TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44007

Hybrid CNN-BiLSTM Model for Enhancing Sentiment Analysis using Text Classification on WhatsApp Group

Megha Agarwal[1], Vinodini Katiyar[2], Vandana Patel[3], Bineet Kumar Gupta[4]

[1,2] IT Department, Dr. Shakuntala Misra National Rehabilitation University, UP-India
[3]Applied Sciences, BN College of Engineering & Technology, Lucknow, India
[4]Department of Computer Science and Information Systems, Institute of Technology, Shri Ramswaroop Memorial University, Barabanki, 225003
meghaagarwal2011@gmail.com,

KEYWORDS

Text classification, WhatsApp group, hybrid CNN-BiLSTM;

ABSTRACT

The rapid expansion of social media has led to the generation of massive volumes of data, emphasizing the need to extract valuable insights, categorize information, and predict user sentiments effectively. Text classification, a prominent domain within natural language processing (NLP), focuses on organizing unstructured textual data into sentiment categories to enhance its interpretability. Achieving high accuracy in sentiment categorization calls for refined and efficient text classification techniques. Although Deep Learning models have considerably advanced this field, there remains room for optimization. This study applies the NLP framework to a WhatsApp group dataset to identify sentiment patterns and evaluates five Deep Learning models: Neural Network, Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Bidirectional LSTM, and Convolutional Neural Network (CNN). Furthermore, it introduces a hybrid CNN–BiLSTM model that integrates feature extraction mechanisms with specific activations, dropouts, filters, kernel sizes, and layered structures to enhance sentiment prediction. The performance of the proposed architecture is benchmarked against existing research. Among individual models, LSTM and BiLSTM achieved the highest accuracy of 81 percent, while the proposed hybrid model attained an improved accuracy of 88 percent on the same dataset, demonstrating superior effectiveness in sentiment classification.

1. INTRODUCTION

Text Classification (TC) denotes the organized process of categorizing textual data into distinct groups based on their intrinsic characteristics[7] [2]. Through automated computational analysis, TC efficiently examines text and assigns it to predefined categories. This process plays a vital role in enabling the processing and extraction of meaningful insights from raw, unstructured textual information. TC systems are typically divided into three main types: rule-based, machine learning-based, and hybrid approaches. Rule-based methods utilize predetermined linguistic or pattern-based rules for categorization, while machine learning systems classify text using learned patterns derived from prior data. Hybrid systems integrate both approaches, leveraging rule-based logic and trained classifiers to enhance classification performance and reliability.

To improve category-level accuracy, increasingly advanced text classification techniques are required .

Corresponding Author: Megha Agarwal, IT Department, Dr. Shakuntala Misra National Rehabilitation University, Lucknow, UP-India
Email: meghaagarwal2011@gmail.com

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44007

Within this domain, Deep Learning (DL) models[5] have shown significant progress and effectiveness . Common DL architectures employed for text classification include Rule-Embedded Neural Networks (ReNNs), Multilayer Perceptrons (MLP), Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN). These models have demonstrated strong capabilities in identifying and organizing textual patterns with high precision. Moreover, word embedding techniques such as Word2Vec[6] and GloVe[13] have further enhanced classification performance across sentence, paragraph, and document levels .

Ongoing research continues to improve the performance of Natural Language Processing (NLP) tasks through the optimization of Deep Learning frameworks . Among these models, CNN, Long Short-Term Memory (LSTM), and Bidirectional LSTM (Bi-LSTM)[1][3][4] have achieved notable results, with reported accuracies reaching 77.4 percent and CNN delivering the most consistent average performance[9][21] . Continuous advancements in these frameworks have significantly improved both accuracy and efficiency in various NLP applications. Researchers have also introduced hybrid DL models to achieve even greater precision.

For instance, the hybrid Bayesian Network–RNN (BN-RNN) approach achieved metrics between 73.4 and 78.4 percent[8]. Similarly, the combined Long Short-Term Memory–Neural Network (LSTM-NN) model [26] reached 66.5 percent accuracy, with precision between 65.9 and 83.7 percent, recall between 79.9 and 81.1 percent, and F1-scores ranging from 71.6 to 80.0 percent. In contrast, the Bi-LSTM–MLP hybrid achieved a strong accuracy of 88.3 percent and an F1-score of 85.8 percent[13], confirming the model's robustness and its potential for advancing text classification performance.

Recent studies have explored the development of hybrid Deep Learning (DL) architectures to improve sentiment classification accuracy. Among these, the MTL-MSCNN-LSTM model—an integrated multi-task and multi-scale approach combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM)—demonstrated superior results, achieving accuracy levels between 82.88 and 88.50 percent[14] . These findings highlight the model's strong capability in accurately handling sentiment analysis across different tasks and data scales, reflecting its potential contribution to advancing sentiment classification research. Ayo et al. further validated this perspective through an extensive evaluation of hybrid models, using performance indicators such as Accuracy, Precision, Recall, and F1-Score[15]. Their analysis confirmed the reliability and effectiveness of hybrid frameworks in addressing diverse classification challenges and improving analytical outcomes.

Integrating CNN and Bidirectional LSTM (BiLSTM) into a single framework enhances both the breadth and depth of data processing[25][21]. CNN modules focus on capturing spatial attributes, while BiLSTM layers extract temporal and contextual dependencies . Studies employing CNN-BiLSTM architectures in conjunction with word embedding techniques like Word2Vec, GloVe, and FastText have achieved promising results[17] . Using an embedding dimension of 300, Softmax activation, dropout and recurrent dropout rates of 0.5 and 0.4, an SGD optimizer, 50 training epochs, 512 filters, a kernel size of 3, max-pooling, BiLSTM, and dense layers, researchers reported an accuracy of 82 percent . Despite these advances, limitations persist, including inefficiencies with long input sequences, redundant convolutional operations in NLP applications , and increased data loss with larger datasets that demand additional computational time and training data[1][3] .

The current research proposes a novel Hybrid CNN-BiLSTM architecture for sentiment analysis using a WhatsApp Group (WAG) dataset. This model integrates distinct feature extraction strategies and multiple convolutional kernel sizes to enhance classification performance. By exploiting CNN's strength in capturing local spatial patterns and BiLSTM's ability to model long-term contextual dependencies, the framework ensures a more comprehensive representation of text features. Unlike conventional LSTM, which relies solely on past information, BiLSTM processes both preceding and succeeding sequences, thereby enriching contextual understanding and improving predictive precision.

Moreover, the model employs kernels of different sizes to capture localized temporal relationships within sequential data effectively. For evaluation, text classification experiments were conducted using a confusion matrix as the primary assessment tool. Considering that WhatsApp messages differ slightly in structure and linguistic patterns from other social media corpora, the dataset underwent specialized preprocessing steps, sentiment labeling, and feature extraction procedures.

2. LITERATURE REVIEW

Several previous studies have examined text classification and sentiment analysis, following systematic methodologies for categorizing textual sentiments. Key contributions from the existing literature are summarized below:

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44007

The first study investigated sentiment analysis in the Thai language using deep learning (DL) models . Its primary objective was to enhance Thai sentiment classification by integrating word embeddings, part-of-speech (POS) tagging, and sentic features[9]. Sentic features play an important role by adding contextual information on polarity and subjectivity, thereby improving the semantic representation of textual sentiments. Three deep learning models were tested: the LSTM model achieved the highest F1-score of 0.726 using POS-tag embeddings and sentic features; the BiLSTM model, utilizing word embeddings with POS one-hot encoding and selected sentic features, attained an F1-score of 74.7 percent; and the CNN model achieved the best overall result with an F1-score of 81.7 percent when trained with word embeddings, POS-tags, and sentic features. The findings suggest that incorporating advanced features and deep learning models could further improve performance in future research.

Another relevant study focused on improving text classification accuracy through a Bi-LSTM model enhanced with Word2Vec embeddings, convolutional neural networks, and attention mechanisms . This model effectively combined multiple strengths: Word2Vec captured rich vector representations of words, CNN extracted spatial and local features, and the attention mechanism prioritized significant parts of the text. Using parameters such as a skip-gram embedding size of 300, a dropout rate of 0.2, a batch size of 128, and the Adam optimizer, the proposed framework achieved an average accuracy of 87.4 percent and an F1-score of 90.1 percent on a dataset containing 13,000 instances[13]. The results demonstrate that combining these techniques leads to more accurate and efficient text classification, especially when larger datasets and extended training durations are available.

A more recent work by Salur and Aydin[17] proposed a hybrid deep learning architecture for sentiment classification designed to handle dataset heterogeneity and class imbalance. Their approach combined CNN components for localized feature extraction with LSTM-based recurrent layers for capturing long-term dependencies. The model also incorporated both character-level and pre-trained embeddings, including Word2Vec, GloVe, and FastText. Using a dataset of tweets related to a Turkish GSM operator, the BiLSTM model with FastText achieved 80.44 percent accuracy, whereas the hybrid CNN-BiLSTM model that included character embeddings improved accuracy to 82.14 percent. These results highlight the hybrid model's superior capability in sentiment classification, particularly for morphologically rich languages such as Turkish, Arabic, and Lithuanian.

In addition, Naqvi et al. conducted sentiment analysis on Urdu text using deep learning frameworks[18]. Their methodology combined CNN for capturing local text features with RNN and LSTM architectures to learn long-term contextual information. Four embeddings—Samar, CoNLL, pretrained, and self-trained FastText—were analyzed for performance comparison. The BiLSTM-ATT model obtained the highest overall accuracy of 77.9 percent, while the LSTM model achieved a maximum precision of 85.16 percent using Samar embeddings. These findings confirm the strength of deep learning approaches in processing sentiment-rich languages like Urdu.

The collective insights from these studies form a strong foundation for ongoing research in sentiment analysis and text classification. They demonstrate how hybrid and deep learning architectures continue to evolve, offering greater adaptability and accuracy across different languages, datasets, and application domains.

A hybrid CNN-BiLSTM architecture was designed employing diverse feature extraction techniques and variable kernel configurations for analyzing the WAG dataset. This model represents an innovative framework that integrates the strengths of both CNN and BiLSTM networks. While CNN layers are adept at capturing localized features within text data, the BiLSTM component effectively models long-term dependencies by leveraging both preceding and succeeding contextual information, thereby improving overall predictive accuracy. To handle variations in input sequence length, padding techniques were applied to ensure consistent data representation. The integration of Rectified Linear Unit (ReLU) and Softmax activation functions further enhanced the model's adaptability and non-linear learning capability. Using an embedding dimension of 300 allowed the model to identify complex and fine-grained semantic relationships within the textual inputs.

The proposed design for sequential data analysis within the WAG dataset utilizes four LSTM layers, with 64 units in the

final layer to optimize learning efficiency. Additionally, convolutional filters of varying kernel sizes were incorporated to

capture distinct local dependencies across sequential features. The primary contribution of this research lies in the

formulation of a hybrid CNN-BiLSTM model that synthesizes multiple feature extraction strategies and kernel configurations, thereby substantially improving the performance and interpretability of sentiment analysis on

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44007

sequential
social media text data.

MATERIALS AND METHODS

This study aims to enhance the performance of the hybrid CNN BiLSTM model, specifically for sentiment analysis in the TC domain. We evaluated this hybrid model thoroughly by considering the advantages and potentials of both the CNN and BiLSTM architectures. The methodology we adopted for this study focuses on crucial aspects, such as selecting appropriate data, performing meticulous labeling, developing robust feature vectors, and formulating the hybrid CNN BiLSTM approach. All these aspects contribute to a more precise and accurate sentiment analysis solution. We applied the proposed model to effectively predict the sentiment polarity of textual data and subsequently classify it based on the determined polarity

A. Dataset

We obtained the research dataset from the "Forum DTC Riau" WhatsApp group, a community comprising 134 owners of Daihatsu Taruna cars in the Riau region[19][20]. We chose to use the "Forum DTC Riau" dataset (DTC) because it aligns with our research objectives and specific geographical characteristics. This dataset helps us understand unique preferences, issues, and perspectives within the region, which could potentially influence user sentiments and viewpoints. For comparison, we also tested the model with the Amazon Product Summaries dataset from Kaggle (PS), which consists of ten thousand data record. We followed similar preprocessing and model testing procedures for this dataset

This WAG was established on 10/11/2018, and the conversation data collected for analysis spanned from 16/3/2023 to 15/3/2023. The data extraction process involved using an OPPO A15 smartphone equipped with an Android Version 10 operating system. The smartphone employed an octa-core processor with 3 Gigabyte RAM capacity to facilitate the data extraction procedure. All conversation data from the WAG were exported in text file (txt) format and subsequently transmitted directly via email.

The text file containing the WhatsApp group data is unstructured data. It contains information about encrypted messages and calls exchanged between the group members, as well as the creation time of the group. This data needs to be readable and understandable by machines (computers), for which NLP is used to comprehend, classify, and extract opinions, sentiments, and emotions from natural language texts or data [19]. The NLP performed in this dissertation consists of two stages: data labelling and text preprocessing.

In this study, the framework and methodology include several main stages: labeling, preprocessing, feature extraction, and the use of the proposed model. The labeling stage involves assigning labels or classifications to the data used, which facilitates the machine-learning process. Next, a preprocessing stage was conducted to clean and prepare the data for further analysis. Subsequently, the feature extraction stage is performed to identify and extract important features from the data, which can aid in modeling and further analysis. Finally, the use of the proposed model involves implementation and testing of the developed model with the objective of achieving the desired results. By following these stages, this study aims to produce a comprehensive and accurate analysis based on carefully prepared and validated data

B. Labeling

The unstructured dataset was derived from text file exports of the WhatsApp Group (WAG) . This dataset includes information such as the group's creation date, details about encrypted messages and calls exchanged among members, and a variety of meta elements including time stamps, contact numbers, member names, emojis, and media placeholders represented as "<Media omitted>," indicating image or multimedia message traces. As the extracted messages were not pre labeled with sentiment values, an emotion based sentiment classification method using the SentiWordNet emotional lexicon was adopted to determine sentiment polarity[5]. The sentiment labeling and data preparation process consisted of the following sequential steps:

The initial stage involved reading and structuring the data using regular expressions. Each line of the unstructured dataset was parsed, and elements were separated using commas as delimiters. The split function was applied to isolate the first item in each group, after which tokenization was performed to extract and organize date, time, author, and message segments. Message contents often included text, emojis, and URLs, which were subsequently categorized into three main components: "Message," "Emoji," and "URL Count." This parsing process utilized commas, hyphens, colons, and spaces as delimiters to ensure accurate token retrieval. The processed data were then organized and managed within

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44007

a pandas DataFrame for further analysis[24].

The next stage involved sentiment labeling focused on the message column. Words within each message were tokenized, and irrelevant terms were removed using the Natural Language Toolkit (NLTK) and Sastrawi libraries. A new column, titled "Message_English," was created to store English translated content of the messages, which was generated using the Google TransTranslator package.

For efficient translation and processing, the dataset was divided into batches of 250 rows. Sentiment features were then derived using the nltk.sentiment.vader module, which produced four key metrics: positive, negative, neutral, and compound scores. Based on the compound score, sentiments were categorized according to the following criteria: a compound value of 0.05 or higher was classified as Positive, −0.05 or lower as Negative, and values between −0.05 and 0.05 as Neutral. The sentiment labeled dataset was saved as "DTCRiau_sentimen.csv" for future experimentation and evaluation.

The labeling of the Amazon Product Summaries (PS) dataset followed a simplified rule based approach. Using the "Score" column, reviews with a score greater than 3 were labeled as positive, those equal to 3 as neutral, and those below 3 as negative.

Through these sequential preprocessing and labeling steps, both datasets were standardized to enable consistent sentiment classification and model training for further analysis.

C. Preprocessing

Text preprocessing is a crucial phase in text classification that involves a series of techniques to prepare and transform textual data for computational analysis. It plays an important role across various domains and languages . Textual information transferred by humans must be converted into a machine-interpretable format before analysis . This stage helps mitigate irregularities or noise that may affect subsequent stages of data processing. Since text data used for classification is often imperfect, preprocessing ensures that labeled data are clean and consistent before model training[5][6].

In this study, several preprocessing steps were performed to improve data quality and ensure readiness for classification. These steps included removing HTML tags and URLs embedded in messages to obtain clearer text content. Negation words were replaced with their antonyms to better capture sentiment nuances and improve interpretability. Neutral sentiment entries were excluded so the dataset focused solely on positive and negative categories. Punctuation marks were eliminated to minimize their influence on model interpretation[7]. Finally, lemmatization was applied to reduce words to their base forms, promoting consistency and linguistic normalization throughout the dataset.

D. Feature Extraction

A key challenge in Natural Language Processing (NLP) lies in developing models capable of comprehending the hierarchical structure of textual information, particularly in classification and feature extraction tasks . Feature extraction seeks to identify key attributes or patterns within the data that contribute to accurate classification. This process helps reduce data dimensionality, creating a more compact representation that facilitates efficient analysis while retaining meaningful information. By selecting and combining the most relevant variables, the model effectively captures the essence of the textual input with reduced computational complexity[23].

In this research, multiple feature extraction techniques were applied. Encoding methods were employed to convert categorical or nominal data into numerical form, ensuring compatibility with machine learning algorithms. Tokenization was carried out to segment sentences into smaller units such as words or tokens. Padding was then used to standardize input sequence lengths, enabling consistent dimensionality across training and testing sets.

E. Proposed Model

After sentiment labeling and data preparation, the datasets were evaluated using individual deep learning algorithms, including Neural Network, Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and Convolutional Neural Network (CNN). Prior to testing, data were thoroughly cleaned by removing redundant elements such as empty entries, URLs, emojis, and punctuation. Feature extraction followed, utilizing tokenization and encoding processes to represent textual data in a structured numeric format.

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44007

Subsequently, the cleaned datasets were partitioned into training and testing subsets. Padding procedures were applied to standardize sequence lengths, facilitating alignment for further analysis and model fitting. The performance of these individual models was then compared with that of the proposed hybrid model.

The hybrid CNN-BiLSTM model was developed following the processes of labeling, preprocessing, and feature extraction as described earlier. The model adopted an embedding strategy in which textual content was represented as low-dimensional numerical vectors, with each vector having a padding size of 300. This configuration, as detailed in Table I, differs from prior implementations and contributes to enhanced model performance in sentiment classification tasks.

TABLE I. HYPERPARAMETER OF THE PROPOSED HYBRID MODEL "SEQUENTIAL"

Layer (type)	Output Shape	Param #
Embedding	(None, 100, 300)	7,500,000
Conv1D	(None, 98, 200)	180,200
Bidirectional	(None, 98, 128)	135,680
Dropout	(None, 98, 128)	0
Bidirectional	(None, 128)	98,816
Dense	(None, 50)	6,450
Dense	(None, 50)	2,550
Flatten	(None, 50)	0
Dense	(None, 100)	5,100
Dense	(None, 2)	202

The proposed model's architecture was compared with existing studies, as summarized in Table II. In each referenced work, model configurations were adapted to align with the characteristics and requirements of the corresponding datasets. For instance, Jang et al. implemented a hybrid CNN-BiLSTM model integrated with the Word2Vec Skip-Gram embedding method to analyze sentiment in clothing and camera product reviews. Salur and Aydin examined Twitter data generated by users of a Turkish GSM operator, employing a CNN-BiLSTM model that incorporated character-level embeddings combined with FastText representations. Similarly, Soumya and Pramod performed sentiment analysis on Malayalam tweets using a hybrid model that fused CNN-BiLSTM and CNN-LSTM components.

The hybrid architecture presented in this research integrates CNN and BiLSTM layers to leverage the distinct advantages of each. CNN layers, equipped with filters and kernels, are responsible for extracting spatial and local textual features, whereas the BiLSTM layers capture long-term contextual dependencies within the text sequence, understanding meaning derived from both past and future contexts. To mitigate overfitting, a dropout layer was introduced following the initial BiLSTM layer. This mechanism randomly deactivates selected neurons during training, reducing dependency on specific features and enhancing generalization. The dropout rate was maintained at 0 to preserve essential neuron activity while preventing the model from overfitting the training set. Additionally, a dense layer with a reduced number of neurons following the BiLSTM layer was included to further constrain the model's capacity and control overfitting.
The combined CNN-BiLSTM structure enables comprehensive and multi-level feature processing, where CNN captures spatial characteristics and BiLSTM handles temporal or sequential features . The overall model development begins with labeling, preprocessing, and feature extraction phases described in earlier sections. The embedding process transforms the input text into low-dimensional numeric vectors with a padding size of 300, allowing consistent input representation. This embedding configuration differs from those implemented in previous works , as detailed in Table II, and contributes to improved adaptability and feature learning in sentiment classification tasks.

TABLE II. PERFORMANCE OF THE EXISTING DL MODEL

Model	Embedding	F1-Score	Accuracy
CNN LSTM CNN Bi LSTM [9]	POS Tagging, sentic vevtor	81.7%	77.4%
CNN BiLSTM [13]	Word2vec Skip Gram	88.0%	87.6%
CNN BiLSTM [17]	Carakter +Fastext	89.0%	82.1%

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44007

CNN BiLSTM, CNN LSTM [26]	Sentiment Tagging, Wordvector	75.0%	85.5%

RESULT AND DISCUSSION

Sentiment labeling was performed using the VADER Sentiment library, which classified the text data into three categories: positive, neutral, and negative. For the objectives of this research, however, only positive and negative sentiments were retained for analysis. The original dataset comprised 5,237 entries. After removing data labeled as neutral, the dataset was reduced to 1,089 records. Further refinement was carried out through the elimination of stopwords to enhance the quality of the remaining data. Following preprocessing, the positive and negative samples were numerically encoded using an encoder, generating new columns to represent sentiment labels. The dataset was then partitioned into training and testing subsets to facilitate performance evaluation. The training set contained 80 percent of the data, while the testing set comprised the remaining 20 percent. The distribution and statistical representation of the partitioned data, utilizing a random_state value of 69, are detailed in Table III.

TABLE III. SPLITTING OF TRAINING DATASET AND TESTING DATASET

Feature	Training Dataset		Testing Dataset	
	DTC	PS	DTC	PS
Length	1,552	7,311	389	1,828
Dentences shape	1,552.0	7,311.0	389.0	1,828.0
Labels shape	1,552.2	7,311.2	389.2	1,828.2

The hybrid model integrates Convolutional Neural Network (CNN) layers for textual feature extraction with Bidirectional Long Short-Term Memory (BiLSTM) layers to capture contextual relationships within sentences. The CNN component, equipped with filters and kernels, identifies key spatial patterns and local dependencies in the text. In contrast, the BiLSTM component processes information bidirectionally, allowing the model to interpret semantic relationships based on both preceding and succeeding word contexts. Non-linearity and overfitting control are achieved through the use of activation functions—specifically the Rectified Linear Unit (ReLU)—and a dropout mechanism that randomly deactivates neurons during model training to improve generalization.

The model employs two activation functions, ReLU and Softmax, with the dropout rate set to 0.5 to minimize overfitting. The Adam optimizer is used for parameter optimization, and training is conducted over 20 epochs. The proposed architecture comprises an embedding layer, a CNN layer, a BiLSTM layer, and four dense layers, as outlined in Table I. The model's overall performance is evaluated using the F1-Score, a comprehensive metric that represents the harmonic mean of precision and recall, providing a balanced measure of classification accuracy.

The proposed model employs multiple filter sizes to capture diverse local dependencies within the textual data, thereby improving the efficiency and depth of feature extraction. The architecture of the developed hybrid CNN-BiLSTM model is llustrated in Fig. 1. To evaluate its effectiveness, comprehensive performance assessments were conducted on both individual deep learning (DL) models and the hybrid DL architecture.

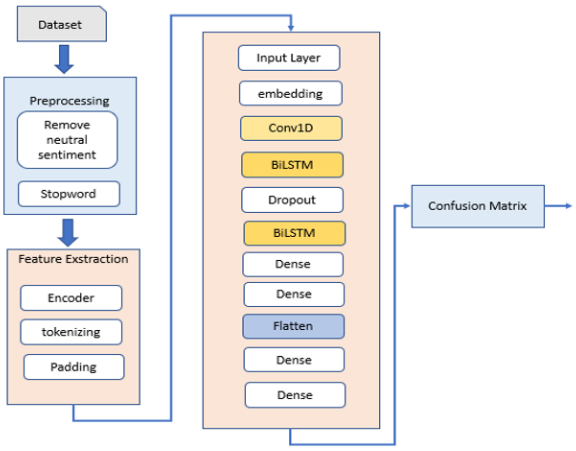

Fig. 1. The proposed architecture of the hybrid model.

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44007

The categorization results are given in the form of a confusion matrix. The model evaluation also gives accuracy, precision, recall, and F1-Score. These results are compared to previous research using the same method to produce better results using slightly different methods or phases. The following performance indicators are used in this work to assess the efficacy of the proposed single-DL and hybrid CNN BiLSTM models.

The PS dataset exhibited a larger sample size in both training and testing partitions, providing stronger representational capacity and diversity. In contrast, the DTC dataset contained fewer samples, enabling faster model training but with reduced capability to generalize complex sentiment patterns (see Table III). Analysis of dataset partitioning results also revealed variations in sentiment distribution: the PS dataset was dominated by positive sentiments, whereas the DTC dataset contained a higher proportion of negative sentiments (see Fig. 2).

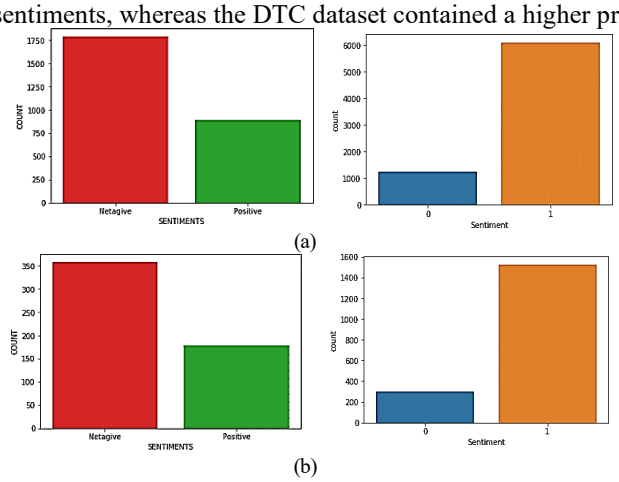

Fig. 2. Sentiment distribution of DTC and PS datasets; (a) Training dataset (b) Testing dataset.

Following the selection of the designated testing and training dataset, further preprocessing steps were employed to facilitate sub sequent analysis. The processing involved tokenization and data padding utilizing specific features, namely vocab_size = 25000, embedding_dim = 300, max_length = 100, trunc_type=post, and oov_tok = "<OOV>". These features were utilized to ensure efficient representation and standardization of the data for subsequent modeling and analysis purposes.

Precision measures the proportion of correctly predicted positive samples relative to all samples classified as positive. The datasets utilized in this research varied considerably in both size and structure. The DTC dataset consisted of 871 training samples and 218 testing samples, while the PS dataset was substantially larger, comprising 7,311 training and 1,828 testing samples. To ensure uniform input representation, both datasets were standardized with a padding length of 100. The larger scale of the PS dataset allowed for improved pattern recognition and stronger model generalization, whereas the smaller DTC dataset facilitated quicker model training but offered limited capacity for capturing complex contextual relationships.

For sentiment classification, the labeled datasets were analyzed using several standalone deep learning models, including a Neural Network, Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and Convolutional Neural Network (CNN). The evaluation process incorporated modifications in the preprocessing and feature extraction phases to assess each model's efficiency. The preprocessing stage involved data cleaning, removal of neutral sentiments, and elimination of stopwords. Feature extraction processes included label encoding, data partitioning, tokenization, and padding. All models were trained and validated using an 80:20 data split to ensure balanced performance evaluation.

TABLE IV. RESULTS DETAILS OF THE SINGLE DL MODEL AND THE PROPOSED MODEL

Dataset	Model	Accuracy	Precision	Recall	F1-Score
DTC	CNN	0.7%	0.68%	0.68%	0.68%
	Simple RNN	0.75%	0.77%	0.73%	0.74%
	LSTM	0.81%	0.83%	0.80%	0.81%
	BiLTSM	0.81%	0.85%	0.79%	0.80%
	Proposed Model	**0.8%**	**0.76%**	**0.82%**	**0.79%**
PS	CNN	0.75%	0.33%	0.25%	0.29%
	Simple RNN	0.80%	0.49%	0.59%	0.51%
	LSTM	0.81%	0.47%	0.50%	0.47%
	BiLTSM	0.82%	0.53%	0.49%	0.51%
	Proposed Model	**0.88%**	**0.76%**	**0.79%**	**0.78%**

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44007

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44007

The evaluation was further extended using the proposed hybrid CNN-BiLSTM model, which outperformed all other tested models by achieving the highest accuracy of 0.88 across both datasets. However, its precision value was slightly lower, at 0.76, compared with the LSTM and BiLSTM models. For the DTC dataset, the model's precision remained at 0.76, whereas the PS dataset demonstrated notably strong results, with a high accuracy and favorable recall and F1-Scores of approximately 0.79 and 0.78, respectively. These outcomes validate the hybrid model's ability to accurately classify sentiment across diverse data sources.

During feature extraction, processes such as label encoding, data partitioning, tokenization, and padding were applied. Each deep learning model was evaluated using an 80:20 data split. Among the single architectures, the BiLSTM model attained the highest accuracy of 0.81, as indicated in Table IV, using 871 training and 218 testing samples from the "Forum DTC Riau" WhatsApp group dataset. Performance improvements were subsequently observed when testing was conducted with the proposed hybrid model.

The hybrid CNN-BiLSTM model, adapted from prior research , incorporated modifications in both the preprocessing and feature extraction stages. Its architecture also introduced variations in activation functions, dropout configuration, filters, and a kernel size of 3, as detailed in Table II and illustrated in Fig. 1. The network was trained for 20 epochs with a batch size of 256. Throughout the training and validation phases, the model's accuracy and loss metrics were monitored and visualized to assess learning progression (see Fig. 3).
For the DTC dataset, the hybrid model initially recorded an accuracy of approximately 0.7382 in epoch 1, which increased to 0.7681 with a validation loss of about 1.2490. A significant accuracy jump to 0.9882 was observed at epoch 7, reaching approximately 0.9977 between epochs 8 and 10. From epochs 11 to 20, accuracy stabilized with minimal fluctuation, indicating model convergence. The final validation accuracy reached approximately 0.8716 (refer to Fig. 3, DTC).
For the PS dataset, the training began with an initial accuracy of 0.8210 and a validation loss of 0.9867. Accuracy improved to 0.8356 by the second epoch, with a corresponding drop in validation loss to 0.7486. Consistent performance gains were observed between epochs 3 and 5, yielding an accuracy of approximately 0.9948 and a validation loss of 0.6281. Beyond epoch 6, both accuracy and validation loss stabilized with minor variations, indicating steady convergence. The final validation accuracy was recorded at approximately 0.8813 (refer to Fig. 3, PS).
These results collectively demonstrate that the proposed hybrid CNN-BiLSTM model provides enhanced accuracy and stability throughout the training phases across both datasets while maintaining robust generalization performance.

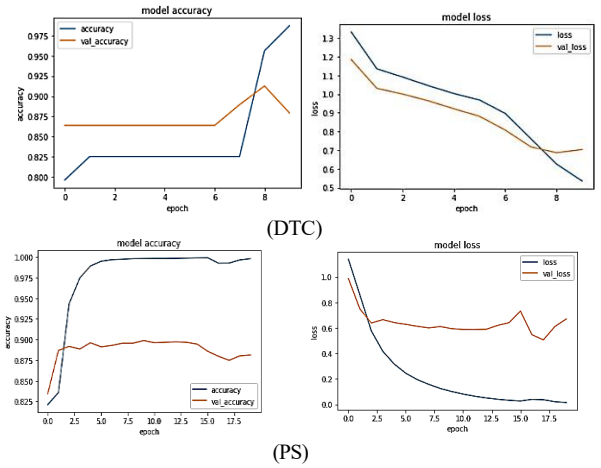

Fig. 3. Accuracy vs. epochs, loss vs epochs plot obtained from the proposed model.

The model's overall performance during both training and testing phases is effectively illustrated in the visual results. Upon completion of training, the proposed hybrid CNN-BiLSTM model achieved a notable test accuracy of 98 percent. Evaluation of the trained model on the testing dataset generated a confusion matrix, as shown in Fig. 4. The hybrid architecture demonstrated the highest accuracy on the PS dataset (0.8813) and a slightly lower accuracy on the DTC dataset (0.8716), suggesting a potential saturation or learning plateau. The variation in convergence rates between the two datasets can be attributed to differences in their structural and linguistic characteristics.

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44007

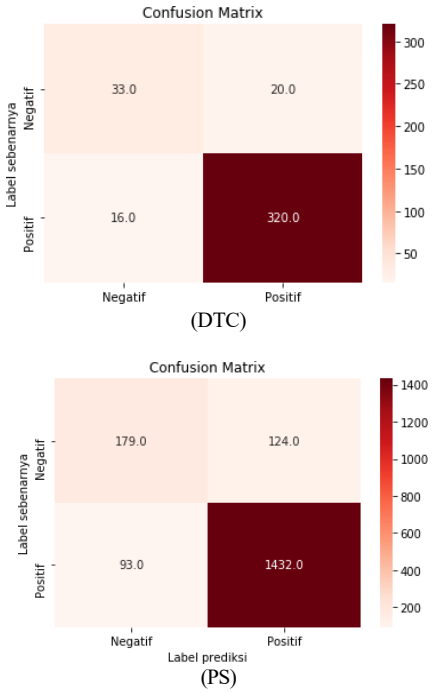

Fig. 4. Confusion matric for the proposed model.

Based on the obtained test results, among the individual deep learning (DL) models, the CNN and BiLSTM architectures demonstrated superior accuracy across the evaluated datasets, as presented in Table IV. The BiLSTM model achieved the highest accuracy, reaching approximately 0.81 on the DTC dataset and 0.82 on the PS dataset. Precision, recall, and F1-scores on the PS dataset were lower compared to those observed for the DTC dataset, with precision values of 0.53 (PS) and 0.85 (DTC), recall values of 0.49 (PS) and 0.79 (DTC), and F1-scores of 0.51 (PS) and 0.80 (DTC). The LSTM model achieved similar accuracy levels, approximately 0.81 on both datasets, with comparable precision, recall, and F1-scores.

The Simple RNN model recorded an accuracy of approximately 0.75 on the DTC dataset and 0.80 on the PS dataset. However, the recall performance was higher for the PS dataset (0.59) than for the DTC dataset (0.73). In contrast, the CNN model produced the lowest results overall, with accuracies of 0.70 on the DTC dataset and 0.75 on the PS dataset. Correspondingly, the precision, recall, and F1-scores of CNN were lower than those of the other models. These outcomes demonstrate better performance compared to the findings reported in , which achieved an accuracy of 77.4 percent.

The hybrid CNN-BiLSTM model exhibited the most promising results, achieving a high accuracy of 0.88 on both DTC and PS datasets, confirming its strength in sentiment classification tasks. Despite its precision value of 0.76—suggesting the potential for false positives—the model achieved robust recall scores of 0.82 (DTC) and 0.79 (PS). The corresponding F1-scores, 0.79 for DTC and 0.78 for PS, indicate a strong balance between precision and recall, highlighting the model's stability and effectiveness. Across all experiments, the hybrid architecture consistently outperformed the individual DL models, as shown in Table IV, and exceeded the performance of models proposed in earlier studies .

Model performance improvements can largely be attributed to the optimization and regularization strategies implemented. Specifically, the adoption of the Adam optimizer with a learning rate of 0.0001, coupled with dropout and L2 regularization techniques, helped control overfitting and enhance model generalization. The inclusion of a dropout layer proved particularly beneficial, as it reduced dependency on specific neural features and improved learning robustness.

Nevertheless, some evidence of overfitting was observed, as the model performed markedly better during training compared to testing. This discrepancy may be due to truncated or untranslated text messages that limited contextual information, thereby reducing labeling accuracy in the lexicon-based sentiment classification stage. Such limitations introduced bias, leading to cases where highly positive terms appeared within negatively labeled samples, making it difficult for the model to generalize unseen data. Mislabelled or ambiguous training examples further contributed to inconsistencies, as they did not adequately capture the intended sentiment patterns. Another possible factor is sub-optimal tuning of hyperparameters, such as dropout rate, embedding dimension, or filter configuration, which could have influenced model convergence and performance.

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44007

CONCLUSION

This research introduces a specialized hybrid CNN-BiLSTM model specifically designed for sentiment analysis of WhatsApp Group (WAG) data. By integrating the complementary strengths of Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks, the model effectively captures both short-term dependencies and long-range contextual relationships within sequential text data. The architecture incorporates comprehensive preprocessing and feature extraction stages, along with a hybrid design featuring activation functions, dropout regularization, multiple filters, and a kernel size of 3 across several layers. Experimental evaluations demonstrate that the proposed CNN-BiLSTM model achieves an outstanding accuracy of 88 percent. Compared to the standalone BiLSTM model, this approach delivers a performance enhancement of seven percentage points, underscoring its efficiency and robustness in analyzing WAG-based sentiment data.

REFERENCES

[1] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, A. Mehmood, and M. T. Sadiq, "Document-level text classification using single-layer multisize filters convolutional neural network," *IEEE Access*, vol. 8, no. Ml, pp. 42689–42707, 2020. doi: 10.1109/ACCESS.2020.2976744

[2] A. Wahdan, S. Hantoobi, S. A. Salloum, and K. Shaalan, "A systematic review of text classification research based on deep learning models in Arabic language," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 6, pp. 6629–6643, 2020. doi: 10.11591/IJECE.V10I6.PP6629-6643

[3] W. Fang, H. Luo, S. Xu, P. E. D. Love, Z. Lu, and C. Ye, "Automated text classification of near-misses from safety reports: An improved deep learning approach," *Adv. Eng. Informatics*, vol. 44, no. March 2019, 101060, 2020. doi: 10.1016/j.aei.2020.101060

[4] Z. Liu, C. Lu, H. Huang, S. Lyu, and Z. Tao, "Hierarchical Multi- granularity attention- based hybrid neural network for text classification," *IEEE Access*, vol. 8, pp. 149362–149371, 2020. doi: 10.1109/ACCESS.2020.3016727

[5] H. Yang, L. Luo, L. P. Chueng, D. Ling, and F. Chin, "Deep learning and its applications to natural language processing," in *Deep Learning: Fundamentals, Theory and Applications*, 2019, pp. 89–109.

[6] R. Joshi, P. Goel, and R. Joshi, "Deep learning for hindi text classification: A comparison," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*, 2020, pp. 94–101. doi: 10.1007/978-3-030-44689-5_9

[7] Q. Li *et al.*, "A survey on text classification: From shallow to deep learning," *IEEE Trans. NEURAL NETWORKS Learn. Syst.*, vol. 31, no. 11, pp. 1–21, 2020.

[8] F. Zaman, M. Shardlow, S. Hassan, and N. Radi, "HTSS : A novel hybrid text summarisation and simplification architecture," *Inf. Process. Manag.*, vol. 57, no. 6, 102351, 2020. doi: 10.1016/j.ipm.2020.102351

[9] K. Pasupa, T. Seneewong, and N. Ayutthaya, "Thai sentiment analysis with deep learning techniques: A comparative study based on word embedding , POS-tag , and sentic features," *Sustain. Cities Soc.*, vol. 50, no. 7, 101615, 2019. doi: 10.1016/j.scs.2019.101615

[10] K. Miok, D. Nguyen-Doan, B. Škrlj, D. Zaharie, and M. Robnik- Šikonja, "Prediction uncertainty estimation for hate speech classification," *Statistical Language and Speech Processing*, pp. 286–298, 2019.

[11] H. Faris, I. Aljarah, M. Habib, and P. A. Castillo, "Hate speech detection using word embedding and deep learning in the arabic language context," in *Proc. 9th International Conference on Pattern Recognition Applications andMethods (ICPRAM 2020)*, 2022, pp. 453–460. doi: 10.5220/0008954004530460

[12] A. Garain, "The titans at semeval-2019 task 6: Offensive language identification, categorization and target identification," in *Proc. 13th International Workshop on Semantic Evaluation (SemEval- 2019)*, 2019, pp. 759–762.

[13] B. Jang, M. Kim, G. Harerimana, S. Kang, and J. W. Kim, "Applied sciences Bi-LSTM Model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism," *Appl. Sci.*, vol. 10, no. 17, 5841, 2020.

[14] N. Jin, J. Wu, X. Ma, K. Yan, and Y. Mo, "Multi-task learning model based on multi-scale CNN and LSTM for sentiment classification," *IEEE Access*, vol. 8, pp. 77060–77072, 2020. doi: 10.1109/ACCESS.2020.2989428

[15] F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, "Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions," *Comput. Sci. Rev.*, vol. 38, 100311, 2020. doi: 10.1016/j.cosrev.2020.100311

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44007

[16] S. Kumar, C. Akhilesh, K. Vijay, and B. Semwal, "A multibranch CNN-BiLSTM model for human activity recognition using wearable sensor data," *Vis. Comput.*, vol. 38, no. 12, pp. 4095–4109, 2021. doi: 10.1007/s00371-021-02283-3

[17] M. U. Salur and I. Aydin, "A novel hybrid deep learning model for sentiment classification," *IEEE Access*, vol. 8, pp. 58080–58093, 2020. doi: 10.1109/ACCESS.2020.2982538

[18] U. Naqvi, A. Majid, and S. A. L. I. Abbas, "UTSA : Urdu text sentiment analysis using deep learning methods," *IEEE Access*, vol. 9, pp. 114085–114094, 2021. doi: 10.1109/ACCESS.2021.3104308

[19] J. Gaglani, Y. Gandhi, S. Gogate, and A. Halbe, "Unsupervised whatsapp fake news detection using semantic search," in *Proc. International Conference on Intelligent Computing and Control Systems, ICICCS 2020*, 2020, pp. 285–289. doi: 10.1109/ICICCS48265.2020.9120902

[20] H. T. Assaggaf, "A discursive and pragmatic analysis of whatsapp text-based status notifications," *Arab World English J.*, vol. 10, no. 4, pp. 101–111, 2019. doi: 10.24093/awej/vol10no4.8

[21] Y. Zhou, Q. Zhang, D. Wang, and X. Gu, "Text sentiment analysis based on a new hybrid network model," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–15, 2022.

[22] B. S. Rintyarna, R. Sarno, and C. Fatichah, "Evaluating the performance of sentence level features and domain sensitive features of product reviews on supervised sentiment analysis tasks," *J. Big Data*, vol. 6, no. 1, 2019. doi: 10.1186/s40537-019-0246-8

[23] H. Aljuaid, R. Iftikhar, S. Ahmad, M. Asif, and M. Tanvir Afzal, "Important citation identification using sentiment analysis of in-text citations," *Telemat. Informatics*, vol. 56, 101492, 2021. doi: 10.1016/j.tele.2020.101492.

[24] N. Chintalapudi, G. Battineni, M. Di Canio, G. G. Sagaro, and F. Amenta, "Text mining with sentiment analysis on seafarers' medical documents," *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 1, 100005, 2021. doi: 10.1016/j.jjimei.2020.100005

[25] Esha Srivastava et al., AI-Driven Predictive Analytics with the Help of IoT for Organizational Change Management, TEJAS Journal of Technologies and Humanitarian Science, ISSN : 2583-5599, V. 04, I.03, July-2025, https://doi.org/10.63920/tjths.43001

[26] A. U. Rehman, A. K. Malik, B. Raza, and W. Ali, "A hybrid CNN- LSTM model for improving accuracy of movie reviews sentiment analysis," *Multimed. Tools Appl.*, vol. 78, no. 18, pp. 26597–26613, 2019. doi: 10.1007/s11042-019-07788-7

[27] S. Soumya and K. V Pramod, "Hybrid deep learning approach for sentiment classification of malayalam tweets," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 4, pp. 891–899, 2022.