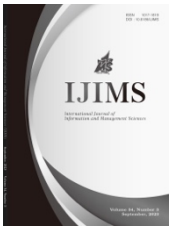
http://ijims.ms.tku.edu.tw/main.php

International Journal of
Information
and
Management
Sciences

International Journal of
Information and Management Sciences

36 (2025), 255-274. DOI:10.6186/IJIMS.202509_36(4).0001



Sentiment Classification in Textual Data Using Combination of Features from Deep Learning Models

Adepu Rajesh[1,2,*] and Tryambak Hiwarkar[1]
[1] Department of Computer Science and Engineering, School of Engineering Sardar Patel University, Balaghat, Madhya Pradesh, India
[2] Department of Computer Science & Engineering, Guru Nanak Institute of Technology, Ibrahimpatnam, Hyderabad, Telangana, India

Keywords

Sentiment Analysis, Textual Data, Deep Learning, Machine Learning, Hybrid Model

Abstract

Sentiment analysis determines the emotional tone of a piece of text i.e., positive, negative, or neutral. In modern era, the comments or sentiments are the ways of communication to express the viewpoints, further, managing and analysing of business based on movies reviews, sentiments play vital role. The use of machine learning (ML) and deep learning (DL) models in sentiment prediction offers significant improvements in sentiment classification. ML methods typically use features extracted from the data, such as word counts, part-of-speech tags, and sentiment lexicons. DL models learns to extract features automatically from the data that are not easily observable by humans, like emotional tone in voice or image sentiment. It can be difficult to extract features that are both informative and discriminative for sentiment analysis. In this work, hybrid DL method is proposed for sentiment analysis on movie reviews, and accuracy of 96.34% is achieved in comparison with existing methods.

1. Introduction

Sentiment analysis (SA), also known as opinion mining, is a natural language processing (NLP) task that involves determining and categorizing the sentiment or emotional tone expressed in a piece of text, speech, or other forms of communication. The primary goal of SA is to understand and extract subjective information from the given content, typically categorizing it into one or more sentiment categories, which commonly includes – 1) Positive - the text expresses a favourable sentiment such as happiness, satisfaction, or approval. 2) Negative - the text conveys an unfavourable sentiment such as anger, sadness, disappointment, or criticism. 3) Neutral - the text does not express any particular sentiment and considered as neutral (Ombabi et al, 2020).

*corresponding author. Email: adepurajeshadepu@gmail.com

In the 1990s, SA saw the emergence of lexicon-based methods, which relied on sentiment dictionaries containing lists of words and their associated sentiment scores. Researchers created lexicons such as "General Inquirer" and "SentiWordNet" to help to classify words and phrases into positive, negative, or neutral categories. These lexicons formed the basis for rule-based SA systems.

Subsequently, it is observed that Social Media analysis uses SA in a productive way. Social media platforms are rich sources of multimodal data includes text, images, and videos. SA can help in tracking public sentiment towards brands, products, events, or political figures. Further, ML and DL models were utilized to analyse text comments, images, and emojis in social media posts to understand the overall sentiment and emotional tone of the discussions (Cen et al. 2020).

Eventually, Support Vector Machine (SVM) was used for classification and regression tasks, shown to be effective for SA on textual and multimodal data, especially when combined with other techniques such as text classification and image recognition. Further, Convolutional Neural Networks (CNNs), a DL algorithm commonly employed for image classification tasks, also, reported to be effective for SA on textual and multimodal data when combined with text classification techniques. Recurrent Neural Networks (RNNs), a sequence data-processing model found more effective for SA (Zhou, 2022). Further, Bi-directional RNN (BRNN), variant of RNN that process both past and future context, also, found suitable for SA. This makes these aforementioned models well suited for SA tasks that require understanding the relative importance of different words or phrases.

This study utilize ML and DL models for SA. Deep learning is specialized form of ML that uses neural networks to understand complex patterns in data. Pre-trained CNNs allow us to use state-of-the-art DL models by drastically reducing time and space complexity (Tripathy et al., 2016).

The problem is to develop an effective SA system that can automatically determine the sentiment (*positive, negative, or neutral*) expressed in textual data. Specifically, the main goal is to build a model capable of classifying text into these sentiment categories with a higher degree of accuracy using ML and/or DL techniques. The main motivation is to propose the hybrid DL model to use embedding techniques for capturing textual semantics, CNN to extract the local dependencies, and LSTM / BiLSTM to handle sequential and long-term dependencies in the text in different settings to achieve higher accuracy.

The some identified challenges is sentiment analysis are:

- To capture subjectivity (Tembhurne & Diwan, 2021), obtaining inherent sentiment, and amount of association in text is difficult.
- In SA, understanding the sarcasm from the textual or multimodal data is challenging wherein context identification is very difficult.
- Model accuracy is less for long text containing a lot of content and correlation between words as compared to shorter texts.

- Limited accuracy of ML models (Bodapati et al., 2019; Dang et al., 2021; Shaukat et al., 2020) and DL models (Dang et al., 2021; Perumalet al., 2023, Tripathy et al., 2016), and (Ullah et al., 2022) for SA task.

The key objectives for SA in textual data using a combination of features from DL models are as follows:

- Development of Effective SA Model: To create a robust model that accurately classifies text into sentiment categories (e.g., positive, negative, neutral).
- Leveraging DL for Feature Extraction: To utilize the DL models such as CNN, LSTM, and BERT to extract meaningful features from textual data.
- Combining Features for Improved Performance: To integrate features from multiple DL models to enhance sentiment classification accuracy.
- Evaluate Performance: To test the model on IMDB dataset and compare results with existing methods.

The textual SA is an extended approach to traditional language-based sentiment analysis. Analyzing the sentiment of people from their reactions and comments can result into automation and major help in sector of business for understanding reviews of the products, and advancements in the product of a company. Further, audio-visual SA model can contribute to real-world industrial applications. The contributions in this work are as follows:

- The word embedding is implemented (Devlin et al., 2018; Mikolov et al., 2013) for better feature representation, and training is performed on large dataset in unsupervised fashion.
- To propose hybrid feature extraction framework using different deep learning models for sentiment analysis.
- To propose the DL-based hybrid model using advantages of CNN – extract spatial local features and RNN – captures temporal features for sentiment analysis on movie review tweets with higher accuracy.
- To compare proposed system with the state-of-the-arts to witness the suitability and applicability in different areas of applications.

The paper is organized as – literature review is presented in Section 2. The datasets, ML, DL and proposed methodology is discussed in Section 3. Section 4 highlighted the results and corresponding discussions, finally conclusion is proposed in Section 5.

2. Related Works

Lately, there has been a surge in user-generated content on social media platforms (Ombabi et al., 2020). A DL model for Arabic sentiment analysis was introduced, utilizing a single-layer CNN coupled with two LSTM layers. The FastText word-embedding model serving as the input layer reinforces the architecture. Through, experiments conducted on a multi-domain corpus, the model exhibited remarkable performance, achieving precision, recall, F1-score, and accuracy scores of 89.10%, 92.14%, 92.44%, and 90.75%, respectively. The findings indicate that SVM emerges as the most effective classifier, showcasing a noteworthy accuracy improvement of up to +3.92%. Additionally, the LSTM component demonstrated a substantial accuracy

enhancement of +11.6%.

According to Hu et al. (2022), a psychological perspective was investigated and jointly modelling sentiment and emotion is feasible and reasonable. The unified multimodal knowledge-sharing framework (UniMSE) was developed to captures knowledge of sentiment and emotion, where input features and output labels were aligned. Moreover, acoustic and visual details were fused with multi-level textual features.

According to Zhou (2022), emotional part of speech features, location information features, and dependency features of words, and word embedding were adopted to design the CNN + BiLSTM with support of attention was experimented. In addition, improvement in the performance of model was reported.

In the realm of sentiment classification, the significance of feature extraction and classifier design cannot be overstated. Addressing sentiment classification within Twitter's Covid-19 dataset, Xiaoyan et al. (2022) introduced the GloVe+CNN+BiLSTM model as a solution. The accuracy achieved was 95.65%, which is due the combination of features from different models. In Rao (2022), multilevel features extraction for sentiment classification was performed. Further, Hassan et al. (2018) proposed CNN and RNN model for movie review and reported the accuracy of 93.3%. In (Prabhu & Nashappa, 2023), weight function along with Bidirectional Encoder Representations from Transformers (BERT) was modelled for movie reviews, the WordNet is utilized to prepare knowledge base and BERT encoding is employed for sentiment computation. The accuracy of 93.66% is achieved for this system.

In (Cen et al., 2020), authors experimented DL method to analyse the sentiment of the movie reviews. For the user, it is a recommendation tool for movie selection, which helps the user make a choice. For film companies, this information could be used for marketing decisions and finding customers. The three DL models such as CNN, RNN and LSTM were employed and comparison is accomplished with Recursive Neural Tensor Network (RNTN). In (Shaukat et al., 2020), emotions or feelings are captured based on documentary information i.e., opinions and factual information. Here, opinions are subjective expressions - people's views, feelings, or sentiments about various aspects of events, objects, or entities. On the other hand, facts are objective statements about events, entities, objects, characteristics, and their attributes. In (Bodapati et al., 2019), sentiments forecasting was performed by examining available reviews. The sequential model for discerning the sentiment in movie reviews was employed and LSTM is experimented on IMDB dataset (www.imdb.com), the authors reported the improved results. In (Perumal et al., 2023), the comparative study of SA on different dataset is presented by using different DL models such as CNN, RNN, LSTM, and Gated Recurrent Unit (GRU). The highest accuracy of 87.04% is reported by GRU on IMDB dataset, and highest accuracy of 92.27% is achieved by LSTM on ARAS dataset (Alemdar et al., 2013). In addition, 99.69% accuracy is reported by CNN on Fruit-30 dataset (Muresan & Oltean, 2018).

Furthermore, the sentiment analysis is also experimented for the multimodal data to predict the significant sentiments (Geng et al., 2022; Gu et al., 2021; Harish et al., 2019; Huddar et al., 2021; Peng et al., 2021; Qi et al., 2022; Seo et al., 2020; Wang et al., 2021; Zou et al., 2022;). These studies highlights the usage of hybrid models for the extraction of feature from the different data modalities such as audio, video, text, and images. Thus, multimodal sentiment

analysis opens the avenue for hybrid model development for the task of sentiment analysis.

2.1 Motivation

After successful literature review, it identified that the models implemented in (Ombabi et al., 2020; Xiaoyan et al., 2022), (Bodapati et al., 2019; Cen et al. 2020; Dang et al., 2021; Shaukat et al., 2020), (Harish et al., 2019; Hassan & Mahmood, 2018), (Domadula & Sayyaparaju, 2023; Sharma et al., 2023; Ullah et al., 2022), and (Prabhu & Nashappa, 2023) utilized the ML methods, combination of word embedding and ML methods, and individual DL model or combination of different DL models. Table 1 shows the model used, dataset, result obtained, findings and scope for further research. Thus, we are having many motivating scenarios such as suitable use of word embedding technique, hybrid model composition, and improvement in model classification accuracy, etc. In this paper, we exclusively targeted to use BERT / Word2Vec embedding for better feature representation, in addition, CNN, LSTM and BiLSTM models are utilised for efficient feature learning. Finally, the proposed hybrid model outperform the state-of-the-arts by achieving higher accuracy, which is discussed in Section 4. In this work, the CNN and LSTM / BiLSTM models are utilized as per the capabilities of these models in SA. The CNN captures the local and spatial features from textual data whereas LSTM / BiLSTM handle long-term dependencies and retaining information from previous words. In addition, BiLSTM process text in forward and backward directions, capturing past and future context. Further, Table 1 present the existing works highlighting the model applied on different datasets and accuracy achieved with future scopes. Thus, we targeted to apply advance DL model to achieve higher accuracy for the SA.

Table 1. *Comparison with Existing Work for Motivation*

Ref.	Model Used	Dataset	Accuracy (%)	Findings	Future Scope
Ombabi et al., 2020	CNN+LSTM	IMDB	90.75%	Low accuracy due to morphological complexity	Improve upon the accuracy, more efficient word representation
Xiaoyan et al., 2022	GloVe+CNN +BiLSTM	COVID-19	95.65%	Improved accuracy, word representation using GloVe	SA in Chinese online text, hybrid model can be applied to other domain
Cen et al., 2020	CNN RNN LSTM	IMDB	88.22% 68.64% 85.32%	Lower accuracy	Improve upon the accuracy, could use hybrid DL models
Shaukat et al., 2020	Lexicon - MLP	IMDB	91.90%	ML method is applied, lower accuracy	Advanced DL models can be applied, accuracy can be improved
Bodapati et al., 2019	LSTM	IMDB	88.46%	Lower accuracy, no suitable embedding is utilized	Hybrid model can be used, accuracy can be improved

Dang et al., 2021	LSTM+CNN +SVM	IMDB	91.10%	Hybrid model used, lower accuracy obtained	Different combination of hybrid model, accuracy can be improved
Harish et al., 2019	ML models	IMDB	83.93%	Hybrid ML model used but accuracy is lower	Lexicon features extraction, improved hybrid model with DL based approaches
Hassan & Mahmood, 2018	CNN+LSTM	IMDB	93.30%	Hybrid model used but accuracy is limited	Machine translation and information retrieval can be performed, accuracy can be improved
Domadula & Sayyaparaju, 2023	BERT	IMDB	90.67%	BERT for feature learning, lower accuracy	Hybrid model can be used, accuracy can be improved, can be used for multi-class classification
Sharma et al., 2023	ML models	IMDB	73.00%	Hybrid ML modes used, very lower accuracy	Advanced DL models can be applied, accuracy can be improved
Ullah et al., 2022	CNN	IMDB	92.00%	Single DL model used, limited accuracy	Hybrid DL models can be applied, accuracy can be improved
Prabhu & Nashappa, 2023	WF-BERT	IMDB	93.66%	Limited accuracy, only BERT is utilized for sentiment learning	Aspect based SA for efficient text classification

3. Materials and Methods

3.1 Dataset

To highlight which model yields superior results in SA, we employed the publicly available IMDB dataset (Lakshmipathi, 2023). This dataset comprises 50,000 reviews sourced from the online movie database, with 25,000 allocated for training and the remaining 25,000 for testing. To maintain a balanced and equitable evaluation, both negative and positive comments are equally represented, each making up 50% of the dataset.

3.2 Preprocessing

We have performed the pre-processing on the IMDB dataset. In data file, we have *html*, ',' , '.', thus, the punctuation marks are removed to achieve interference-free text that is not yet recognized by the computer. Therefore, we also need to convert our text into a one-dimensional vector; hence, Word2Vec embedding is adopted. There are only 30 reviews per movie, as reviews for the same movie tend to have correlated ratings. Furthermore, the train and test sets contain a disjoint set of movies so memorizing a particular movie term and their associated labels would have no significance. A negative review is given a score of $<= 4$ out of 10 while a positive one holds a score of $>= 7$, and a neutral review has scores from > 4 and < 7. For evaluation and model selection, we used k-fold cross validation to take different portions of the training set to be utilized as the validation set.

3.3 Proposed Methodology

The proposed system for SA in this work compromises three stages, i.e. Data Collection, Data Preprocessing, Feature representation, Feature learning and classification. Fig. 1 shows the proposed system for the task of SA on movie review dataset. The details of each stage is presented as follows.

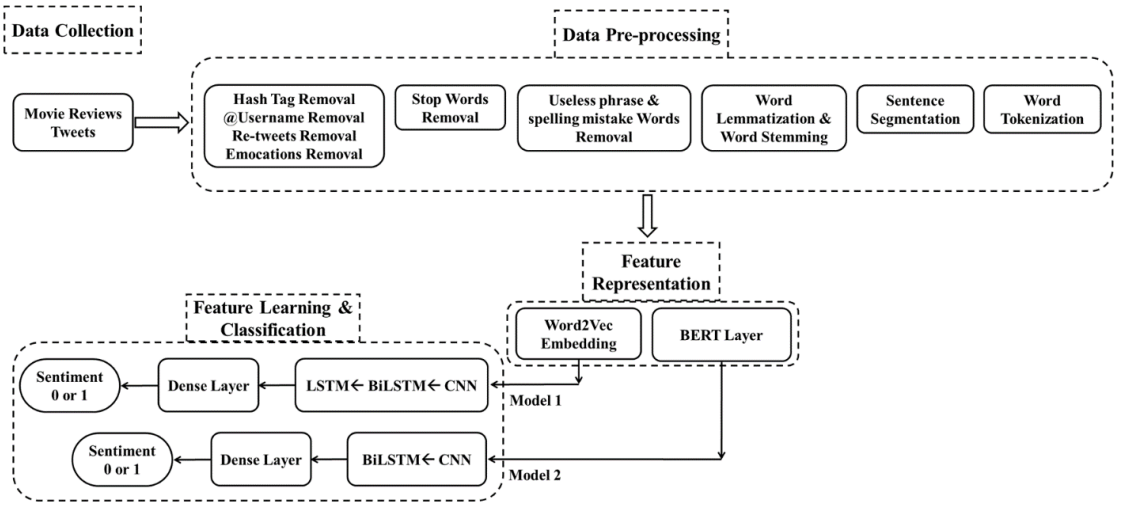


Fig. 1 *System Diagram for Sentiment Analysis*

Data collection: The IMDB movie review dataset is utilized for the task of SA. The movie's reviews are stored based on sentiment polarity.

Data preprocessing: The pre-processing of data implies the processing of raw data into a more convenient format, which could be fed to a classifier for better classification. Here, the raw data is extracted from twitter using an API. In twitter terminology, there are various common useless phrases and spelling mistakes present in the data, which need to be removed to boost the accuracy of the resultant data. These could be summoned up as follows:

• *Hash tags*: These are very common in tweets, Hash tags (#topic) represent a topic of interest about which the tweet is being written.

• *@Usernames*: These represent the user mentions in a tweet.

• *Retweets (RT):* As the name suggests retweets are used when a tweet is posted twice by same or different user.

• *Emoticons*: These are very commonly found in the tweets. Using punctuations facial expressions are formed in order to represent a smile or other expressions, these are known as emoticons or emoji's.

• *Stop words*: Stop words are those words, which are useless when it comes to SA. Words such as '*it*', '*is*', '*the*', '*him*', '*both*', etc. are known as stop words.

• Word *Stemming:* In stemming, we strip the word and remove its prefixes and suffixes; finally, root word is derived. This is performed using NLTK (Natural Language Tool Kit) library in Python.

• *Word tokenization*: In tokenization, we convert sentence into separate words called as tokens. In proposed system, word level tokenization is achieved to offer word's sentiment polarity.

Feature representation: In the proposed system, Word2Vec and BERT are utilized for better feature representation for the text. The advantages offered by Word2Vec are – simple, fast training, no tagging, suitable for any dataset length, captures semantic details, and unsupervised method. Further, BERT offers advantages as – efficient representation of contextual information, attention mechanism, understand tones of expression, bi-direction behaviour to capture past and future context, and unsupervised method.

Feature learning and classification: We developed two ensemble models (Model 1 and Model 2) for feature learning which combines CNN+BiLSTM+LSTM and CNN+BiLSTM. The CNN extract the spatial features called to be local features, LSTM/BiLSTM is used to extract the temporal features to learn the long-term dependencies found in the local features. This is followed by drop-out layers, pooling layers, and lastly, classification score is generated by the dense layer.

3.3.1 Machine Learning Models

In this section, we have explored the ML classifier algorithms for the sentiment analysis but the ultimate goal is to develop DL model for accurate sentiment classification.

Bayesian logistic regression: This process involves feature selection and optimization to facilitate text categorization. It incorporates a Laplace prior to prevent overfitting and generates sparse predictive models tailored for text data. Logistic regression estimation $P(c|f)$ is expressed in a parametric form as:

$$P(c|f) = \frac{1}{z(f)} exp\left(\sum_i \lambda_{i,c} F_{i.c}(f,c)\right) \quad (1)$$

Here, $z(f)$ functions as a normalization operation, λ represents a weight parameter vector for a specific set of features, and $F_{i,c}$ is a binary function. This function activates when a particular feature is present and the sentiment associated with it is postulated in a specific manner.

Naïve Bayes: This classifier operates on probabilistic principles, relying on a robust assumption of conditional independence. It excels in classifying categories with interdependent features. Sentiment class probabilities are computed using Bayes' theorem. Naïve Bayes is known for its simplicity, yielding decent results, although it may not perform as effectively as some other classifiers.

$$P(X|y_i) = \prod_{i=1}^{m} P(x_i|y_i) \tag{2}$$

X is a feature vector defined as $X = \{x_1, x_2, ...x_m\}$ and y_j is a class label.

Support Vector Machine: SVMs are supervised models with associated learning algorithms that analyse data used for classification and regression analysis. It makes use of the concept of decision planes that define decision boundaries.

$$g(x) = w^T \phi(x) + b \tag{3}$$

x is feature vector, w is weights of vector and b is bias vector. $\Phi()$ is the non-linear mapping input space to high dimensional feature space. SVM can be used for pattern recognition.

Maximum Entropy Classifier: This classifier takes no assumptions regarding the relations between features; it always tries to maximize entropy of a system by computing its conditional distribution of its class labels.

$$P_\lambda(y|X) = \frac{1}{z(f)} exp(\sum_i \lambda_i F_i(X, y)) \tag{4}$$

'X' is the feature vector and 'y' is the class label. $Z(X)$ is the normalization factor and λ_i is the weight coefficient $F_i(X, x)$ which is the feature function which is defined as;

$$f_i(X, x) = \begin{cases} 1, & X = x_i \; and \; y = y_i \\ 0, & otherwise \end{cases} \tag{5}$$

3.3.2 Deep Learning Models

In this section, we discuss on the various DL models utilized for the sentiment classification.

Recursive Neural Tensor Networks: RNTNs are a type of neural network architecture employed in NLP. These networks adopt a tree-like structure featuring a neural network at every node. RNTNs could be employed for tasks such as boundary segmentation, distinguishing between positive and negative word groupings, and extending their applicability to entire sentences. Word vectors play a pivotal role as feature vectors, forming the foundation for sequential classification. These word vectors are subsequently organized into subphrases, and these subphrases are aggregated into full sentences, which can then undergo SA and other relevant metrics. In the realm of neural network-driven text analysis, words can be effectively represented as continuous parameter vectors. These word vectors encapsulate not only the essence of the individual word but also encapsulate contextual information derived from the surrounding words, encompassing usage and other semantic nuances. It's worth noting that, while deep learning includes the implementation of Word2Vec, the incorporation of RNTNs is

not currently part of our operational framework.

Convolutional Neural Network: CNN represents a type of feedforward neural network and stands as one of the most well established domains in the application of DL algorithms. It possesses a robust capacity for learning distinctive features. CNN comprised of three key layers: convolutional layer, pooling layer, and fully connected layer (O'Shea & Nash, 2015). In (Diwan & Tembhurne, 2022), sentiment analysis on various modalities using CNNs is surveyed and challenges, issues, and suitability of CNNs for SA is explored. The CNN for sentiment analysis is presented in Fig. 2.

The operation of convolution layer in forward direction is represented by Eq. (6), where W represent the weight matrix at each layer, X correspond to input vector and bias b at each layer.

$$C = conv(W, X) + b \tag{6}$$

Further, the prediction into different category is given by Eq. (7). In addition, error term is computed from the prediction and back propagation algorithm is utilized to achieved the optimal weights.

$$Y = \emptyset(C) \tag{7}$$

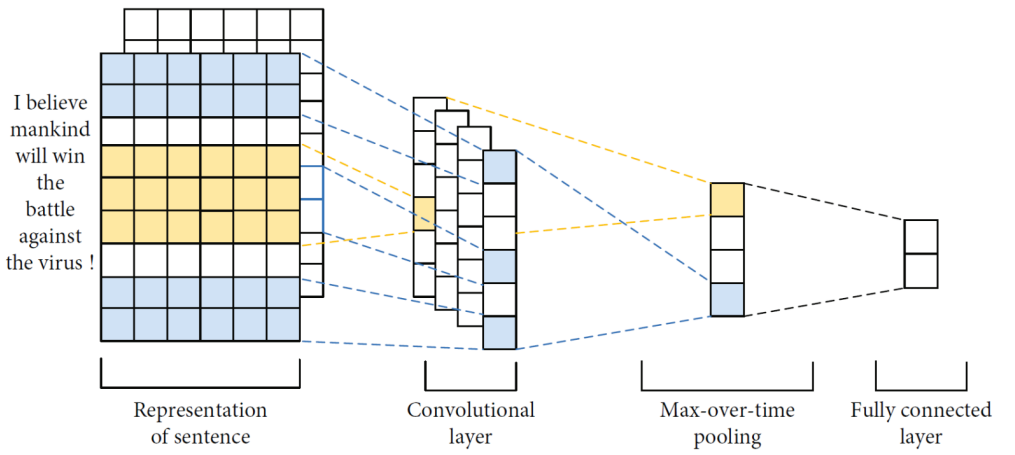


Fig. 2 *CNN for Feature Sentiment (Xiaoyan et al., 2022)*

Recurrent Neural Network: RNN (Sherstinsky, 2020) is a type of neural network designed for processing sequential data. It iteratively operates along the sequence's temporal dimension, with all nodes interconnected in a chain-like fashion. RNNs possess shared memory and parameters, making complete and advantageous for studying the nonlinear properties of sequences. These cyclic neural networks find applications in NLP, including tasks such as speech recognition, language modelling, machine translation, and time series forecasting, etc. Eq. (8) shows the RNN cell equations with internal state (ht).

$$h_t = f(Vx_t + Uh_t + b) \tag{8}$$

Here, f() is the activation function (typically the hyperbolic tangent), U and V are the weights corresponding to hidden state (h), bias is b, and x represents input vector.

In (Tembhurne & Diwan, 2021), a comprehensive review of SA is presented using RNNs for unimodal and multimodal data. This review offers the utilization of RNNs, customized RNNs, and different challenges in the development of DL models.

Long Short-Term Memory: LSTM (Sherstinsky, 2020) is the superior RNN model to handle the long-term dependencies easily. Over the past few years, LSTM have gained significant traction in the realm of NLP tasks such as language analysis and text classification. Total three layers makes one LSTM cell i.e. input gate, forget gate, and output gate. The LSTM cell diagram is presented in Fig. 3.

These gates are designated to perform the specific task to control the information flow. Firstly, input gate fetches the past internal state and current input to update the latest internal state. Secondly, forget gate is responsible to forget the irrelevant information. Finally, output gate produce the results based on the internal state (more details see (Sherstinsky, 2020)). The LSTM cell equations are shown as follows;

$$h_t = \tanh(C_t) \times O_t \tag{9}$$

$$C_t = \sigma\left(f_t \times C_{t-1} + i_t \times \widehat{C}_t\right) \tag{10}$$

$$\widehat{C}_t = tanh(d_t U + h_{t-1} V) \tag{11}$$

$$O_t = \sigma(d_t U_o + h_{t-1} V_o) \tag{12}$$

$$f_t = \sigma\left(d_t U_f + h_{t-1} V_f\right) \tag{13}$$

$$i_t = \sigma(d_t U_i + h_{t-1} V_i) \tag{14}$$

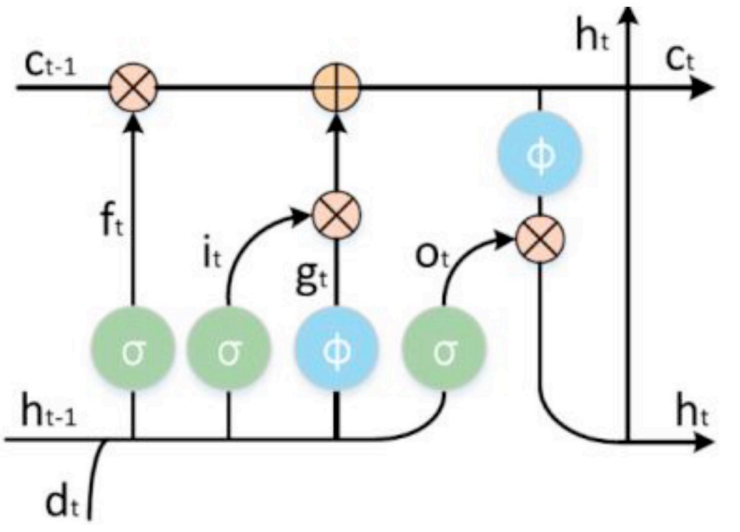


Fig. 3 *LSTM Cell (Bodapati et al., 2019)*

BERT (Bidirectional Encoder Representations from Transformers): BERT (Devlin et al., 2018) designed to handle bidirectional representation of textual data where joint conditioning is considered for context in left and right. It is simple and powerful model for context representation consisting multiple layers of bidirectional Transformer encoder. The pre-trained BERT model for the fine-tuning is shown in the Fig. 4.

The embedding vector is prepared for each word i.e. $word_1$, $word_2$, ..., $word_n$, these embedding vector is a dense network to acquire sentiment classification. The weight is assigned to each word; further, weighted average is computed to enhance the classification performance. A weighted matrix is obtained after the training, the sentiment score is found in the form of class probability i.e. P_c. The categories of sentiment for classification can be visualize as $SC = \{sc_1, sc_2, \ldots\ldots, sc_n\}$. To determine class P_c for specific category in BERT, Eq. (15) can be used.

$$P_c = W_f \times Vec = \{P_{c_1}, P_{c_2}, \ldots\ldots, P_{c_m}\} \tag{15}$$

Here, W_f is weighting function, Vec vector representation and m is total sentiment categories.

Thus, word embedding is computed which is the average of all embedding terms. This, weighted average increase the performance of classification, reported by the research in NLP.

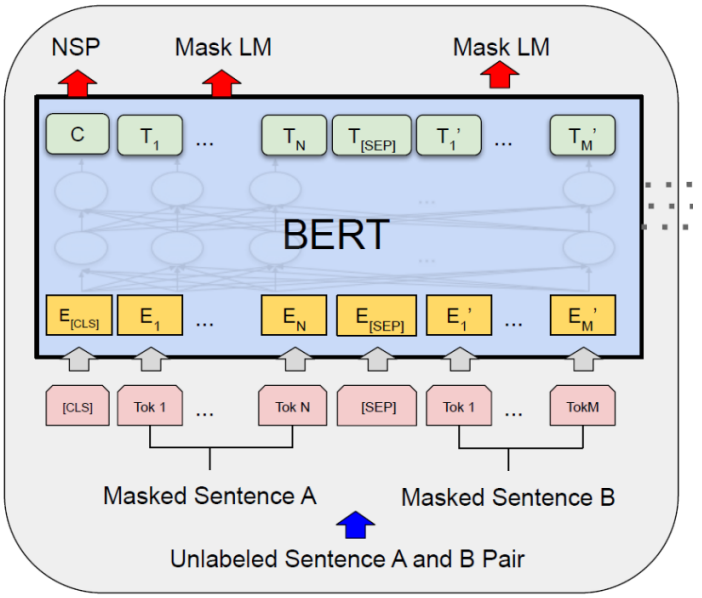


Fig. 4 *Pretrained BERT Model (Devlin et al., 2018)*

Fig. 5 shows the proposed CNN, BiLSTM and LSTM hybrid model (Model 1) for the sentiment detection from movie reviews. After preprocessing performed on the raw data, the Word2Vec embedding is applied for feature representation. Further, batch normalization is used for faster and stable training in neural network, which helps in reliability and efficiency of neural network. The dropout of 0.5 is applied to reduce the irrelevant parameters for further processing. Further, CNN layer is adopted to extract the spatial local features, which is followed by dropout layer (0.5). Next, global maxpooling is utilized to select most efficient features and data volume further reduce to achieve dimensionality reduction. Then, BiLSTM layer is

employed to learn the long-term dependencies in the text in both the directions, further; LSTM is used to acquire the final learned features and context of text data. Finally, dense layer is utilized to perform the classification.

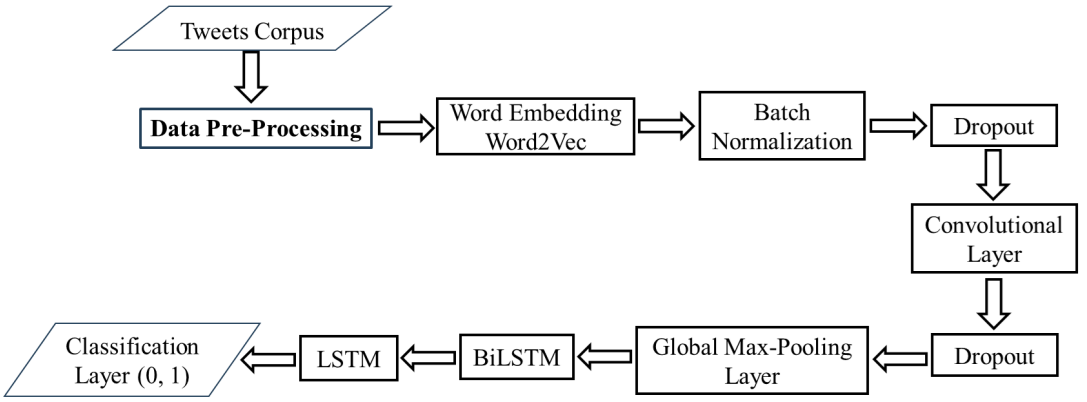


Fig. 5 ***Model 1:*** *CNN-BiLSTM-LSTM Model for SA*

Fig. 6 present another hybrid model (Model 2) which uses BERT to offer the efficient word embedding, and CNN and BiLSTM are utilized for feature extraction and learning to detect the sentiment from the input data. Here, BERT embedding is performed first for feature representation and CNN is applied for prominent and local feature extraction, followed by BiLSTM for temporal feature learning to understand the context of the text.

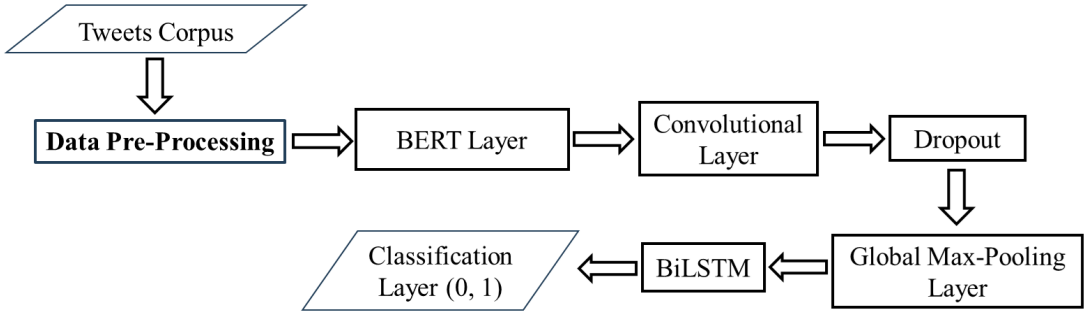


Fig. 6 ***Model 2:*** *BERT-CNN-BiLSTM Model for SA*

4. Results and Discussion

To evaluate the performance of proposed models for SA, the following performance metrics can be utilized, but for this work, we only investigated on accuracy for sentiment analysis.

Accuracy: Accuracy of a classifier indicates how accurately the classifier predicted the result. It can be expressed as follows:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \tag{16}$$

Precision: precision shows how often the result that is being predicted by the classifier, when it indicates true is correct. The formula for precision is:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{17}$$

Recall: It indicates the true positive rate of the classifier.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{18}$$

F1-Score: F1-Score measure provides a single score that provides the combined measure of the previous two measures i.e., precision and recall and can be defined as a harmonic mean of both of these.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{19}$$

The IMDB dataset is employed for experimentation - contains 50,000 movie reviews, evenly split into 25,000 positive and 25,000 negative reviews, all composed in English. These reviews come in various file sizes, ranging from 1KB to 15KB. It's worth noting that there are no ratings included within the text files. We divided the dataset into an 80% training set and a 20% testing set for the analysis.

Table 2 summarizes the specifications of the environment used to build and run the proposed models. Table 3 presents the hyperparameter used for the training of the proposed models.

Table 2. ***Development Environment specification***

Component	Configuration
OS	Windows 11
Memory	8 GB
Processor	Ryzen 5
Development tool	Python 3.11.4
Libraries	Keras and Tensorflow

Table 3. ***Hyperparameters for SA Models***

Parameter(s)	Values
No. of Epochs	5
Batch Size	32
Optimizer	Adam
Drop-out	0.5
Verbose	2
Activation Function	Sigmoid , Relu

Table 4 provides a comprehensive overview of proposed work and related works. The increased accuracies of 3.84% using MLP, 8.2% using CNN is reported in comparison with (Dang et al., 2021), respectively. Further, 10.62% improvement in accuracy by using CNN and LSTM over RNTN (Bodapati et al., 2019).

We also experimented Naïve Bayes, SVM, MLP, RNTN, CNN, LSTM and CNN+LSTM on different dataset comprising 752 negative and 1301 positive reviews, totalling 2053 reviews. In contrast, (Bodapati et al., 2019) utilized another dataset of English movie reviews extracted from (Pang and Lee, 2005), consisting of 11,855 individual sentences extracted from movie reviews. It's noteworthy that CNN and LSTM hybrid model has surpassed all other models. This outcome is promising, especially when considering the considerably larger number of processed reviews. Furthermore, there's potential for achieving even higher accuracy by enhancing the word embedding technique.

Table 4 helps in identifying the significance of independent ML and DL models in SA. These, results offers the direction in investigating the ensemble of DL models with efficient word embedding techniques.

Table 4. *Comparison with Existing Works*

Previous work			Proposed Models (50k Review files)	
SVM (Dang et al., 2021)	2035 review files	82.90%	MLP	86.74%
NB (Dang et al., 2021)		81%	CNN	89.20%
			LSTM	86.64%
RNTN (Bodapati et al., 2019)	11,855 sentences	80.70%	CNN-LSTM	91.32%

Table 5 shows the performance analysis of proposed SA models on IMDB dataset. The improved accuracy of 9.34% is achieved by the proposed model in comparison with (Jnoub et al., 2020). Further, accuracies improvement of 5.34% in (Shaukat et al., 2020), 5.56% in (Dang et al., 2021), 24.22% in (Rizal & Purbolaksono, 2022), 4.34% in (Ullah et al., 2022), 5.67% in (Domadula & Sayyaparaju, 2023), 23.34% in (Sharma et al., 2023), 9.3% in (Perumal et al., 2023) are achieved in comparison with proposed SA models. Thus, the proposed model outperforms the state-of-the-arts results.

Table 5. *Comparison with Existing Work on IMDB Dataset*

Ref.	Model Used	Accuracy (%)
Jnoub et al., 2020	SNN	87.00%
	CNN	81.00%
Shaukat et al., 2020	MLP	91.00%
Dang et al., 2021	CNN	87.30%
	LSTM	85.10%
	CNN+LSTM	89.20%
	LSTM+CNN	89.40%
	CNN+LSTM+BERT	90.70%
	LSTM+CNN+BERT	90.70%
Rizal and Purbolaksono, 2022	Word2Vec+Naive Bayes	72.23%
Ullah et al., 2022	CNN	92.00%
Domadula and Sayyaparaju, 2023	BERT	90.67%
Prabhu et al., 2023	WF-BERT	93.66%
Sharma et al., 2023	SVM	73.00%
Perumal et al., 2023	CNN	86.69%
	BiLSTM	85.95%
	GRU	87.04%
Proposed Model	Word2Vec+CNN+BiLSTM+LSTM	94.26%
	BERT+CNN+BiLSTM	96.34%

The details discussions on obtained results in comparison with existing models are highlighted as follows:

- The Model 1 and Model 2 are very efficient in feature representation and classification accuracy in comparison with (Ombabi et al., 2020), 3.51% to 5.59% increment in accuracy is witnessed. This is due to the utilization of superior word embedding i.e. Word2Vec / BERT, and BiLSTM to acquire context information in both the directions.
- In (Xiaoyan et al., 2022), similar approach is utilized with GloVe embedding, which is unable to learn the context based on word ordering. However, in proposed model context aware embedding is utilized in the form of BERT thus increase in accuracy of 0.69% is reported with efficient training.
- The performance comparison with (Cen et al., 2020), reported the enhanced accuracy of 11.02% by proposed model, which is due to feature combination from the different DL models. In (Cen et al., 2020), only individual models of CNN, RNN and LSTM are experimented for movie reviews.
- In (Shaukat et al., 2020), lexicon and MLP based model is investigated; this model achieved an accuracy of 91% which is 5.34% lower than the proposed model. The efficient word embedding and utilization of DL models identifies the improvement in the accuracy. Further, LSTM with dense network is experimented in (Bodapati et al., 2019) for movie reviews, due to simple model, only 88.46% of accuracy is achieved, which is 7.88% lower based on the proposed model. Here, individual model not able to report the higher performance.

- A hybrid model of LSTM+CNN (Dang et al., 2021) is applied for movie reviews but limited accuracy of 91.10% is achieved, here, 3 layers of CNN and 3 layers of dense network were utilized, which make the model complex and takes more time for training. However, proposed model is representative for the spatial local features and temporal features without additional layers and 5.24% improvement in accuracy is recorded.
- Further, BERT (Domadula & Sayyaparaju, 2023) along with softmax is investigated for movie reviews but lower accuracy of 90.67% is achieved which is due to single BERT model but the proposed model achieved more accuracy by using different DL models after BERT embedding. Subsequently, 93.20% accuracy is reported in (Hassan & Mahmood, 2018) by applying CNN+LSTM which is 3.14% lower than the proposed model, this improvement is witness due the utilization of efficient embedding i.e. BERT.
- In (Ullah et al., 2022), two layer CNN is developed for movie reviews where filters are reduced in 2nd CNN layer, this model is able to achieve accuracy of 92% which is lower than 4.34% in comparison with proposed model.
- Lastly, accuracy of 93.66% is obtained by using weighted BERT model in (Prabhu et al., 2023), which is competitive in comparison with proposed model with accuracy of 96.34%. This improvement in accuracy (2.68%) is determine by the additional layers of CNN and BiLSTM for feature learning in proposed model.

5. Conclusion

In this paper, we have demonstrated a system for the analysis of textual data using different methods such as ML and hybrid DL to achieve the better accuracy. Here, we conducted sentiment analysis on IMDB dataset of 50,000 movie reviews using different models such as MLP, CNN, LSTM, and hybrid deep models – Word2Vec+CNN+BiLSTM+LSTM (Model 1) and BERT+CNN+BiLSTM (Model 2). We employed the Word2Vec / BERT technique for word embedding in data representation, which offers efficient feature extraction by the proposed model for SA. Our findings clearly indicate that the Model 1 and Model 2 exhibits superior performance of 94.26% and 96.34%, respectively when compared to state-of-the-arts.

In future, we aim to perform similar analysis in order to increase the accuracy of the model for multimodal sentiment analysis and other areas in which sentiment analysis can be applied.

References

Alemdar, H., Ertan, H., Incel, O. D., & Ersoy, C. (2013). ARAS human activity datasets in multiple homes with multiple residents. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops* (pp. 232-235). IEEE.

Bodapati, J. D., Veeranjaneyulu, N., & Shareef, S. N. (2019). Sentiment Analysis from Movie Reviews Using LSTMs. *Ingénierie des Systèmes d Inf.*, *24*(1), 125-129. DOI: https://doi.org/10.18280/isi.240119.

Cen, P., Zhang, K., & Zheng, D. (2020). Sentiment analysis using deep learning approach. *Journal of Artificial Intelligence, 2*(1), 17-27. DOI: https://doi.org/10.32604/jai.2020.010132.

Chen, C., Ling, P., & Dong, Y. (2022). A multimodal sentiment analysis model based on semantic enrichment. In *2022 4th International Conference on Frontiers Technology of Information and Computer (ICFTIC)* (pp. 106-111). IEEE. DOI: 10.1109/ICFTIC57696.2022.10075170.

Dang, C. N., Moreno-García, M. N., & De la Prieta, F. (2021). Hybrid deep learning models for sentiment analysis. *Complexity*, *2021*, 1-16. DOI: https://doi.org/10.1155/2021/9986920.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. DOI: https://arxiv.org/abs/1810.04805.

Diwan, T., & Tembhurne, J. V. (2022). Sentiment analysis: a convolutional neural networks perspective. *Multimedia Tools and Applications*, *81*(30), 44405-44429. DOI: https://doi.org/10.1007/s11042-021-11759-2.

Domadula, P. S. S. V., & Sayyaparaju, S. S. (2023). Sentiment analysis of IMDB movie reviews: a comparative study of Lexicon based approach and BERT Neural Network model.

Geng, F., Yang, H., Wu, C., & Li, J. (2022, April). Multimodal sentiment analysis based on multi-head self-attention and convolutional block attention module. In *2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)* (pp. 262-266). IEEE. DOI: 10.1109/AEMCSE55572.2022.00059.

Gu, D., Wang, J., Cai, S., Yang, C., Song, Z., Zhao, H., & Wang, H. (2021). Targeted aspect-based multimodal sentiment analysis: An attention capsule extraction and multi-head fusion network. *IEEE Access*, *9*, 157329-157336. DOI: 10.1109/ACCESS.2021.3126782.

Harish, B. S., Kumar, K., & Darshan, H. K. (2019). Sentiment analysis on IMDb movie reviews using hybrid feature extraction method.

Hassan, A., & Mahmood, A. (2018). Convolutional recurrent deep learning model for sentence classification. *IEEE Access*, *6*, 13949-13957. DOI: 10.1109/ACCESS.2018.2814818.

Hu, G., Lin, T. E., Zhao, Y., Lu, G., Wu, Y., & Li, Y. (2022). Unimse: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*.

Huddar, M. G., Sannakki, S. S., & Rajpurohit, V. S. (2021). Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM. *Multimedia Tools and Applications*, *80*, 13059-13076. DOI: https://doi.org/10.1007/s11042-020-10285-x.

Jnoub, N., Al Machot, F., & Klas, W. (2020). A domain-independent classification model for sentiment analysis using neural models. *Applied Sciences*, *10*(18), 6221. DOI: https://doi.org/10.3390/app10186221.

Lakshmipathi, N. (2023). IMDB Dataset of 50k Movie Reviews. https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews. Accessed on 4th November 2023.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Muresan, H., & Oltean, M. (2018). Fruit recognition from images using deep learning. *Acta Universitatis Sapientiae, Informatica*, *10*(1), 26-42.

Ombabi, A. H., Ouarda, W., & Alimi, A. M. (2020). Deep learning CNN–LSTM framework for Arabic sentiment analysis using textual information shared in social networks. *Social Network Analysis and Mining*, 10, 1-13. DOI: https://doi.org/10.1007/s13278-020-00668-1.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*.

Peng, C., Zhang, C., Xue, X., Gao, J., Liang, H., & Niu, Z. (2021). Cross-modal complementary network with hierarchical fusion for multimodal sentiment classification. *Tsinghua Science and Technology*, *27*(4), 664-679. DOI: 10.26599/TST.2021.9010055.

Perumal, T. H. I. N. A. G. A. R. A. N., Mustapha, N. O. R. W. A. T. I., & Mohamed, R. A. I. H. A. N. I. (2023). A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU. https://arxiv.org/ftp/arxiv/papers/2305/2305.17473.pdf.

Prabhu, R., & Nashappa, C. S. (2023). A dynamic weight function based BERT auto encoder for sentiment analysis. *International Journal of Applied Science and Engineering, 21*(1), 1-10.

Qi, Q., Lin, L., Zhang, R., & Xue, C. (2022). Medt: Using multimodal encoding-decoding network as in transformer for multimodal sentiment analysis. *IEEE Access*, *10*, 28750-28759. DOI: 10.1109/ACCESS.2022.3157712.

Rao, L. (2022). Sentiment Analysis of English Text with Multilevel Features. *Scientific Programming*, *2022*, 1-10. DOI: https://doi.org/10.1155/2022/7605125.

Rizal, S., & Purbolaksono, M. D. (2022). Sentiment Analysis on Movie Review from Rotten Tomatoes Using Word2Vec and Naive Bayes. In *2022 1st International Conference on Software Engineering and Information Technology (ICoSEIT)* (pp. 180-185). IEEE. DOI: 10.1109/ICoSEIT55604.2022.10030009.

Seo, S., Na, S., & Kim, J. (2020). HMTL: Heterogeneous modality transfer learning for audio-visual sentiment analysis. *IEEE Access*, *8*, 140426-140437. DOI: 10.1109/ACCESS.2020.3006563.

Sharma, H., Pangaonkar, S., Gunjan, R., & Rokade, P. (2023). Sentimental Analysis of Movie Reviews Using Machine Learning. In *ITM Web of Conferences* (Vol. 53). EDP Sciences. DOI: https://doi.org/10.1051/itmconf/20235302006 .

Shaukat, Z., Zulfiqar, A. A., Xiao, C., Azeem, M., & Mahmood, T. (2020). Sentiment analysis on IMDB using lexicon and neural networks. *SN Applied Sciences*, *2*, 1-10. DOI: https://doi.org/10.1007/s42452-019-1926-x.

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, 132306. DOI: https://doi.org/10.1016/j.physd.2019.132306.

Tembhurne, J. V., & Diwan, T. (2021). Sentiment analysis in textual, visual and multimodal inputs using recurrent neural networks. *Multimedia Tools and Applications*, *80*, 6871-6910. DOI: https://doi.org/10.1007/s11042-020-10037-x.

Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, *57*, 117-126. DOI: https://doi.org/10.1016/j.eswa.2016.03.028.

Ullah, K., Rashad, A., Khan, M., Ghadi, Y., Aljuaid, H., & Nawaz, Z. (2022). A Deep Neural Network-Based Approach for Sentiment Analysis of Movie Reviews. *Complexity*, *2022*. DOI: https://doi.org/10.1155/2022/5217491.

Wang, X., He, J., Jin, Z., Yang, M., Wang, Y., & Qu, H. (2021). M2lens: Visualizing and explaining multimodal models for sentiment analysis. *IEEE Transactions on Visualization and Computer Graphics*, *28*(1), 802-812. DOI: https://doi.org/10.1109/TVCG.2021.3114794.

Xiaoyan, L., Raga, R. C., & Xuemei, S. (2022). GloVe-CNN-BiLSTM model for sentiment analysis on text reviews. *Journal of Sensors*, *2022*, 1-12. DOI: https://doi.org/10.1155/2022/7212366.

Zhou, Z. G. (2022). Research on sentiment analysis model of short text based on deep learning. *Scientific Programming*, *2022*. DOI: https://doi.org/10.1155/2022/2681533.

Zou, W., Ding, J., & Wang, C. (2022). Utilizing BERT Intermediate Layers for Multimodal Sentiment Analysis. In *2022 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6). IEEE. DOI: 10.1109/ICME52920.2022.9860014

Adepu Rajesh
Department of Computer Science and Engineering, School of Engineering,
Sardar Patel University, Balaghat, Madhya Pradesh, India
Department of Computer Science & Engineering, Guru Nanak Institute of Technology,
Ibrahimpatnam, Hyderabad, Telangana, India
E-mail address: adepurajeshadepu@gmail.com
Major Area(s): Artificial Intelligence, Deep Learning and Information Processing.

Tryambak Hiwarkar
Department of Computer Science and Engineering, School of Engineering
Sardar Patel University, Balaghat, Madhya Pradesh, India
E-mail address: drtahiwarkar@gmail.com
Major Area(s): Soft Computing, Machine Learning and Big Data analytics.

(Received February 2024; accepted December 2025)