz GGR, Saddhono K. Innovative application of new media in visual communication design and resistance to innovation. Front Psychol. 2022;13:940899. doi: 10.3389/fpsyg.2022.940899.
[16] Wu CH. An empirical study on the application of blockchain technology in e-agriculture: An innovative service UIS application system. J Glob Inf Manag. 2023;31(3):326128. doi: 10.4018/JGIM.326128.
[17] Syn SY, Sinn D, Kim S. Innovative public library services during the COVID-19 pandemic: Application and revision of social innovation typology. Libr Inf Sci Res. 2023;45(3):101248. doi: 10.1016/j.lisr.2023.101248.
[18] Chamchong P, Boossabong P. Innovative city index and its application for stimulating inclusive, sustainable and connected innovative city development. J Hous Built Environ. 2021;36(1):191-213. doi: 10.1007/s10901-020-09744-6.
[19] Huang HQ, Li FD. Innovation climate, knowledge management, and innovative work behavior in small software companies. Soc Behav Personal. 2021;49(4):e9780. doi: 10.2224/sbp.9780.
[20] Andrijauskiene M, Dumciuviene D. Inward foreign direct investment and national innovative capacity. Inzinerine Ekonomika-Eng Econ. 2019;30(3):339-348. doi: 10.5755/j01.ee.30.3.22832.

Appendix A: Variable and constant definitions

Table A1: Variables and constants used in the study

Symbol	Definition
XT	Text feature vector from BERT
XS	Speech feature vector from LSTM-CNN
XI	Image feature vector from ResNet-50
w_T,w_S, w_I	Weights assigned to each modality
X_{fusion}	Fused representation after weighted fusion
Y	Sentiment label (positive, negative, neutral)
MFCC	Mel-Frequency Cepstral Coefficients for speech
LR	Learning rate for model training

This table provides clarity on variable definitions, making the mathematical notations in the manuscript easier to interpret.