Multimodal Sentiment Analysis of Twitter Data Using Early Fusion Along with a Fully Connected Neural Network and Multilayer Perceptron Framework

T. S. Kaveri
Department of Information Science and Engineering, JSS Science and Technology University, JSS TI Campus, Mysore, Karnataka, India
kaveri@jssstuniv.in

B. S. Harish
Department of Information Science and Engineering, JSS Science and Technology University, JSS TI Campus, Mysore, Karnataka, India
bsharish@jssstuniv.in (corresponding author)

C. K. Roopa
Department of Information Science and Engineering, JSS Science and Technology University, JSS TI Campus, Mysore, Karnataka, India
ckr@jssstuniv.in

M. S. Kendagannaswamy
Department of Information Science and Engineering, JSS Science and Technology University, JSS TI Campus, Mysore, Karnataka, India
kswamy@jssstuniv.in

Received: 4 October 2025 | Revised: 22 November 2025, 3 December 2025, and 4 December 2025 | Accepted: 5 December 2025

 DOI: https://doi.org/10.48084/etasr.15303

ABSTRACT

Multimodal sentiment analysis of social media data remains challenging due to the complex integration of textual and visual information. A Fully Connected Neural Network and Multilayer Perceptron-based Sentiment Analysis (SA-FCNNMP) framework is proposed for effective multimodal sentiment classification of Twitter data. The approach combines textual and visual data to improve sentiment comprehension and is tested on the Multi-View Sentiment Analysis (MVSA) Single Twitter Dataset, which consists of 4,869 images labeled with their corresponding texts. The approach begins by cleaning textual data through stop-word removal using Bayesian Boundary Trend Filtering (BBTF), preserving only meaningful words. Subsequently, features are extracted from text and images independently to leverage their unique properties. The input textual data are represented using Word2Vec embeddings that encode semantic relationships, whereas image features are extracted using a ResNet-50 convolutional neural network. These features are then fused at an early stage using Hierarchical Multi-Scale Feature Fusion (HMSFF), which incorporates information across all scales and modalities into a consistent representation. The fused features are channeled through a Fully Connected Neural Network and Multilayer Perceptron (FCNNMP), which is optimized using the data to classify sentiment more effectively. The SA-FCNNMP model categorizes sentiments as positive, negative, or neutral. Performance is evaluated using accuracy, precision, recall, sensitivity, and computational time. To maximize learning and prediction robustness, the model uses both cross-entropy loss, which works well for multi-class classification problems, and Mean Squared Error (MSE) loss, which captures finer-grained variations in sentiment distribution. Compared with existing state-of-the-art methods, experimental results demonstrate that the proposed SA-FCNNMP model outperforms existing methods in multimodal sentiment analysis on Twitter data.

Keywords-multimodal sentiment classification; Twitter; fully connected neural network; multilayer perceptron; Bayesian Boundary Trend Filtering (BBTF); Word2Vec; ResNet-50

I. INTRODUCTION

With the rapid multiplication of mobile devices and internet services [1], the world has entered a multimedia big data era. Every phone call, message, picture, or online search contributes to the generation of big data. Every single action creates massive amounts of data every day. There are around 4.2 billion active social media users throughout the world [1]. The easy availability of multimedia data has resulted in a big data revolution. With the rise of new technologies, social networking, internet services, and the extensive expansion of usage of Web 2.0, people, especially the younger generations, spend a lot of time on the Internet to communicate and share their emotions and opinions with each other. This has led to an increase in user-generated content and self-opinionated data [2-4]. Human emotions are nuanced, often expressed through body language, facial expressions, and words, all of which are considered in emotion recognition. This complexity also extends to online communication, where people interact or convey information in the form of audio, video, images, and text. Thus, analyzing the sentiment behind these different modalities is a challenging task. Since there is a wide range of social media users, it has become a trend to use multimodal data for analytics and to apply the inference obtained from it for decision-making.

The process of gathering data from blog posts, websites, and online communication platforms and analyzing them to make informed business decisions is known as Social Market Analysis (SMA) [5-7]. In the modern world, using social media has become very usual. In addition to compiling user-shared comments and favorites, SMA serves as a platform for numerous advertising businesses [8]. People can connect and share their interests, thoughts, knowledge, and life occasions on different kinds of networking sites [9]. Social news enables users to share external news and content links. Media sharing enables users to distribute their images and videos, whereas microblogging permits users to publish brief articles, and blogs and forums give people a space to share their thoughts on specific topics and have meaningful conversations with others who are interested in the same things [10]. SMA has the capacity to collect information from these places, interpret it, make decisions, and assess how well those decisions worked out on social media.

Social media intelligence, social media listening, monitoring social networks, social competitive evaluation, image analytics, sentiment analysis, and customer sentiment analysis are some of the terms that SMA employs for this [11]. Marketing and the broad use of social data to make predictive judgments are only two examples of numerous possibilities. Some of the techniques are designed to generate opinions, map occurrences, and delve deeply into data [12]. These calculations can also be done in services such as business, advertisements, education, and machine learning-based predictions. Nowadays, it has become a trend to use multimodal data for analytics in marketing approaches and tactics [13]. Services like enterprises, schools, predictive modeling, and advertisements can also use these computations [14]. Since more individuals and companies intend to share personal information on social media, the spread of data is only anticipated to increase [15]. A company will ultimately learn more about its audience via this content, particularly on social media platforms like Facebook, Instagram, and Twitter. SMA is currently used for opinion mining, review, and impact [16]. These socioeconomic determinants of psychological wellness are significant, and these social factors are important indicators of mental health as well.

Initially, only text was used for sentiment analysis. Text data usually provide detailed and clear explanations and precise articulation of ideas. However, text struggles when it comes to complicated expressions like irony or sarcasm because a single modality usually gives incomplete information. With the emergence of multimodal sentiment analysis, not only text data but also the combination of text, image, audio, and video formats are integrated to provide additional context in analyzing sentiments. Text and visual data play complementary as well as contradictory roles because visual data conversely provide better understanding of the context and visual evidence, thus adding more clarity in classifying a sentence as positive, negative, or neutral. Despite significant progress in sentiment analysis, existing methods face challenges in effectively integrating heterogeneous modalities such as text and images for sentiment classification. Many approaches either focus solely on text or treat the two modalities independently, leading to suboptimal performance due to incomplete sentiment representation. Moreover, the current methods lack an efficient fusion strategy that can capture the complex interactions between text and image features at an early stage of the learning process [17-22]. This reduces the precision and stability of sentiment classification models, particularly in datasets such as the Multi-View Sentiment Analysis (MVSA) Single Twitter Dataset, in which both written and visual information are needed.

Numerous datasets are available for sentiment analysis that comprise of visual and textual data. The MVSA-Single and the MVSA-Multiple datasets [23, 24], consisting of 4,869 and 19,665 image–text pairs, respectively, are benchmark datasets, particularly focusing on Twitter data. These are restricted to social media environments. Although MVSA-Multiple has a larger size, it often contains posts with multiple images, introducing noise and complexity in aligning textual and visual content. MVSA-Single, on the other hand, contains a more manageable dataset of 4,869 posts, each with a single image, ensuring better alignment between text and image modalities, which improves model learning and interpretability. Certain datasets like Memotion 1.0 and Memotion 2.0 [25, 26] have around 7,000 and 10,000 images, respectively, along with their metadata and multilabel annotations. Memotion 1.0 provides a five-point sentiment scale, whereas Memotion 2.0 classifies the data as positive, negative, or neutral. Although they are valuable datasets, both exhibit higher class imbalance, which may result in bias during training. The Twitter-15 and Twitter-17 datasets [27] are widely used benchmark datasets for stance detection, rumor verification, and multimodal sentiment

analysis. Although they are useful, these datasets are specific to rumors and controversial topics, which may not generalize to other domains. The Multilingual Aspect-based Sentiment Analysis Dataset (MASAD) contains 38,532 data entries in multiple languages. MASAD, although large, is primarily geared towards Indian languages and is limited to only two sentiment classes, positive and negative, reducing its usefulness for balanced sentiment modeling [28].

Feature extraction is an important aspect in the analysis and classification of sentiments. For text feature extraction, approaches such as BERT and DistilBERT [29] exhibit the ability to maintain semantic context but at the cost of substantial computational power. For image processing, models such as DenseNet and Vision Transformer (ViT) [30, 31] display better accuracy but at the expense of processing speed, clearly indicating an urgent demand for other models. Multimodal fusion strategies are essential to improve the performance of models in understanding complex data. Multimodal sentiment analysis employs four primary fusion strategies with distinct characteristics and trade-offs [32].

Early fusion combines features before classification, effectively capturing uniform data patterns and ensuring stable properties, but struggles with heterogeneous inputs and may lose modality-specific information [33]. Intermediate fusion combines features from different modalities after partial extraction, thus enabling enhanced cross-modal interactions and effective noise reduction. However, this approach requires careful feature selection and balance between sub-models, resulting in high computational overhead [34]. Late fusion processes modalities independently and then combines their final predictions to determine the overall sentiment. This approach preserves modality-specific characteristics and requires less computing power. However, it risks losing critical cross-modal correlations [35]. Hybrid fusion combines all approaches for maximum flexibility and diversity but requires extremely high computational resources and complex architectural design [36].

To overcome these shortcomings, this study proposes a multimodal Sentiment Analysis approach based on early fusion and Fully Connected Neural Networks and Multilayer Perceptrons (SA-FCNNMP), which allows the mutual benefits of text and image features to be exploited. By applying early fusion at the feature level and employing a Fully Connected Neural Network and Multilayer Perceptron (FCNNMP) model, the method aims to improve feature interaction and sentiment distinction. This study is inspired by the need for more precise, efficient, and holistic sentiment analysis tools to process real-life social media data. The major contributions of the proposed work are summarized as follows: The SA-FCNNMP model is proposed for multimodal sentiment analysis. Features from the text and image modalities are fused using an early fusion strategy through Hierarchical Multi-Scale Feature Fusion (HMSFF) [21]. The proposed FCNNMP model then classifies the sentiments into negative, neutral, and positive categories, and its performance is evaluated against existing approaches.

II. METHODOLOGY

With the rapid evolution of social media platforms such as Twitter and in response to the identified limitations of existing datasets and methods for feature extraction and fusion, the analysis of rich multimodal data has become crucial for understanding public sentiment, brand perception, and social trends. Traditional sentiment analysis methods primarily rely on text-based data, often neglecting the valuable information embedded in accompanying images. Recent advancements in deep learning [37] and multimodal fusion techniques have enabled more effective sentiment analysis by jointly analyzing both textual and image data. The overall workflow of the proposed multimodal sentiment analysis framework is illustrated in Figure 1.

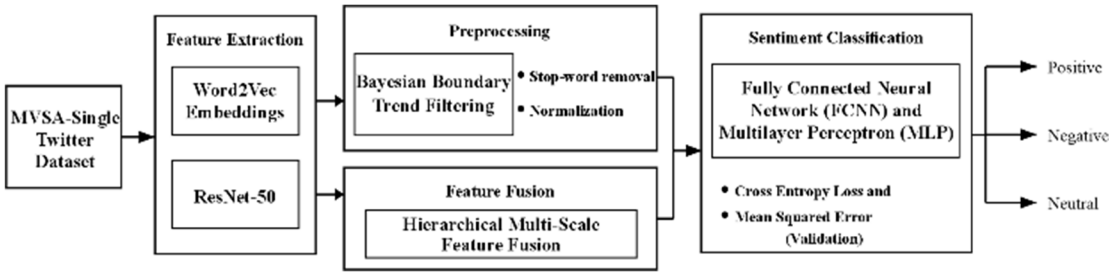

Fig. 1. Block diagram of the proposed SA-FCNNMP method.

A. Dataset

The MVSA-Single dataset [23] is a benchmark dataset designed for research in multimodal sentiment analysis, particularly focusing on Twitter data. It consists of 4,869 tweet–image pairs, where each entry includes a short text post (tweet), an associated image, and a sentiment label. The dataset captures the natural integration of visual and textual content in social media, enabling the development of models that analyze both modalities simultaneously. Each sample in the MVSA-Single dataset has been annotated by a single human annotator and assigned one of three sentiment categories: neutral, negative, or positive, depending on the combined interpretation of the image and its accompanying text. The dataset reflects real-world social media behavior, where sentiment is often conveyed through a blend of words and visuals, making it valuable for building and evaluating multimodal sentiment analysis models. Its balanced and diverse distribution of sentiment classes allows for robust performance comparisons across various deep learning and fusion-based sentiment classification approaches. Table I presents the distribution of positive, negative, and neutral samples in the MVSA-Single dataset.

TABLE I. DATASET STATISTICS FOR THE MVSA-SINGLE DATASET

Category	Count
Total image–text pairs	4,869
Positive samples	2,325
Negative samples	1,218
Neutral samples	1,326

B. Pre-Processing Using Bayesian Boundary Trend Filtering

During the pre-processing stage, Bayesian Boundary Trend Filtering (BBTF) [38] is applied to the textual data to enhance their quality and reliability before classification. BBTF is a statistical technique designed to refine datasets by approximating the underlying boundaries and trends within each feature. One of its primary roles in this context is to perform stop-word removal, which eliminates non-informative or frequently occurring words that do not contribute meaning to sentiment analysis. BBTF evaluates the probability of varying boundary conditions for every feature using a Bayesian framework, allowing it to filter out noise (including stop-words) while preserving significant semantic trends. This is particularly effective when dealing with social media text data, which often contain outliers, slang, or inconsistencies that can negatively affect model performance. The method works by estimating a smooth trend from noisy input data, balancing accuracy and smoothness through a regularization term that reduces error, is described in (1):

$$\min_{\varphi \in \Re^o} \sum_{j=1}^{o} \left(\varphi_j - z_j\right)^2 + \gamma \left\| E_o^{(l+1)} \varphi \right\|_1 \quad (1)$$

where z_j represents the probability density function, φ_j represents the scale parameter, E_o represents the half-normal distribution, $(l+1)$ represents the piecewise polynomials, and γ represents the input points. This formulation describes the goal of estimating a smooth trend from noisy data, achieving a balance between accuracy and smoothness by decreasing error and using a regularization term. The difference operator matrix is given in (2):

$$E_o^{(l+1)} = E_{o-l}^{(1)} E_o^{(l)} \quad (2)$$

where $E_o^{(l)}$ represents the data points. This recursive formulation improves the ability to identify minor changes in the trend. The probability density function z_j is given in (3):

$$z_j = \varphi_j + \omega_j, \;\; \omega_j \sim q(.) \quad (3)$$

where ω_j denotes independent errors, and $q(.)$ signifies the Laplace density. This demonstrates that each data point is interpreted as a real value plus some noise. The noise is expected to follow a probability distribution appropriate for data. The estimation of the upper boundary trend is given in (4):

$$q\left(z_j \middle| \varphi_j, \rho^2\right) = \sqrt{\frac{2}{\pi\rho^2}} \exp\left(-\frac{1}{2\rho^2}\left(z_j - \theta_j\right)^2\right) 1_{\{z_j \le \varphi_j\}}\left(z_j\right) \quad (4)$$

where q represents the local parameter, ρ^2 represents the error variance, $\sqrt{2/(\pi\rho^2)}$ is a global parameter, and $z_j - \theta_j$ represents the normal prior. This formulation determines the likelihood of observing specific data values under given conditions and ensures that only numbers within an acceptable range are used in trend estimates. The same concept is presented in (5):

$$E = \left(\frac{J_{l+1}}{E_o^{(l+1)}} P\right) \quad (5)$$

where E represents the difference operator, J_{l+1} represents the identity matrix, and P represents the zero matrix. This generates a matrix that identifies points where major changes in data trends occur and combines components to provide effective trend filtering during pre-processing. Finally, BBTF is used to refine the quality and reliability of the input data, which are then fed into the feature extraction phase.

C. Feature Extraction

Feature extraction is performed separately for the text and image modalities. For textual data, Word2Vec embeddings [39] are employed to capture semantic features, whereas for images, ResNet-50 [40] is used to extract deep visual features.

1) Textual Feature Extraction

For the textual modality, Word2Vec embeddings are used to extract meaningful semantic features from the tweet text [41]. Word2Vec is a widely used word embedding technique that converts words into dense, fixed-length vectors based on their context in a large corpus of text. The model can be trained using two architectures: Continuous Bag of Words (CBOW) and Skip-gram. In the Skip-gram model, the objective is to maximize the probability of context words w_{t+j} given a target word w_t, which is defined in (6):

$$\max \sum_{t=1}^{T} \sum_{-c \le j \le c, j \ne 0} \log P\left(w_{t+j} | w_t\right) \quad (6)$$

where T is the total number of words in the corpus, c is the context window size, and $P\left(w_{t+j} | w_t\right)$ is modeled using SoftMax over all vocabulary words. The conditional probability is estimated using (7):

$$P(w_O | w_I) = \frac{\exp(v'_{w_O} . v_{w_I})}{\sum_{w=1}^{W} exp(v'_w . v_{w_I})} \quad (7)$$

where v_{w_I} and v'_{w_O} are the input and output vector representations of the input word w_I and output word w_O, and W is the vocabulary size. In this approach, Word2Vec captures the contextual relationships among words by placing semantically similar words close together in the high-dimensional vector space. During the pre-processing phase, the tweet text undergoes cleaning, tokenization, and stop-word removal using BBTF. The processed text is then passed through the Word2Vec model, where each word is mapped to its corresponding vector representation. These word vectors are commonly aggregated by averaging across the sequence to form a composite sentence-level representation, which encapsulates the tweet's overall semantic and sentiment-bearing characteristics. This embedding-based representation serves as an effective input for subsequent fusion and classification, enabling the model to understand nuanced linguistic patterns, slang, and context-specific meanings commonly found in social media data.

2) Visual Feature Extraction

For the visual modality, deep features are extracted from the associated tweet images using a sophisticated convolutional neural network with 50 layers, called ResNet-50. The ResNet-50 model is well-known for its superior performance in image recognition tasks [40]. ResNet-50 introduces the concept of residual learning, where shortcut connections are used to mitigate the vanishing gradient problem and enable training of deeper networks without sacrificing performance. The images I are first resized and normalized according to the input requirements of the ResNet-50 architecture. Once pre-processed, the features of each image are extracted from one of its deeper layers, typically just before the final classification layer, as defined in (8):

$$y = F(x, \{W_i\}) + x \tag{8}$$

where $F(x, \{W_i\})$ represents the residual transformation (a sequence of convolution, batch normalization, and ReLU operations), x is the block's input, y is the output, and W_i are the learnable weights in the convolutional layers. The input x is added to the result of the transformation F, which helps training deeper networks without vanishing gradients. After passing through the network, the final deep visual feature vector $V_{img} \in \mathbb{R}^d$ is extracted from a fully connected layer, as described in (9):

$$V_{img} = F_{ResNet}(I) \tag{9}$$

These features represent high-level abstractions, such as objects, textures, facial expressions, colors, and other visual elements that may carry sentiment information. ResNet-50 is pre-trained, and possesses a strong ability to generalize across various image types and content styles. The resulting deep feature vectors effectively summarize the visual sentiment cues present in the images, which are essential for multimodal analysis. These extracted image features are then integrated with textual features during the fusion stage, enabling a comprehensive multimodal sentiment classification.

D. Hierarchical Multi-Scale Feature Fusion

The extracted features from both modalities are fused using an early fusion strategy through HMSFF [42], enabling comprehensive feature integration. Once the semantic features from the text modality and the deep visual features from the image modality are extracted, they are integrated using HMSFF. This technique is designed to effectively combine multimodal features at different levels of abstraction, allowing the model to capture both fine-grained and high-level interactions between text and image data. Unlike traditional fusion methods that treat all features uniformly, HMSFF fuses features hierarchically across multiple scales, preserving critical contextual and structural information from each modality. The process begins by aligning the dimensions of the text feature vector $F_t \in \mathbb{R}^{d_t}$ and the image feature vector $F_i \in \mathbb{R}^{d_i}$, typically through a linear transformation or projection into a common space $\mathbb{R}^d$. This is formulated in (10):

$$F_t'' = W_t F_t + b_t, \;\; F_i'' = W_i F_i + b_i \tag{10}$$

where W_t and W_i are learnable weight matrices, b_t and b_i are biases, and $F_t'', F_i'' \in \mathbb{R}^d$ are the transformed feature vectors.

Next, the transformed features are concatenated and passed through a series of hierarchical fusion layers that progressively integrate information from coarse to fine levels, as described in (11):

$$F_{fused}^l = \sigma\left(W_f^l [F_t'', F_i'']\right) + b_f^l \tag{11}$$

where $[F_t'', F_i'']$ denotes the concatenation of the aligned features, W_f^l and b_f^l are the weights and biases at hierarchical level l, and $\sigma(.)$ is a non-linear activation function, such as ReLU. This process is repeated across multiple layers $l = 1,2,\ldots L$, enabling multi-scale interaction learning. Both modalities contribute proportionally, and to suppress irrelevant information, attention mechanisms or gating functions are incorporated within HMSFF, dynamically adjusting feature weights based on their contextual importance. The final fused representation F_{fused} is obtained by aggregating outputs from different levels, as described in (12):

$$F_{fused} = \sum_{l=1}^{L} \alpha^l F_{fused}^l \tag{12}$$

where α^l are learnable coefficients representing the contribution of each hierarchical level. The fused output is then passed to the sentiment classification phase.

E. Sentiment Classification Using Fully Connected Neural Network and Multilayer Perceptron

In the proposed sentiment analysis framework, the fused multimodal features are passed into a novel FCNNMP model for effective sentiment classification [43]. The FCNNMP model is specifically designed to create a boundary between the neutral, negative, and positive sentiment categories while maintaining efficiency and accuracy in high-dimensional, resource-constrained environments. The architecture of FCNNMP consists of an input layer, multiple hidden layers, and an output layer. The input layer receives the fused feature vector $x \in \mathbb{R}^d$. Each hidden layer performs a linear transformation followed by a non-linear activation, typically the ReLU function. Mathematically, the output of a hidden layer l is computed as in (13):

$$h^{(l)} = f\left(W^{(l)} h^{(l-1)} + b^{(l)}\right) \tag{13}$$

where $W^{(l)}$ and $b^{(l)}$ are the weight matrix and bias vector of layer l, and $f(.)$ is the activation function. The final hidden layer connects to the output layer, which contains three neurons corresponding to the three sentiment classes. The output layer uses the softmax function to convert raw scores into class probabilities, calculated as in (14):

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^{3} e^{z_j}} \tag{14}$$

where $\hat{y}_i$ is the probability of class i. To train the model efficiently, FCNNMP employs a combination of cross-entropy loss and Mean Squared Error (MSE). The cross-entropy loss is defined in (15):

$$L_{CE} = -\sum_{i=1}^{3} y_i \, log(\hat{y}_i) \tag{15}$$

The MSE is given in (16):

$$L_{MSE} = \frac{1}{3} \sum_{i=1}^{3} (y_i - \hat{y}_i)^2 \tag{16}$$

MSE captures finer-grained sentiment differences and ensures robustness in cases of ambiguous sentiments. In some cases, a hybrid loss function combining both can be employed as in (17):

$$L_{total} = \alpha L_{CE} + (1 - \alpha) L_{MSE} \quad (17)$$

Here, α balances the contribution of each component. The model parameters are updated via backpropagation using gradient descent: $\vartheta \leftarrow \vartheta - \eta \forall_0 L_{total}$, ensuring that the network learns to minimize classification error while preserving nuanced sentiment distinctions. This FCNNMP design enables the proposed model to achieve both high accuracy and fine-grained sentiment resolution in multimodal Twitter data classification.

III. RESULTS AND DISCUSSION

The experimental results of the SA-FCNNMP method are discussed in this section. The experiments are performed in Python and assessed using several performance metrics, including accuracy, precision, recall, f1-score, specificity, error rate, and computational time. The performance of the SA-FCNNMP approach is analyzed and compared with existing techniques.

A. Performance Measures

The performance of the proposed method is evaluated using the metrics described below.

1) Accuracy

Accuracy is a commonly used criterion for evaluating the overall performance of a classification method. It is defined as the ratio of correctly predicted instances (true positives and true negatives) to the total number of instances in the dataset. It is calculated as:

$$\text{Accuracy} = \frac{TP+TN}{FN+TP+FP+TN} \quad (18)$$

where FN denotes False Negative, FP False Positive, TP True Positive, and TN True Negative.

2) Precision

Precision measures the exactness and reliability of the model in generating accurate results. It quantifies the proportion of correctly predicted positive instances out of all instances predicted as positive. It is given by:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (19)$$

3) Recall

Recall measures the model's ability to correctly identify all relevant positive instances. Higher recall indicates better detection of positive cases. It is defined as:

$$\text{Recall} = \frac{TP}{FN+TP} \quad (20)$$

4) F1-Score

The F1-score evaluates the model's performance by balancing precision and recall. It is defined as the harmonic mean of precision and recall, with a value of 1 indicating perfect performance and 0 indicating the poorest performance, as shown in (21):

$$\text{F1} - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (21)$$

5) Specificity

Specificity measures the proportion of actual negative instances that are correctly identified. It complements recall by evaluating the model's ability to correctly reject negative cases, as shown in (22):

$$\text{Specificity} = \frac{TP}{FN+TP} \quad (22)$$

B. Performance Analysis

Figures 2–6 present the experimental results of the SA-FCNNMP method. In addition, its performance is compared with existing sentiment classification models.

Figure 2 shows the training and validation accuracy of SA-FCNNMP over epochs. The proposed model achieves a peak validation accuracy of 84.20%. During the early stages of training, the training accuracy fluctuated and temporarily fell below the validation accuracy around 50 epochs, indicating that the model had not yet fully captured the patterns in the data. However, as training progressed, both training and validation accuracy improved, reaching their most balanced and effective point at 100 epochs, where the validation accuracy peaked.

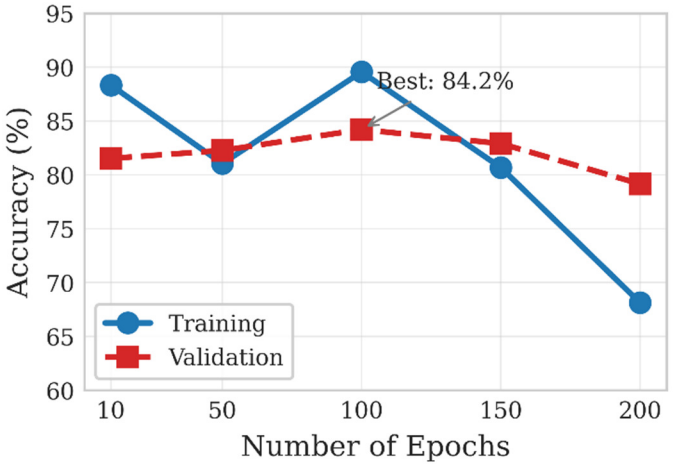

Fig. 2. Training and validation accuracy of SA-FCNNMP over epochs.

Figure 3 shows the training and validation precision of SA-FCNNMP over epochs. Both metrics improved as the number of epochs increased, peaking at epoch 100, where training precision reached 0.90 and validation precision reaches 0.87. This epoch falls within the optimal precision range (0.85–0.90), demonstrating effective learning and generalization.

Figure 4 shows the training and validation recall of SA-FCNNMP over epochs. Initially, both training and validation recall scores were relatively modest, with gradual improvement seen up to epoch 100. At this point, validation recall peaked impressively, approaching 0.90, which indicates the model is particularly effective at identifying correct instances without missing relevant patterns in the data.

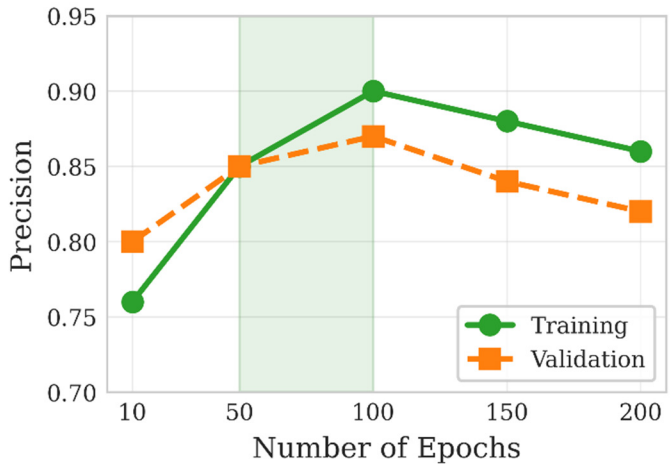

Fig. 3. Training and validation precision of SA-FCNNMP over epochs.

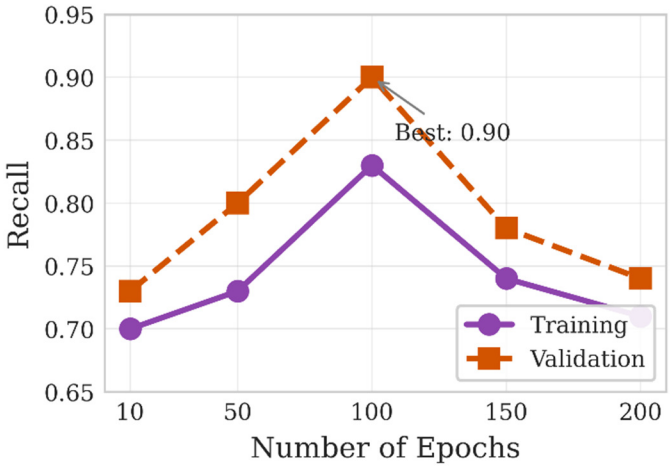

Fig. 4. Training and validation recall of SA-FCNNMP over epochs.

Figure 5 shows the sensitivity of the SA-FCNNMP model over training epochs. Sensitivity improves steadily up to 100 epochs, reaching a stable and strong level. Pushing beyond this point appears to reduce the model's effectiveness, likely due to overfitting or overspecialization on the training data. For sentiment analysis tasks, where detecting subtle emotional cues is essential, maintaining this balance is particularly important. Figure 6 depicts the computational time (in seconds) for different sentiment classification methods. The proposed SA-FCNNMP model demonstrates the lowest computational time among the compared approaches, thanks to its compound scaling strategy, which optimally balances model reliability with minimal resource consumption. The reduced computational time ensures cost-effective and energy-efficient performance, making the proposed model highly suitable for time-sensitive applications.

Figure 7 illustrates the progression of Cross-entropy loss over training epochs. The SA-FCNNMP model exhibits a consistently faster and more stable decrease in cross-entropy loss compared to the others, indicating superior learning and classification capability. The gap in performance becomes increasingly evident after the early epochs, where the baseline models converge more slowly, suggesting that SA-FCNNMP more effectively optimizes classification boundaries in the multimodal sentiment space.

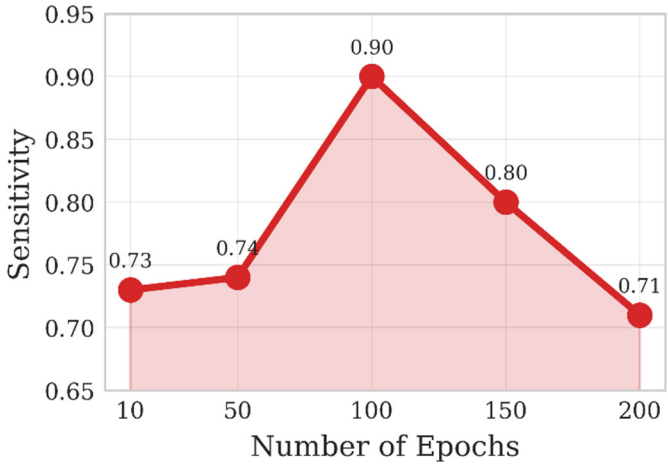

Fig. 5. Sensitivity of SA-FCNNMP over training epochs.

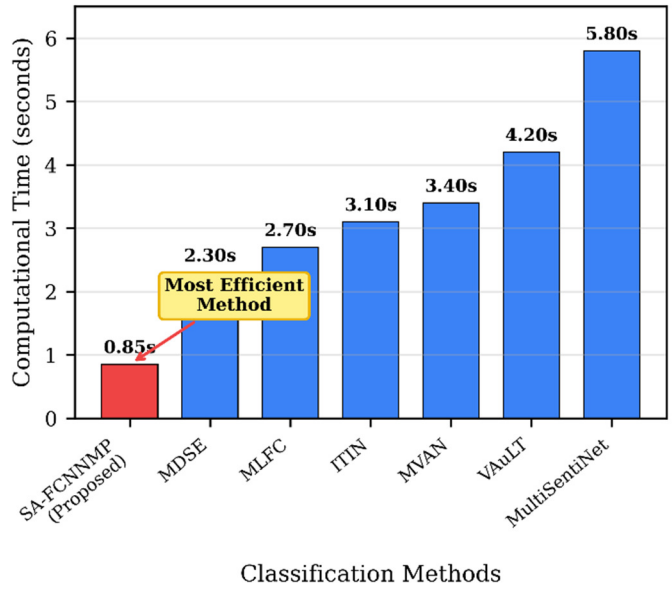

Fig. 6. Computational time of different sentiment classification methods.

Figure 8 presents the MSE loss curves across training epochs. The proposed SA-FCNNMP model achieves significantly lower MSE values throughout the training process, highlighting its enhanced ability to capture subtle differences in sentiment intensity and distribution. While all models show some improvement over time, SA-FCNNMP consistently outperforms the baseline methods, especially in later epochs, reflecting its robustness and precision in modelling finer-grained sentiment nuances in the multimodal Twitter dataset. From Figures 7 and 8, it is noticed that the loss may increase due to overfitting or a distribution shift. Further, an increasing loss over epochs can occur when the model begins to overfit or when it is evaluated on a validation/test set, even though training accuracy improves. However, the problem of overfitting will be addressed in future work.

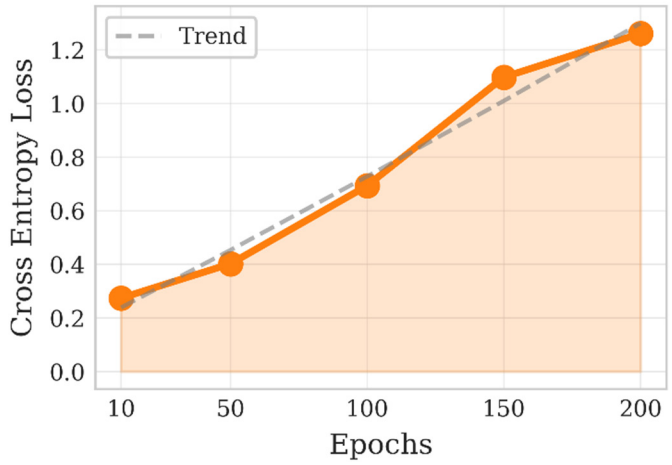

Fig. 7. Cross-entropy loss of SA-FCNNMP over training epochs.

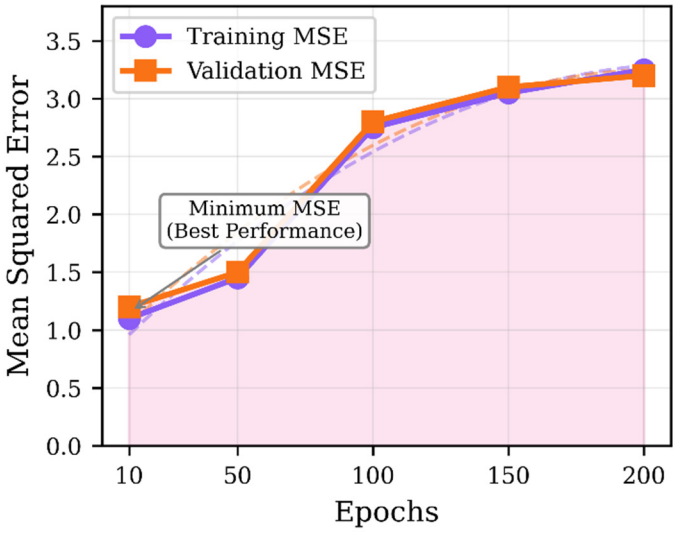

Fig. 8. MSE of SA-FCNNMP over epochs.

C. *Feature Importance Results*

Table II and Figure 9 present a statistical comparison of the proposed SA-FCNNMP method with various existing approaches, evaluating model performance in terms of accuracy. The proposed method consistently outperforms the existing approaches, indicating that the improvements achieved by SA-FCNNMP are statistically significant. These findings demonstrate the effectiveness of the proposed strategy in leveraging multimodal features for sentiment classification using the MVSA-Single dataset.

TABLE II. STATISTICAL COMPARISON OF SA-FCNNMP WITH EXISTING METHODS ON THE MVSA-SINGLE DATASET [44]

Model	Accuracy (%)
MultiSentiNet	69.84
CNN-Multi	61.20
DNN-LR	61.42
VAuLT	72.80
MVAN	72.98
ITIN	75.19
MLFC	75.33
MultiPoint	70.13
MDSE	76.22
CIGNN	75.11
MMJL	76.98
OTE	73.24
CMCA	68.55
HSTEC	71.60
NaiveCat	72.46
Proposed method	84.20

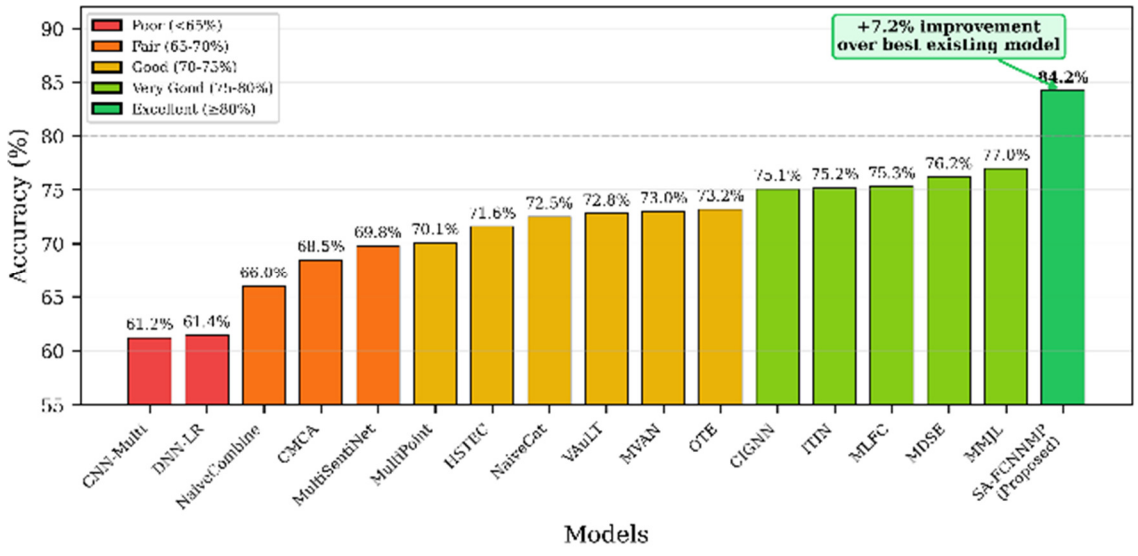

Fig. 9. Statistical comparison of the proposed SA-FCNNMP method versus existing methods.

IV. CONCLUSION AND FUTURE SCOPE

With rapid advancements in multimodal sentiment analysis, the proposed Sentiment Analysis using Fully Connected Neural Network and Multilayer Perceptron (SA-FCNNMP) model demonstrates strong potential for analyzing Twitter data by effectively integrating textual and visual modalities in a unified fusion and classification pipeline. The model achieves improved performance in classifying positive, negative, and neutral sentiments, leveraging advanced techniques such as Bayesian Boundary Trend Filtering (BBTF), Word2Vec embeddings, ResNet-50, and Hierarchical Multi-Scale Feature Fusion (HMSFF). Key challenges observed include intermodal imbalance, noise removal, missing modalities, synchronization issues, and the selection of the optimal fusion strategy.

In the proposed approach, feature extraction is performed separately for text and image modalities. The extracted features are then integrated using an early fusion strategy via HMSFF, followed by classification using a Fully Connected Neural Network and Multilayer Perceptron (FCNNMP) to categorize sentiments as negative, neutral, or positive. The model is evaluated using multiple performance metrics, demonstrating promising results compared to existing state-of-the-art methods.

Several avenues exist for further improvement. The problem of handling overfitting could be a significant issue that can be considered as future work. Incorporating temporal and contextual features could enhance the representation of dynamic sentiment changes. Expanding the sentiment categories to finer-grained or emotion-based classes may enable deeper analysis. Additionally, integrating transformer-based architectures could improve feature extraction and contextual understanding. Addressing challenges such as intermodal imbalance, noise removal, lack of modalities, and synchronization requires further innovative research, especially on the development of more adaptive fusion strategies that can fully exploit the strengths of different modalities. Finally, the experimental findings indicate consistent improvements in accuracy and precision over state-of-the-art methods, highlighting the ability of SA-FCNNMP to capture more complex emotional cues across modalities and rendering it a highly effective tool for real-world sentiment analysis on social media.

REFERENCES

[1] W. Aljedaani *et al.*, "Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of US airline industry," *Knowledge-Based Systems*, vol. 255, Nov. 2022, Art. no. 109780, https://doi.org/10.1016/j.knosys.2022.109780.

[2] M. Bibi *et al.*, "A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis," *Pattern Recognition Letters*, vol. 158, pp. 80–86, June 2022, https://doi.org/10.1016/j.patrec.2022.04.004.

[3] Md. M. Rahman and M. N. Islam, "Exploring the Performance of Ensemble Machine Learning Classifiers for Sentiment Analysis of COVID-19 Tweets," in *Sentimental Analysis and Deep Learning: Proceedings of ICSADL 2021*, Songkhla, Thailand, 2021, pp. 383–396, https://doi.org/10.1007/978-981-16-5157-1_30.

[4] A. Alwehaibi, M. Bikdash, M. Albogmi, and K. Roy, "A study of the performance of embedding methods for Arabic short-text sentiment analysis using deep learning approaches," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, Part B, pp. 6140–6149, Sept. 2022, https://doi.org/10.1016/j.jksuci.2021.07.011.

[5] M. Qorib, T. Oladunni, M. Denis, E. Ososanya, and P. Cotae, "Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset," *Expert Systems with Applications*, vol. 212, Feb. 2023, Art. no. 118715, https://doi.org/10.1016/j.eswa.2022.118715.

[6] N. Aslam, F. Rustam, E. Lee, P. B. Washington, and I. Ashraf, "Sentiment Analysis and Emotion Detection on Cryptocurrency Related Tweets Using Ensemble LSTM-GRU Model," *IEEE Access*, vol. 10, pp. 39313–39324, 2022, https://doi.org/10.1109/ACCESS.2022.3165621.

[7] Md. A. Babu, M. Ahammad, M. Mahmud, and Md. S. Uddin, "Social Media as a Market Prophecy: Leveraging ML Algorithms for Predicting Market Trends and Demand," *Transportation Research Procedia*, vol. 84, pp. 137–144, Jan. 2025, https://doi.org/10.1016/j.trpro.2025.03.056.

[8] S. Tabinda Kokab, S. Asghar, and S. Naz, "Transformer-based deep learning models for the sentiment analysis of social media data," *Array*, vol. 14, July 2022, Art. no. 100157, https://doi.org/10.1016/j.array.2022.100157.

[9] G. Kaur and A. Sharma, "A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis," *Journal of Big Data*, vol. 10, no. 1, Jan. 2023, Art. no. 5, https://doi.org/10.1186/s40537-022-00680-6.

[10] E. Lee, F. Rustam, P. B. Washington, F. E. Barakaz, W. Aljedaani, and I. Ashraf, "Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets Using Stacked Ensemble GCR-NN Model," *IEEE Access*, vol. 10, pp. 9717–9728, 2022, https://doi.org/10.1109/ACCESS.2022.3144266.

[11] K. Pasupa and T. Seneewong Na Ayutthaya, "Hybrid Deep Learning Models for Thai Sentiment Analysis," *Cognitive Computation*, vol. 14, no. 1, pp. 167–193, Jan. 2022, https://doi.org/10.1007/s12559-020-09770-0.

[12] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network," *IEEE Access*, vol. 10, pp. 21517–21525, 2022, https://doi.org/10.1109/ACCESS.2022.3152828.

[13] A. P. Rodrigues and N. N. Chiplunkar, "A new big data approach for topic classification and sentiment analysis of Twitter data," *Evolutionary Intelligence*, vol. 15, no. 2, pp. 877–887, June 2022, https://doi.org/10.1007/s12065-019-00236-3.

[14] I. E. Fattoh, F. Kamal Alsheref, W. M. Ead, and A. M. Youssef, "Semantic Sentiment Classification for COVID-19 Tweets Using Universal Sentence Encoder," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, Oct. 2022, Art. no. 6354543, https://doi.org/10.1155/2022/6354543.

[15] T. H. Jaya Hidayat, Y. Ruldeviyani, A. R. Aditama, G. R. Madya, A. W. Nugraha, and M. W. Adisaputra, "Sentiment analysis of twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier," *Procedia Computer Science*, vol. 197, pp. 660–667, Jan. 2022, https://doi.org/10.1016/j.procs.2021.12.187.

[16] F. K. Sufi and I. Khalil, "Automated Disaster Monitoring From Social Media Posts Using AI-Based Location Intelligence and Sentiment Analysis," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 4, pp. 4614–4624, Aug. 2024, https://doi.org/10.1109/TCSS.2022.3157142.

[17] A. Sarirete, "Sentiment analysis tracking of COVID-19 vaccine through tweets," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 11, pp. 14661–14669, Nov. 2023, https://doi.org/10.1007/s12652-022-03805-0.

[18] L. G. Singh and S. R. Singh, "Sentiment analysis of tweets using text and graph multi-views learning," *Knowledge and Information Systems*, vol. 66, no. 5, pp. 2965–2985, May 2024, https://doi.org/10.1007/s10115-023-02053-8.

[19] P. Rakshit and A. Sarkar, "A supervised deep learning-based sentiment analysis by the implementation of Word2Vec and GloVe Embedding techniques," *Multimedia Tools and Applications*, vol. 84, no. 2, pp. 979–1012, Jan. 2025, https://doi.org/10.1007/s11042-024-19045-7.

[20] S. Zhang, Y. He, L. Li, and Y. Dou, "Multimodal sentiment analysis with BERT-ResNet50," in *Second International Conference on Algorithms, Microchips, and Network Applications*, Zhengzhou, China, 2023, vol. 12635, pp. 232–236, https://doi.org/10.1117/12.2679113.

[21] Y. Li, H. Ding, Y. Lin, X. Feng, and L. Chang, "Multi-level textual-visual alignment and fusion network for multimodal aspect-based sentiment analysis," *Artificial Intelligence Review*, vol. 57, no. 4, Mar. 2024, Art. no. 78, https://doi.org/10.1007/s10462-023-10685-z.

[22] Y. Qi and Z. Shabrina, "Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach," *Social Network Analysis and Mining*, vol. 13, no. 1, Feb. 2023, Art. no. 31, https://doi.org/10.1007/s13278-023-01030-x.

[23] T. Niu, S. Zhu, L. Pang, and A. El Saddik, "Sentiment Analysis on Multi-View Social Data," in *22nd International Conference on Multimedia Modeling*, Miami, FL, USA, 2016, pp. 15–27, https://doi.org/10.1007/978-3-319-27674-8_2.

[24] N. Xu and W. Mao, "MultiSentiNet: A Deep Semantic Network for Multimodal Sentiment Analysis," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, Singapore, 2017, pp. 2399–2402, https://doi.org/10.1145/3132847.3133142.

[25] A. Alsayat, "Improving Sentiment Analysis for Social Media Applications Using an Ensemble Deep Learning Language Model," *Arabian Journal for Science and Engineering*, vol. 47, no. 2, pp. 2499–2511, Feb. 2022, https://doi.org/10.1007/s13369-021-06227-w.

[26] X. Yang, S. Feng, D. Wang, and Y. Zhang, "Image-Text Multimodal Emotion Classification via Multi-View Attentional Network," *IEEE Transactions on Multimedia*, vol. 23, pp. 4014–4026, 2021, https://doi.org/10.1109/TMM.2020.3035277.

[27] T. Zhu, L. Li, J. Yang, S. Zhao, H. Liu, and J. Qian, "Multimodal Sentiment Analysis With Image-Text Interaction Network," *IEEE Transactions on Multimedia*, vol. 25, pp. 3375–3385, 2023, https://doi.org/10.1109/TMM.2022.3160060.

[28] M. A. M. Hamidi, A. Y. Taqa, and Y. I. Ibrahim, "A Systematic Review of Multimodal Sentiment Analysis Based on Text-Image Fusion: Trends, Models, and Research Gaps," *Sinkron : jurnal dan penelitian teknik informatika*, vol. 9, no. 2, pp. 987–999, Apr. 2025, https://doi.org/10.33395/sinkron.v9i2.14840.

[29] K. Wang and Y. Zhang, "Topic Sentiment Analysis in Online Learning Community from College Students," *Journal of Data and Information Science*, vol. 5, no. 2, pp. 33–61, May 2020, https://doi.org/10.2478/jdis-2020-0009.

[30] B. Bhavana, C. Chaitanya, and B. M. G, "Multimodal Question Answering with DenseNet and BERT for Improved User Interaction," in *2025 Fourth International Conference on Smart Technologies, Communication and Robotics*, Sathyamangalam, India, 2025, pp. 1–7, https://doi.org/10.1109/STCR62650.2025.11019326.

[31] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, and X. Luo, "TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis," *Pattern Recognition*, vol. 136, Apr. 2023, Art. no. 109259, https://doi.org/10.1016/j.patcog.2022.109259.

[32] Z. Liu, L. Cai, W. Yang, and J. Liu, "Sentiment analysis based on text information enhancement and multimodal feature fusion," *Pattern Recognition*, vol. 156, Dec. 2024, Art. no. 110847, https://doi.org/10.1016/j.patcog.2024.110847.

[33] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, "Multimodal sentiment analysis based on fusion methods: A survey," *Information Fusion*, vol. 95, pp. 306–325, July 2023, https://doi.org/10.1016/j.inffus.2023.02.028.

[34] W. Zou, J. Ding, and C. Wang, "Utilizing BERT Intermediate Layers for Multimodal Sentiment Analysis," in *2022 IEEE International Conference on Multimedia and Expo*, Taipei, Taiwan, 2022, pp. 1–6, https://doi.org/10.1109/ICME52920.2022.9860014.

[35] J. Li *et al.*, "Hybrid Multimodal Feature Extraction, Mining and Fusion for Sentiment Analysis," in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, Lisbon, Portugal, 2022, pp. 81–88, https://doi.org/10.1145/3551876.3554809.

[36] K. L. Tan, C. P. Lee, K. M. Lim, and K. S. M. Anbananthen, "Sentiment Analysis With Ensemble Hybrid Deep Learning Model," *IEEE Access*, vol. 10, pp. 103694–103704, 2022, https://doi.org/10.1109/ACCESS.2022.3210182.

[37] S. Kumari and M. P. Singh, "A Deep Learning Multimodal Framework for Fake News Detection," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 16527–16533, Oct. 2024, https://doi.org/10.48084/etasr.8170.

[38] T. Onizuka, F. Iwashige, and Shintaro Hashimoto, "Bayesian boundary trend filtering," *Computational Statistics & Data Analysis*, vol. 191, Mar. 2024, Art. no. 107889, https://doi.org/10.1016/j.csda.2023.107889.

[39] M. Grohe, "word2vec, node2vec, graph2vec, X2vec: Towards a Theory of Vector Embeddings of Structured Data," in *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, Portland, OR, USA, 2020, pp. 1–16, https://doi.org/10.1145/3375395.3387641.

[40] B. Koonce, "ResNet 50," in *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*, Berkeley, CA, USA: Apress, 2021, pp. 63–72, https://doi.org/10.1007/978-1-4842-6168-2_6.

[41] G. R. Kishore, B. S. Harish, and C. K. Roopa, "Unenhanced Sparse Vector-based Embedding Method for Sentiment Analysis," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 21225–21231, Apr. 2025, https://doi.org/10.48084/etasr.10098.

[42] X. Huo *et al.*, "HiFuse: Hierarchical multi-scale feature fusion network for medical image classification," *Biomedical Signal Processing and Control*, vol. 87, Jan. 2024, Art. no. 105534, https://doi.org/10.1016/j.bspc.2023.105534.

[43] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (Technology) is Power: A Critical Survey of 'Bias' in NLP," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020, pp. 5454–5476, https://doi.org/10.18653/v1/2020.acl-main.485.

[44] L. Gong, X. He, and J. Yang, "An Image-Text Sentiment Analysis Method Using Multi-Channel Multi-Modal Joint Learning," *Applied Artificial Intelligence*, vol. 38, no. 1, Dec. 2024, Art. no. 2371712, https://doi.org/10.1080/08839514.2024.2371712.

AUTHORS PROFILE

T. S. Kaveri received the B.E. degree in Information Science and Engineering from Visvesvaraya Technological University (VTU), and an M.Tech. degree in Data Science and Engineering from JSS Science and Technology University, Mysuru, Karnataka, India. She is currently working as an Assistant Professor and a Researcher in the Department of Information Science and Engineering at JSS Science and Technology University. Her research interests include Machine Learning, Artificial Intelligence, and Text Mining.

B. S. Harish received the Ph.D. degree in Computer Science from the University of Mysore, India. He is currently a Professor in the Department of Information Science and Engineering, JSS Science and Technology University, India. He was a Visiting Researcher with DIBRIS—Department of Informatics, Bio Engineering, Robotics and System Engineering, University of Genoa, Italy. He has been invited as a resource person to deliver various technical talks on data mining, image processing, pattern recognition, and soft computing. He is also serving and served as a reviewer for international conferences and journals. He has published articles in more than 100 international reputed peer-reviewed journals and conferences proceedings. He successfully executed AICTERPS Project, which was sanctioned by AICTE, Government of India. His research interests include Machine Learning, Text Mining, and Computational Intelligence.

C. K. Roopa received the B.E. degree in Information Science and Engineering and the M.Tech. degree in Computer Engineering from Visvesvaraya Technological University, Belagavi, Karnataka, India, and the Ph.D. degree from the University of Mysore, India. She is currently an Associate Professor in JSS Science and Technology University. She is serving as a reviewer for many conferences and journals. She is a life-time member of ISTE and CSI. Her research interests include Medical Image Analysis, Biometrics, and Text Mining.

M.S. Kendagannaswamy received his B.E. degree in Computer Science and Technology from Visvesvaraya Technological University (VTU), and his M.Tech. degree in Software Engineering from JSS Science and Technology University, Mysuru, Karnataka, India. He currently serves as an Assistant Professor at JSS Science and Technology University. Previously, he worked as a Researcher in the Department of Information Science and Engineering at JSS STU. His research interests include Machine Learning and Artificial Intelligence Applications in Remote Sensing Analysis.