Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Transforming Multimodal Sentiment Analysis and Classification with Fusion-Centric Deep Learning Techniques

Vinitha V[1], Dr. S. K. Manju Bargavi[2]
[1]*Research scholar, Department of Computer science and IT, Jain University, Karnataka, India, vinithafive@gmail.com*
[2]*Professor, Department of Computer science and IT, Jain University, Karnataka, India, b.manju@jainuniversity.ac.in*

ARTICLE INFO

Received: 17 Oct 2024

Revised: 15 Dec 2024

Accepted: 22 Dec 2024

ABSTRACT

Multimodal Sentiment Analysis (MSA) has has become an important field of research, integrating information since text, visuals, video, and speech modalities to derive thorough physiological insights. Despite substantial advancements, current methodologies frequently regard various modalities uniformly, neglecting the preeminent impact of text during sentiment analysis and ignoring address of redundant and irrelevant data generated during multimodal fusion. This study proposes the Enhanced Multi-modal spatiotemporal attention network (EMSAN), to integrate key features across modalities designed to develop the robustness and generalization of sentiment and emotion predictionfrom video data. It consists various phases such as multimodal feature extraction, fusion, and detection of sentiment polarityto integrate key features across modalities. Extensive experiments carried out on the publicly available Multimodal Emotion Lines Dataset (MELD) that show that the suggested method performs with an accuracy of 92.28% in capturing complicated sentiment and emotion. The comparison showed that the suggested method worked better than other baseline models, which made it possible to develop sentiment analysis in a number of different multimodal frameworks.

Keywords: Sentiment analysis, attention mechanism, natural language processing, artificial intelligence, deep learning

1. INTRODUCTION

Since the advent of digital insights, multimodal sentiment analysis (MSA) technology has been essential in the fast-expanding field of emotional computation for evaluating complex human emotions and opinions. In order to obtain integrated sentiment, MSA seeks to integrate disparate data sources and modalities, including text, audio, and video [1]. This cross-field evolution manages acoustic signal processing with computer vision and natural language processing technique. Its applications range from psychology, to social media monitoring for companies, to marketing analytics, to human-computer interaction studies [2]. In addition, automatic and consistent emotion recognition in multimodal data plays an important role in the interaction and customizing the content since it gives insightful knowledge about the human emotive dynamics in the digitalization age [3]. MSA is believed to be significant in human emotion to interpret since it considers textual, auditory, and visual facets of information. Thereby this multisensory approach enables the extraction of sentiment across each modality [4]. Machine learning (ML) is the ability of a machine or a program to increase its proficiency at completing certain tasks through exposure to data and experiences. Supervised, semi-supervised, and unsupervised machine learning are the three categories. Such techniques can automate and handle very large amounts of data, making them a suitable fit for SA [5].ML models can generate a variety of classifiers that address the challenges of removing information and improving polarity accuracy and classification, as well as performing effective sophisticated models from one modality or multiple modality data, such as audio, video, and textual information for sentiment analysis [6].

Many inherent complications in the realm of MSA hinder its full potential in real-world applications. The intricacy of joining and processing data from various modalities, each with distinct traits and cues, is the main cause of these complications. For example, the processing of textual data necessitates complex semantic analysis, while visual data and audio require the examination of spatial and temporal patterns, respectively [7]. In the same way, data alignment is an important experiment. Modalities don't always line up perfectly over time, which makes it hard to get synchronized and coherent multimodal features.

Real-world constraints, including the discrepancy amongst modalities or the existence of incomplete or missing data in one or additional modalities, extend this problem even more. These features propose technical experiments and may lead to less-than-optimal sentiment analysis results, misrepresenting or ignoring the complications and nuances of human emotions [8]. The proposed approach uses aenhanced multi-modal spatiotemporal attention network technique to pre-process various input data types, such as text, audio, and video, through a number of data pre-processing steps to prepare it for additional processing. The objectives of this study consists of:

- The Enhanced Multi-Modal Spatiotemporal Attention Network is employed to develop a comprehensive technique for multimodal sentiment categorization that focuses on attention mechanisms.
- Combining different modalities enhances the model's comprehension of context and visual cues, to apply specific pre-processing methods that are suited to every modality, improving the input data's quality and applicability.
- The effectiveness of EMSAN is demonstrated by extensive evaluations utilizing the publicly available Interactive MELD data set. The experimental results show how effectively the technique captures complicated sentiment patterns across multiple domains.

2. RELATED WORK

Multimodal sentiment analysis attracted a lot of interest because it can comprehensively detect emotions by integrating text, audio, and visual data. The fusion of these modalities to enhance accuracy is the primary focus of recent advancements, as opposed to the isolation of early approaches. Methods such as attention mechanisms, Fusion at the level of decisions and feature-level fusion have been explored to enhance the capture of correlations among modalities. It was combined with a new type of interaction network [9] using an effective classification technique. Several studies have used the encoder layer of the transformer to generate high-level sentimental semantic encodings from audio and textual series. Then, a multimodal attention mechanism with bimodal feature interaction fusionconsiders intra- and inter-modal correlation (context dependency) has been proposed. This allows the model to better process and infer affective semantics. This sentiment classification is achieved through the inclusion of the fused features into the classification layer. Emotion recognition and attention-based multimodal sentiment analysis is a distinctive architecture [10]. These architectures dictate how much information can be passed between modalities and how much discriminating information can be streamed within modalities, i.e. in text, audio, and visual contexts. Based on these methods, a more sophisticated hierarchic algorithm can then be obtained with the intermediated fusion so as to find hierarchic relations about the situations at the trimodal, bimodal levels. Ultimately, this system integrates four different methods at the decision level to make this system capable of multimodal SA and emotion recognition.A Deep Emotional Arousal Network (DEAN) that incorporates the temporal dependency into the transformer's parallel design and may replicate emotion coherence is introduced [11]. Three mechanisms constitute the DEAN method that has been suggested. For instance, a multi-modal Bi-LSTM process has been enhanced to mimic using a multi-modal gating block to replicate the activation mechanism in the human emotion arousal technique, the cognitive comparator has been developed to cross-modal transformer can be developed to mimic the functions of the perceptive investigation human method. An unique hybrid deep learning neural network architecture utilizing enhanced text representations, termed Fasttext CNN with LSTM [12].

Deep learning framework can handle straight way of getting rid of features directly from the text by using CNN, which has multiple convolutional layers for this approach, to get best possible results [13]. The similar technology for word embedding that based on n-grams is applied to extract a machine-level representation of each word. model level that fuses feature levels with pre-trained models based on emoticons images text inputs [14].In these algorithms compared to traditional sentiment analysis is based on variables like emoticon information and image properties to find more sensitive variations in emotions like hyper positive, hyper negative and neutral. This approach is also used in both scenarios, where feature fusion models mix up features for images and text to give a more extensive representation. a transformer-based encoder-decoder translation network with a multimodal extension, following a unified encoder-decoder model where text is the main data type and text, images, and audio provide secondary information. In order to improve quality and facilitate multi-modal feature fusion, a modalities reinforcement cross-attention unit for the weighted fusion of different modal data has been designed to mitigating the adverse impact on non-natural language data. Additionally, the dynamic filter mechanism isolates the informational errors occurring in cross-modal communication to improve subsequent output [15].

3. METHODOLOGY

This study specifically concentrates on the architecture and formulation of the Enhanced multi-modal spatiotemporal attention framework, which is intent on performing sentiment analysis by utilizing multimodal methods. It consists of various techniques such as parameters tuning, classification, multimodal fusion extraction, preprocessing and predicting sentiment as indicated in Figure 1.

Figure 1. Block diagram of Enhanced Multi-modal spatiotemporal attention network

More and more people are sharing their opinions online via video rather than text. Consequently, multimodal sentiment analysis (MSA) applying several modalities has become an essential subject of research. In this work, Multi modal sentiment analysis achieves higher performance and reduces the error rates by means of deep learning approaches in several step to processes including sentiment polarity identification and multimodal feature extraction and fusion. This paper employs a multi-modal deep learning model for automated sentiment and emotion classification utilizing text, audio, and video data.

Pre-processing: Initially, the preprocess the diverse input Text, audio, and video data are all carried out through various phases of data preprocessing to ensure compatibility with subsequent processing. Preprocessing is an essential phase of video analysis, as it ensures that the data utilized in deep learning methods is clear, consistent, and optimized for performance. The extraction of frames is the initial step in the procedurewhich involves the separation of individual frames from the video. These frames are then translated into a series of images for analysis. The video frames are subsequently enhanced with a Gaussian-adaptive bilateral filter, which dynamically adjusts

the spatial and range weights using a Gaussian function. Then, the frames are resized to a standard dimension of 224x224 pixels, which is compatible with the input requirements of a variety of deep learning techniques. Gaussian Adaptive Bilateral Filtering (GABF): is applied to smooth frames while preserving edges as per Eq. (1) :

$$I'(x,y) = \frac{1}{W}\sum_{p\in\Omega} I(p) \cdot G_s(||p-(x,y)||) \cdot G_r(|I(p)-I(x,y)|)(1)$$

where:

- G_s is the spatial weight
- G_r is the range weight

The removal of noise is a critical step in audio processing to ensure that the signals being analysed are as precise and distinct as possible. The Kalman Filter (KF) is a recursive state-space estimation algorithm that is used to denoise, smooth, and predict signals in time-series data. It is one effective model for this purpose. It is effective in real-time filtering, noise reduction, and speech enhancement during audio preprocessing as per Eq. (2)

$$x_k = Ax_{k-1} + w_k, y_k = Hx_k + v_k \quad (2)$$

where:

- x_k is the hidden clean signal,
- y_k is the noisy observed signal,
- A is the state transition matrix,
- w_k, v_k are noise terms.

Text pre-processing is an essential phase in the preparation of textual data for analysis in NLP tasks. The process begins with reduction expansion, which involves the cleaning, tokenization, and lemmatization of text data. The importance of preprocessing is substantiated by the following: Social media text types may exhibit a variety of grammatical and semantic issues, noise, and differences in style due to factors such as size, informal language, and typing speed. (2) Trend detection for classifiers is facilitated by data standardization. (3) Input requirements for Word Embedding layers and other classifiers must be satisfied by the texts. The Natural Language Toolkit (NLTK) is used for completing this process as per Eq. (3)

Tokenization: Split text into words/tokens

T={t1,t2,...,tn} (3)

Lemmatization: Convert words to base form:

$$ti' = Lemma(ti)$$

where ti′ is the lemmatized word.

- **Multimodal Feature Extraction**

  Next, multimodal fusion approaches can be used to merge high-level text, audio, and video information while applying the proposed attention strategy to accurately and efficiently categorize emotions and sentiments. In video processing, the self-attention mechanism employed by vision transformers (ViTs) varies from the convolution used in CNNs in that it evaluates the significance of distinct patches or segments of an image based on their interrelations. This device is used to detect global dependencies between image patches, allowing the technique to extract complicated characteristics and interactions between them.

An assumed matrix $X \in \mathbb{R}^{n\times m}$ containing n rows and m columns while every row is realized for a fattened image patch and is divided into 3 equal copies. All of these 3 matrices are correspondingly multiplied by an early set of weights $W_Q \in \mathbb{R}^{m\times d}, W_K \in \mathbb{R}^{m\times d}$ and $W_V \in \mathbb{R}^{m\times d}$, whereas d refers to an embedding size, to provide a query $(Q = W_Q X \in \mathbb{R}^{n\times d})$, key $(K = W_K X \in \mathbb{R}^{n\times d})$ and value $(V = W_V X \in \mathbb{R}^{n\times d})$ matrix. In the subsequent step, an attention matrix is computed as per Eq. (4) as

$$A(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

The softmax transformation of $Z = \frac{QK^T}{\sqrt{d_k}} \in \mathbb{R}^{n\times n}$The weights of attention that show which set of image patches (tokens)in the sequence interact with one another are represented by the matrix. Complex multi-head structures can integrate different visual segments essential to classification tasks by using self-attention as a base. This eliminates the requirement for explicit feature engineering. This makes them more adaptable to different kinds of images.

In audio analysis, autoencoder-decoder architecture and feature extraction are essential for obtaining meaningful sound data representations. Raw audio waveforms are converted into numerical representations by feature extraction so that machine learning models can process them. Zero Crossing Rate (ZCR), which calculates the frequency at which the signal crosses zero amplitude, and Root Mean Square (RMS) energy, which measures signal intensity, are examples of common time-domain features. Spectrograms, the Short-Time Fourier Transform (STFT), and Mel-Frequency Cepstral Coefficients (MFCCs) are the examples of frequency-domain features that aid in capturing the frequency components of audio signals.These architectures are very helpful for learning high-level audio representations, anomaly detection, and denoising. This procedure is further improved by Variational Autoencoders (VAEs) and Convolutional Autoencoders (CAEs), which produce significant latent representations.This technique enables the extraction of rich, multidimensional characteristics, which are crucial for accurate audio detection and classification.The audıosıgnal is encoded into latent variablesas per Eq. (5)and Eq. (6) as:

$$z = \mu + \sigma \cdot \epsilon, \epsilon \sim \mathcal{N}(0,1) \quad (5)$$

where μ and σ are learned mean and variance. The reconstructed feature:

$$A_f = \text{Decoder}(z) \quad (6)$$

Both high-level and low-level feature extraction are used in text processing to capture each aspect of language. By breaking down text into sub-word units, XLNet tokenization makes it easier to extract features at a fine level while maintaining context and meaning—two crucial components for understanding intricate language systems [20]. A cutting-edge pre-trained model for natural language understanding, XLNet (Extreme Multi-head Attention Network) was developed as an extension of the Transformer architecture. With the use of this autoregressive technique, XLNet can understand bidirectional context without the constraints of masked language modeling, allowing for more sophisticated text analysis and increasing performance in jobs like sentiment analysis.

natural language comprehension, and text categorizationas per Eq. (7).

XLNet is used to extract context-aware features:

$$T_f = \text{XLNet}(T) \quad (7)$$

where T_f is sequence length

- **Multi-Modal Fusion Strategies:** Fusion methods are required in Multimodal Sentiment Analysis (MSA) to integrate features from many modalities (e.g., text, audio, and visual) and increase sentiment prediction precision. Many fusion procedures can be separated into three groups: hybrid, late, and early fusion. Each alternative has various advantages and can be chosen based on the objective, the data features, and the architecture used Using a Multi-Modal Spatiotemporal Attention Network (MSTAN) to highlight salient features across modalities. Feature representations from various modalities are supplied into spatiotemporal attention layers as per Equations (8), (9), (10), and (11) as follows.
- Spatial Attention (Frame-Wise Focus on Key Regions)
- Compute spatial attention weights:

$$W_s = \text{Softmax}\left(\text{Conv}_{1\times1}(V_f)\right) \quad (8)$$

- Applying attention:

$$F_s' = W_s \odot V_f \quad (9)$$

Temporal Attention (Focus on Important Time Steps)

- Attention score computation:

$$\alpha_t = \text{Softmax}(W_t \cdot H_t) \quad (10)$$

- Weighted sum over time:

$$F_t' = \sum_{t=1}^{T} \alpha_t H_t \quad (11)$$

Additionally, the MSTAN approach classifies and identifies sentiment using Dense net 201. A deep learning architecture for convolutional neural networks is called DenseNet, or Dense Convolutional Network. By establishing a feed-forward connection between each layer and every other layer, it signifies a paradigm shift. DenseNetguarantees that each layer receives input from all preceding layers and outputs to all following layers, in contrast to conventional CNNs, which have a single link between successive layers. A deep learning model that works well for classification. Dense connections improve gradient flow and encourage feature reuse. Perfect for a range of classification issues and datasets. Lastly, the MSTAN technique's hyperparameters were adjusted using Adam Optimizer-based hyperparameter tuning, which produced sentiment and emotion predictions.The feature vector is passed through DenseNet-201 for classification as per Eq. (11), Eq. (13) stated below:

$$F_{\text{final}} = \text{DenseNet}\left(\left[F_s', F_t', T_f, A_f\right]\right) \quad (12)$$

Apply fully connected layers followed by softmax activation:

$$y = \text{Softmax}(W F_{\text{final}} + b) \quad (13)$$

where y is the sentiment class prediction.

4. RESULTS AND DISCUSSION

This section focuses mostly on details of implementation and evaluation measures prior to the to the compilation of experimental data. The performance of the suggested model is then compared to that of the comparator approaches. We then conduct additional test to validate the model's effectiveness.

A. Dataset description: In this work, the stimulation validation of Enhanced Multi-Modal Spatiotemporal Attention Network has been tested utilizing the dataset. Derived from EmotionLines dataset, Multimodal Emotion Lines Dataset (MELD) sorts emotions and generates emotion-cause pairings. Since MELD focusses more on conversations, it is valuable for real-time dialogue applications like customer service chatbots and healthcare virtual agents. MELD includes text, audio, and visual facial expressions among almost 13,000 utterances covering many emotions. Each utterance is accompanied by sentiment annotation (positive, negative, and neutral).

B. Experimental setup: In this study, we used the MELD dataset to train, test, and validate the classifier. The remaining 30% of the data was utilized for testing, and the remaining 70% was used for training. a system running 16 GB of RAM, a 2.4 GHz Intel(R) processor and an Nvidia P100 GPU was used for the study.

C. Comparative study with various Baseline models: In order to precisely assess the methodology's efficacy, one can contrast it with the most cutting-edge techniques in the field. GLFN, BERT-based Multi-semantic Learning (BERT-MSL), Relational Graph Convolutional Network (R-GCN), Average Pooling-Bidirectional Encoder Representations from Transformer (BERT), and Bidirectional Gated Recurrent Unit (Bi-GRU) were all evaluated.

- Text-centered Fusion Network with cross-modal Attention (TeFNA) [16]: TeFNA integrates cross-modal attention and mutual information toward effectively model unaligned, multimodal temporal data.

- Oppositional Grass Bee optimization (OGBEE)[17]: This study uses image, text, audio, and video modalities to investigate sentiments that have been retrieved from web recordings. A feature-level fusion technique is used to integrate features gathered from multiple modalities.

- Hierarchical Self-Attention Fusion-Contextual Self Attention Temporal Convolutional Network Multi Branch Memory(H-SATF-CSAT-TCN-MBM) [18]: Multi-modal information was fused, and multi-branch memory systems were used. In order to increase the model's effectiveness in long-term activities, the self-Attention Fusion framework (H-SATF) and the CSAT-TCN gathered both internal and external correlates of multi-modal information. As shown below in Table 1.

Comparative Methods

Table1. Accuracy Comparison on various Models

Various Techniques	**Accuracy**
TEFNA	87.34%
OGBEE	76.23%
H-SATF-CSAT-TCN-MBM	90.33%
PROPOSED MSTAN	92.28%

D. Evaluation metrics: The overall classifying efficiency was evaluated using a number of evaluation metrics. Various evaluation criteria were evaluated to examine the sentiments using recall, accuracy, and precision,F1 score. The following criteria were used to assess the efficacy of the system:

- **Accuracy:**Thisscore assesses the overall prediction accuracy of the model. The total number of samples in the data set determines the percentage of correctly identified samples.

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

- **Precision:**It is capable of predicting positive sentiments, as it is accurately predicted. The ratio of true positives to the true positive rate compares all other true positives and false positives.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- ***Recall*:**The measure of its ability to identify each positive attitude in the data. The ratio represents the proportion of true positives to the sum of true positives and true negatives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score:**It is a unique metric that combines precision and recall, thereby capturing both properties. It is particularly beneficial when dealing with an imbalanced dataset in which one class is significantly more prevalent than the other.

$$\text{F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The Figure 2 shows the Classification report for various evaluation metrics consisting of precisions, recall,F1 score which indicates that this technique had an effective outcome and obtained an accuracy of 92.28%. The Figure 3. shows the comparison metrics for various performance measures. The results indicated that this technique had an effective outcome compared with other baseline models.

Classification Report:

	precision	recall	f1-score	support
0	0.81	0.87	0.84	255
1	0.74	0.70	0.72	266
2	0.78	0.83	0.81	243
3	0.83	0.74	0.78	262
4	0.74	0.74	0.74	247
5	0.70	0.74	0.72	263
6	0.82	0.79	0.80	264
macro avg	0.77	0.77	0.77	1800
weighted avg	0.77	0.77	0.77	1800

Test accuracy: 0.9228

Figure 2: Classification Report of MSTAN

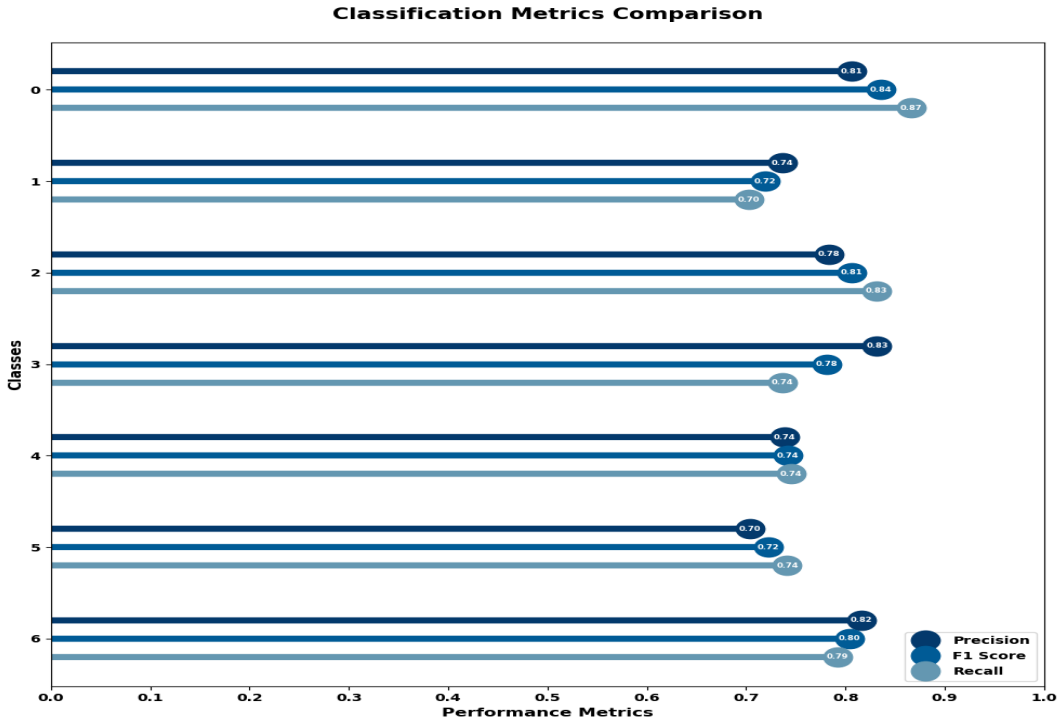

Figure 3: Comparison of various classification metrics

E. Confusion matrix Heatmap: It shows how well the model has classified different with emotional states (like "happy," "sad," or "neutral") based on the actual and predicted labels. The confusion matrix heatmap adds a visual layer, where the intensity of each cell represents how many times a particular combination of true and predicted labels occurred. The below figure 4 illustrates the confusion matrix heatmap, exhibiting the accurate categorization and identification of each class label. The higher the number of correct predictions, the darker the diagonal cells (representing true positives).

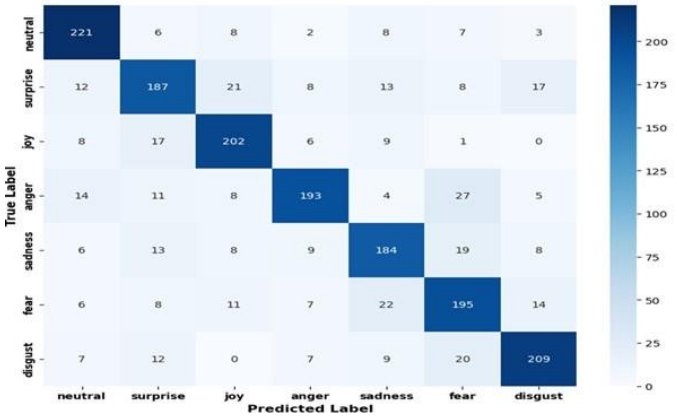

Figure 4. illustrates the confusion matrix heatmap.

Ablation experiments were conducted informally, during which we experimented with a variety of loss functions. In the case of a high learning rate, a lower number of epochs was necessary, and the reverse was also true. Optimal outcomes were achieved by decreasing the learning rate to le-4 and increasing the number of epochs to 50 from the default. For 50 epochs, the model maintained its maximum level of accuracy 92.28% as illustrated in Figure 5. The accuracy results of Training and Validation accuracy exhibit an upward trend, which suggests that the MSTAN technique system has the ability to alter performance across various iteration counts. In addition, the Training and Validation accuracy remains more consistent throughout the epochs, which implies that the EMSAN technique has superior performance and less overfitting. This guarantees that hidden instances will be predicted uniformly.

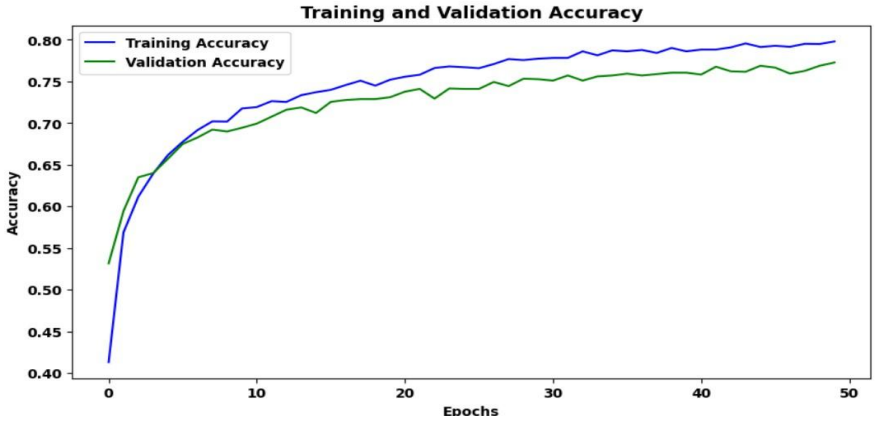

Figure 5: Accuracy curve for video using MSTAN technique

5.CONCLUSION

This study mostly examines the formulation and design of the various EMSAN technique which is used to identify and categorize sentiments through the use of multimodal methods. In order to accomplish this, technique which is implemented to preprocess the diverse Input data, comprising text, audio, and video, undergoes multiple stages of preprocessing to guarantee compatibility with subsequent processing. The suggested model attention method facilitates the integration of high-level text, audio, and video information through multimodal fusion techniques, leading to precise and successful classification of emotions and sentiments. Moreover, the MSTAN methodology employs variants of CNN method to precisely identify and detect sentiment and obtained an accuracy of 92.28%. The hyperparameter optimization was ultimately performed with the Adam Optimizer to adjust the hyperparameters of this methodology. The results are assessed based on several characteristics, and extensive experimental analysis is conducted. The comparison investigation revealed that the MSTAN technique is superior compared Thus, to different baseline approaches so enabling the development of sentiment analysis applications within a variety of multimodal frameworks.

REFERENCES

[1] Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion, 91, 424-444.

[2] Liu, Z., Zhou, B., Chu, D., Sun, Y., & Meng, L. (2024). Modality translation-based multimodal sentiment analysis under uncertain missing modalities. Information Fusion, 101, 101973.

[3] Yadav, A., & Vishwakarma, D. K. (2023). A deep multi-level attentive network for multimodal sentiment analysis. ACM Transactions on Multimedia Computing, Communications and Applications, 19(1), 1-19.

[4] Das, R., & Singh, T. D. (2023). Image–Text Multimodal Sentiment Analysis Framework of Assamese News Articles Using Late Fusion. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(6), 1-30.

[5] Islam, K. M., Reza, M. S., &Yeaser, M. D. (2021). Sentiment analysis using Natural Language Processing (NLP) & deep learning (Doctoral dissertation, Brac University).

[6] Mahendhiran, P. D., &Kannimuthu, S. (2018). Deep learning techniques for polarity classification in multimodal sentiment analysis. International Journal of Information Technology & Decision Making, 17(03), 883-910.

[7] Lu, Q., Sun, X., Long, Y., Gao, Z., Feng, J., & Sun, T. (2023). Sentiment analysis: Comprehensive reviews, recent advances, and open challenges. IEEE Transactions on Neural Networks and Learning Systems.

[8] Aslam, A., Sargano, A. B., & Habib, Z. (2023). Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks. Applied Soft Computing, 144, 110494.

[9] Dui, Y., & Hu, H. (2024). Social media public opinion detection using multimodal natural language processing and attention mechanisms. IET Information Security, 2024(1), 8880804.

[10] Aslam, A., Sargano, A. B., & Habib, Z. (2023). Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks. Applied Soft Computing, 144, 110494.

[11] Zhang, F., Li, X. C., Lim, C. P., Hua, Q., Dong, C. R., & Zhai, J. H. (2022). Deep emotional arousal network for multimodal sentiment analysis and emotion recognition. Information Fusion, 88, 296-304.

[12] Tejaswini, V., Sathya Babu, K., & Sahoo, B. (2024). Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(1), 1-20.

[13] Perti, A., Sinha, A., & Vidyarthi, A. (2024). Cognitive hybrid deep learning-based multi-modal sentiment analysis for online product reviews. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(8), 1-14.

[14] Kusal, S., Panchal, P., & Patil, S.A. (2024). Pre-Trained Networks and Feature Fusion for Enhanced Multimodal Sentiment Analysis. 2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon), 1-7.

[15] Wang, F., Tian, S., Yu, L., Liu, J., Wang, J., Li, K., & Wang, Y. (2023). TEDT: transformer-based encoding–decoding translation network for multimodal sentiment analysis. Cognitive Computation, 15(1), 289-303.

[16] Huang, C., Zhang, J., Wu, X., Wang, Y., Li, M., & Huang, X. (2023). TeFNA: Text-centered fusion network with crossmodal attention for multimodal sentiment analysis. *Knowledge-Based Systems*, *269*, 110502.

[17] Bairavel, S., & Krishnamurthy, M. (2020). Novel OGBEE-based feature selection and feature-level fusion with MLP neural network for social media multimodal sentiment analysis. *Soft Computing*, *24*(24), 18431-18445.

[18] Huang, C., Chen, J., Huang, Q., Wang, S., Tu, Y., & Huang, X. (2025). AtCAF: Attention-based causality-aware fusion network for multimodal sentiment analysis. *Information Fusion*, *114*, 102725.