Cross-Platform Hate Speech Detection Using an Attention-Enhanced BiLSTM Model

Muzammil Hussain
Department of Software Engineering, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, Jordan
m.hussain@ammanu.edu.jo

Waqas Sharif
Department of Computer Science, The Islamia University of Bahawalpur, Bahawalpur, Pakistan
waqas.sharif@iub.edu.pk

Muhammad Rehan Faheem
Fakulti Kecerdasan Buatan dan Keselamatan Siber, Universiti Teknikal Malaysia Melaka, 76100 Melaka, Malaysia
rehan@utem.edu.my (corresponding author)

Yazeed Alsarhan
Faculty of Information Technology, Al-Ahliyya Amman University, Amman, Jordan
y.alsarhan@ammanu.edu.jo

Hany A. Elsalamony
Department of Computer Science, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, Jordan | Department of Mathematics, Faculty of Science, Helwan University, Cairo, Egypt
h.salamony@ammanu.edu.jo

Received: 9 July 2025 | Revised: 13 August 2025, 5 September 2025, 17 September 2025, and 27 September 2025 | Accepted: 28 September 2025

 DOI: https://doi.org/10.48084/etasr.13249

ABSTRACT

Hate speech is rapidly spreading across digital platforms, appearing in diverse forms driven by regional, cultural, and linguistic differences. This growing trend presents serious challenges to social harmony and online safety. Existing hate speech detection models often fall short because they rely on limited and homogeneous datasets, making them less effective in real-world, culturally diverse settings. Handling large-scale, diverse datasets adds notable complexity to capturing contextual nuances, as different populations and cultures demonstrate unique language patterns and expressions. This study addresses the necessity for a more universal solution by proposing a deep learning model trained on an extensive and diverse dataset comprising 0.842 million samples collected from various digital platforms. The approach combines a Bidirectional Long Short-Term Memory (BiLSTM) model with a self-attention mechanism to capture contextual depth. Various data embedding techniques were used to assess their impact, along with data resampling and standard Natural Language Processing (NLP) pre-processing steps. The proposed model achieved 0.93 accuracy with an F1-score of 0.92, outperforming several baseline and state-of-the-art models. This work provides a comprehensive and scalable framework for the detection of hate speech across various online platforms.

Keywords-hate speech detection; NLP; deep learning; BiLSTM; SMOTE

I. INTRODUCTION

Hate speech refers to any form of communication that targets individuals or groups based on characteristics such as gender, caste, ethnicity, religion, race or nationality [1]. Social media has become one of the most popular tools for communication, enabling millions of users to share their thoughts and connect globally. Figure 1 shows the active users across different social media platforms [2]. This exponential growth reflects society's increasing dependence on digital platforms for communication, entertainment, and information dissemination. This wide reach and easy access have also

turned it into a common space for spreading harmful and offensive content. The freedom to express opinions is often misused to post messages that attack others based on identity.

Many surveys have shown that online hate is on the rise, with approximately 41% of Americans reporting some form of online harassment by 2020 [3]. Several other incidents, such as the attacks in Myanmar, Charlottesville, and Christchurch, indicated that online hate speech can have serious offline consequences [4]. The persistent presence of this kind of content undermines the credibility of digital platforms and erodes public trust. It also presents challenges for promoting online communities that are genuinely inclusive and respectful.

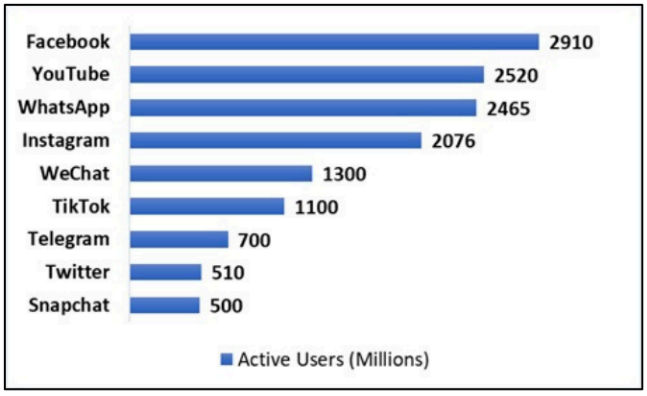

Fig. 1. Active number of users on social media platforms.

Given the huge and ongoing volume of user-generated content on social media platforms, the rapid proliferation of hate speech on digital platforms has underscored the need for scalable automated detection systems. These systems have the potential to identify hate speech, which is still a challenge due to the linguistic diversity of online communication patterns, with social media users often using informal language structures, abbreviations, colloquial expressions, and code switching between languages. Many users disseminate hateful content through sarcasm, coded language, or implicit discriminatory remarks that rule-based approaches cannot detect [5].

However, the majority of studies rely on single-source or limited multi-source dataset combinations that limit the ability of a model to handle linguistic diversity, contextual nuances, and cultural communication patterns. A comprehensive detection model trained on a wide multi-source dataset collected from multiple online platforms and communities was developed in this study to effectively address these limitations to work well in different online settings and to improve the generalization of hate speech detection systems. The key contributions of this study are:

- This study uses a large and diverse dataset consisting of 0.842 million samples collected from 18 publicly available sources. The dataset covers different digital platforms, regions, and linguistic contexts, which helps improve the generalization of the model.
- A BiLSTM-based model is proposed with a self-attention mechanism to capture contextual patterns in hate speech more effectively.
- Various word embedding techniques are applied to evaluate their effect on model performance.
- Standard NLP pre-processing and data resampling methods are used to improve the quality and balance of the data during training.

Hate speech detection has gained more focus in recent years due to the increase in harmful content on digital platforms. Authors in [6] investigated multi-class hate speech classification using ten binary datasets and various linguistic features. Their best performing model, CAT Boost, achieved 0.89 accuracy and 0.87 F1-score. Authors in [7] developed a hybrid approach that combines Ant Lion Optimization (ALO) and Moth Flame Optimization (MFO) with traditional machine learning techniques, achieving improved accuracy results of 0.92 and 0.90, respectively. Authors in [8] created a sentiment-lexicon-based approach that reached an F1-score of 0.65 in analyzing online discussions. Authors in [9] developed a hierarchical deep learning model that integrates Convolutional Neural Networks (CNN), BiLSTM and Bidirectional Encoder Representations from Transformers (BERT) achieving higher F1-scores for hate, abuse and neutral categories.

Another study [10] used Graph Auto-Encoders (GAE) for unsupervised detection, outperforming TF-IDF in multilingual environments. In [11], the authors introduced HateNet with SubDQE augmentation and Weighted Drop-Edge within a Graph Convolutional Network (GCN) framework, achieving 0.84 accuracy. Transformer-based models have become popular because of their strong contextual capabilities. Authors in [12] used ensemble fusion of classifiers including Embeddings from Language Models (ELMo), BERT, and CNN, resulting in a 13% increase in F1-score. Authors in [13] assessed various Twitter-specific BERT variants using a multi-ideology dataset and attained the highest F1-score of 0.72. In [14], the authors proposed a Dual Contrastive Learning (DCL) model using span-level semantics and focal loss, achieving an F1-score of 0.84. For low-resource languages, HateMAML combines episodic meta-learning with multilingual encoders (mBERT/XLM-R), demonstrating a 3% improvement over baselines across eight languages [15].

Several studies improved classification accuracy by incorporating auxiliary features and explainability methods. In [16], emotion and sentiment features were fused using multi-task learning, revealing the strong connection between hateful content and negative emotions like anger. In [17], syntactic dependency graphs were used with RoBERTa and BiLSTM, leading to an increase in Macro-F1 by up to 3.88%. BiCapsHate employed capsule networks and a Hatebase lexicon, surpassing models like HateBERT and ToxicBERT with an average F1-score of 0.88 [18]. Comparative studies and feature-level improvements have also been examined. In [19], the authors compared ML and DL models on an expert-labeled dataset, with BiLSTM and GloVe embeddings achieving strong performance results. Authors in [20] integrated user features

(demographics, network, emotions) with tweet content, boosting the F1-score by up to 0.32.

While many studies have contributed valuable insights to hate speech detection through machine learning, deep learning, and graph-based methods, most depend on limited or platform-specific datasets and have difficulty generalizing across different digital environments [21]. Few studies have examined cross-platform challenges with large-scale, diverse data or investigated model architectures that comprehensively capture contextual and cultural differences. These gaps emphasize the need for larger, more comprehensive datasets and stronger models capable of managing the complexity of real-world hate speech across various platforms and communities.

II. MATERIALS AND METHODS

Hate speech detection is a particular task within the broader field of text classification. The main goal of hate speech detection was to create automated systems that can identify texts with hateful or discriminatory language. This task belongs to the area of text classification in NLP, where machine learning statistical models are trained on labeled datasets to categorize text content.

Hate speech detection can be formalized as follows: A dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$ is given containing a piece of text with $y_i \in 0,1$ indicating the label (1 for hate speech, 0 for non-hate speech). The aim is to learn a mapping $f: x_i \rightarrow y_i$ that minimizes the classification error. Several factors contribute to the success of hate speech detection systems, such as: a) Quality and diversity of training data, b) selection of appropriate model architectures, c) optimization techniques.

In addition, robust data preprocessing and feature engineering techniques are also crucial for identifying relevant patterns and nuances in language and developing effective models often involves iterative experimentation with preprocessing methods, feature representations and model architectures, as well as exploring alternative strategies such as fine-tuning of deep neural networks, employing ensemble techniques or integrating contextual information through advanced language models.

A. Dataset Exploration

The dataset used for experiments was organized into three sets including raw data, pre-processed data and data augmentation files [22]. The raw data consisted of 842,335 samples including 708,641 non-hate sentences and 133,694 hate sentences. These raw data were pre-processed by removal of non-English samples and stop words, lemmatization etc. The original dataset was imbalanced with 82.1% non-hate 17.9% hate samples . This outstanding imbalance may lead to a number of issues with text classification tasks. The model might favor the most common sample classes leading to high accuracy but low precision for the minority class. It can cause the hate speech class to be misclassified which is essential for the task.

To address this issue, resampling techniques were applied using undersampling and augmentation methods. For undersampling, the majority class (non-hate) was undersampled to reduce its size in the dataset. This method randomly selects samples from the majority class (non-hate) to make it equal to the sample size of the minority class (hate). In addition, an augmentation technique such as contextual word embedding was employed to increase the sample size of the minority class as follows:

- Substitution and Insertion Methods: Words in the hate sentences were substituted or new words were inserted based on contextual embeddings provided by BERT to create additional variations of the hate samples.
- Synonym Augmentation: Synonym replacement was conducted using WordNet embeddings where words in the hate samples were replaced with their synonyms to create new sentences that keep the original meaning but increase the dataset size.

The re-sampling techniques used in this study work as follows: Undersampling maintained the statistical representation by randomly selecting the sample from the majority class. BERT-based augmentation technique provides additional context to generate meaningful variations that still preserve the nature of hate speech. After that, synonym augmentation with WordNet increases the linguistic variety without affecting meaning, so key patterns important for hate speech detection are preserved. The resulting dataset for experiments had 361,594 non-hate samples and 364,525 hate speech samples. In the next stage, dataset analysis of word frequencies was carried out to determine the distribution of words within the data samples. The aim of word frequency analysis is to find variations in the length of sentences. The results of this analysis show that the number of words in each sample varied considerably. Figure 2 shows the word distribution for all samples where the words in the samples range from 10 to 200+. Literature on NLP shows that BLASTM is more efficient at capturing longer text sequences because of its ability to capture long strings.

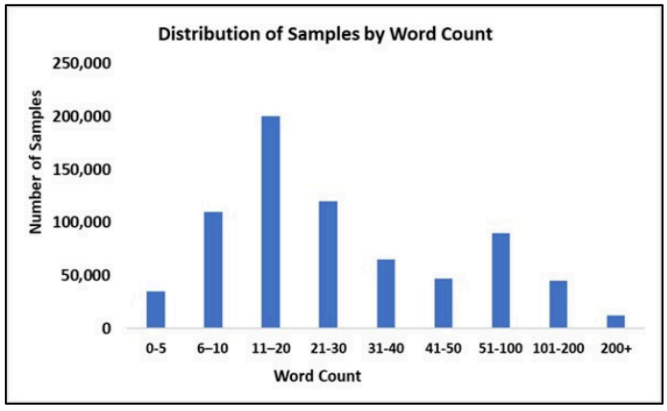

Fig. 2. Word distribution in the dataset.

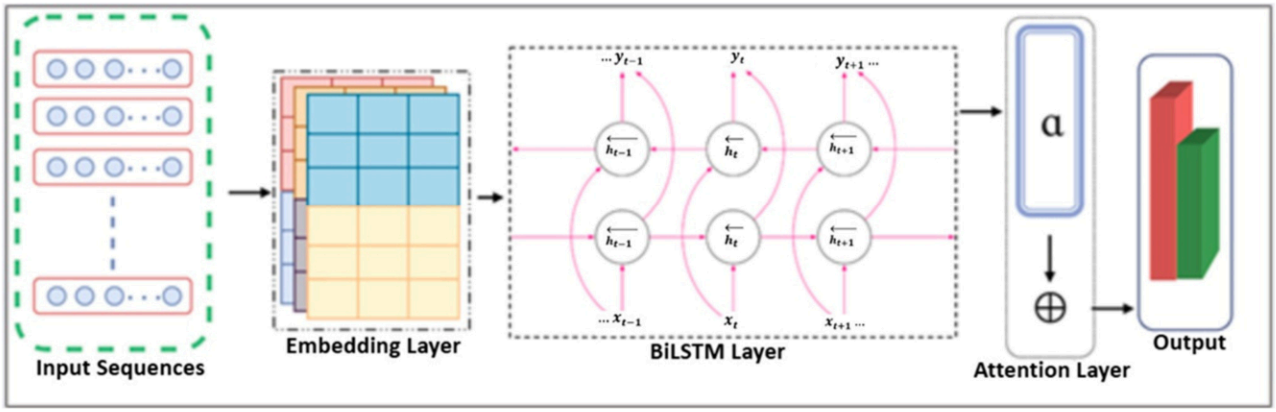

Fig. 3. Architecture of the proposed BiLSTM mode.

B. BiLSTM Model

BiLSTM extends the standard LSTM designed to capture long-term dependencies in sequential data. It utilizes a second LSTM layer that processes the input sequence in reverse order. This allows the network to take both past and future context into account when analyzing data. The reason for using BiLSTM is its proven effectiveness in capturing context from long texts. The dataset was processed through several layers designed to extract meaningful features and ultimately produce a prediction. Figure 3 exhibits the architecture of the proposed BiLSTM model.

1) Input Layer

The proposed model has an input layer that accepted the sequences of textual data. The dimension of input data is equal to the maximum length of the text sequences.

2) Embedding Layer

The first processing layer transformed the input text into dense vector representations. It has been configured with an embedding dimension of 50 so each word of input text data was represented as a 50-dimensional vector. The embedding layer learns semantic relationships between words during training which are critical for text understanding.

3) Bidirectional LSTM Layer

The output from the embedding layer was processed through a bidirectional LSTM layer. This layer captured the contextual information from both preceding and following words in the sequence. The hidden units in the BiLSTM are configured to 32, meaning each LSTM contains 32 units to process the sequence in forward and backward directions. The BiLSTM generated a unified vector by combining both forward and backward contextual information to provide complete understanding of the sequence. A conventional LSTM layer processed the input sequence sequentially at each time step maintaining a hidden state that captures information from earlier time steps. The advantage of LSTMs is their ability to maintain information for long sequences which addresses the vanishing gradient problem commonly faced by traditional Recurrent Neural Networks (RNNs).

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\widetilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh(C_t)$$

where (x_t) is the input at time step (t), (h_{t-1}) is the hidden state from the previous time step, (C_t) is the cell state at the time step (t). The (f_t), (i_t) and (o_t) are the forget, input and output gates, respectively.

4) Dropout Layers

Multiple dropout layers within the proposed architecture have been used to avoid the overfitting. These dropout layers randomly deactivated the neuron units during the learning process. The proposed architecture was incorporated with 50% deactivation rate.

5) Dense and Output Layers

The network architecture comprised a series of fully connected layers that progressively enhanced the extracted features: a dense layer with 32 units and ReLU activation, a dense layer with 16 units and ReLU activation, and finally, a dense layer with a single unit and sigmoid activation function for generating probability scores on binary classification tasks. The network architecture comprised a series of fully connected layers that progressively improve the extracted features: an initial dense layer with 32 units and ReLU activation function and a second dense layer with 16 units and ReLU activation for continued feature improvement. We empirically tested several architectures and chose this configuration (32→16 neurons) with a funnel design that gradually compresses learned representations, while preserving classification performance, using ReLU activation for computational efficiency and to prevent the vanishing gradient problem, with a single unit in the final output layer with the sigmoid activation function that produces probability scores for binary classification.

C. Attention Mechanisms

The dataset used in this study contains sentences of 300+ words. The dataset was collected from a variety of digital

platforms, which poses a challenge in detecting hate speech. The BiLSTM models are good at capturing long-range dependencies. In very long sequences, BiLSTM encounters information bottleneck issues where distant contextual information may be lost or compressed. A self-attention method is used to tackle this issue. The self-attention approach enables the model to focus on the most relevant parts of the sequences, regardless of the length of the sequence or position in the sequence. This is important to identify patterns of hate speech that may be spread by long-form social media content.

The attention mechanism solved this issue by enabling models to pay more attention to specific parts of the sentence. It assigned different weights to each input elements based on the relevance of that element according to the context. The attention mechanism calculates the weight values for each input element in a sequence using scoring functions applied to hidden states h and input representations, followed by softmax normalization to ensure proper weight distribution, which leads to accurate attention assignment across sequence elements.

After the BiLSTM layer processes the input sequence, it produces concatenated hidden states. $H = [h^1, h^2, \dots, h_n]$ where each $h_i = [h_f^i;\ h_b^i]$ combined the forward and backward LSTM hidden states at position i. These bidirectional hidden states captured comprehensive contextual information from both directions in the sequence. Given the BiLSTM hidden states, this research's attention mechanism computes attention scores through:

$$e_i = W_a^T \tanh(W_h h_i + b_a)$$

where $W_a \in \mathbb{R}^{dx1}$ and $W_h \in \mathbb{R}^{dxh}$ are learned parameters, b_a is the bias term, and d is the attention dimension.

The attention weights were normalized using SoftMax and the final context vector was computed by:

$$a_i = \frac{\exp(e_i)}{\sum_{j=1}^{n} \exp(e_j)}$$

$$c = \Sigma_i^{=1n}\, \alpha_i h_i$$

This is crucial for detecting hate speech because discriminatory words (like slurs and threats) can appear anywhere in longer posts. Regardless of where these crucial terms appeared in sequences of more than 300 tokens, our model was able to concentrate on them thanks to the attention weights.

III. EXPERIMENTS AND RESULTS

To evaluate the performance of the proposed model, the dataset was randomly split into 80% training and 20% testing sets to ensure fair evaluation. Standard NLP preprocessing steps were performed before training, including converting to lowercase, removing stopwords, cleaning punctuation, and tokenizing. Data balancing methods were also applied to address class imbalance between hate and non-hate samples. The model was trained using four different embedding layers to identify the optimal one for such a diverse dataset. The performance of model was evaluated using standard metrics: accuracy, precision, recall and F1 score. Table I summarizes the key hyperparameters used in the experiments.

TABLE I. HYPERPARAMETER SETTINGS

Parameter	Value
Embedding Size	50
Dropout	0.5
Epochs (Max)	30
Batch Size	256
Optimizer	Adam
Learning rate	0.0001

Experiments were conducted with batch sizes from 32 to 512, learning rates from 0.0001 to 0.01 and dropout rates from 0.1 to 0.5. The selected values (learning rate: 0.0001, batch size: 256, dropout: 0.5, epochs: 50) achieved optimal convergence and generalization while preventing overfitting on our hate speech dataset. Table II displays the performance across both hate and non-hate classes. The FastText, Word2Vec and Trainable embeddings had the highest accuracy of 91%. All embedding techniques achieved 0.90 or more on the F1-score. This shows a balanced precision and recall performance for both classes.

TABLE II. BILSTM RESULTS WITH VARIOUS EMBEDDING TECHNIQUES

Embedding	Class	Precision	Recall	F1-Score	Acc
GloVE	Hate	0.87	0.94	0.90	0.90
	Non-hate	0.93	0.86	0.90	
FastText	Hate	0.88	0.94	0.91	0.91
	Non-hate	0.94	0.87	0.90	
Word2Vec	Hate	0.88	0.94	0.91	0.91
	Non-hate	0.94	0.87	0.90	
Trainable	Hate	0.88	0.94	0.91	0.91
	Non-hate	0.93	0.87	0.90	

The recall value for the hate class was improved for all embedding methods. With the highest recall of 0.94, the FastText embedding technique is optimum in identifying all hate samples. FastText, Word2Vec, and Trainable embeddings achieved nearly identical performance, with a slight advantage in precision and recall over GloVe. Although GloVe performed slightly lower recall for the non-hate class (0.86), its F1-score remained competitive because of a higher precision (0.93). These results verified that the proposed model is reliable across various embedding types. FastText, Word2Vec, and trainable embeddings provided the most consistent and slightly improved performance, making them better suited for generalized detection tasks across diverse data.

TABLE III. RESULTS OF THE BILSTM MODEL WITH AN ATTENTION MECHANISM

Embedding	*Class*	*Precision*	*Recall*	*F1-Score*	*Acc*
GloVE	Hate	0.88	0.95	0.91	0.92
	Non-hate	0.94	0.88	0.91	
FastText	Hate	0.89	0.96	0.93	0.94
	Non-hate	0.95	0.89	0.92	
Word2Vec	Hate	0.89	0.95	0.92	0.93
	Non-hate	0.95	0.88	0.92	
Trainable	Hate	0.89	0.95	0.92	0.93
	Non-hate	0.94	0.88	0.91	

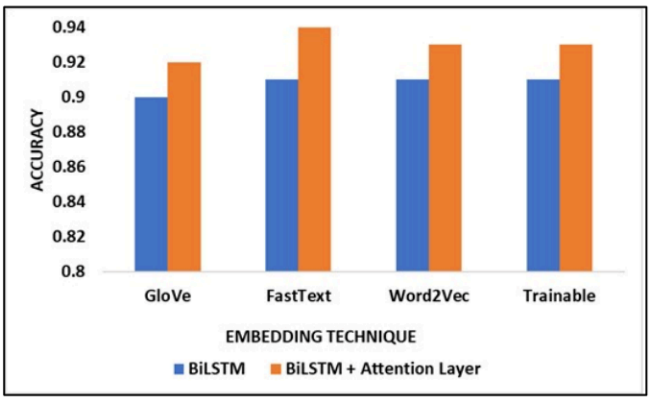

Fig. 4. Performance comparison with and without the attention layer.

In the next phase of the analysis, an attention mechanism was incorporated into the BiLSTM model. The goal of this attention mechanism was to improve BiLSTM's ability to focus on the most essential parts of very long sequences. The findings of this analysis are shown in Table III and Figure 4.

FastText embeddings achieved the highest precision of 0.89 for the hate class and a recall of 0.96. Similarly, Word2Vec embeddings achieved a precision of 0.89 for the hate class, along with a recall of 0.95, an F1-score of 0.92, and an accuracy of 0.93. Trainable embeddings also performed well, with a precision of 0.89, a recall of 0.95, an F1-score of 0.92, and an accuracy of 0.93. Figure 5 shows that combining attention mechanisms with BiLSTM models improves accuracy by 2% to 3% when processing longer text sequences, highlighting the value of attention-based enhancements in capturing important context from lengthy inputs.

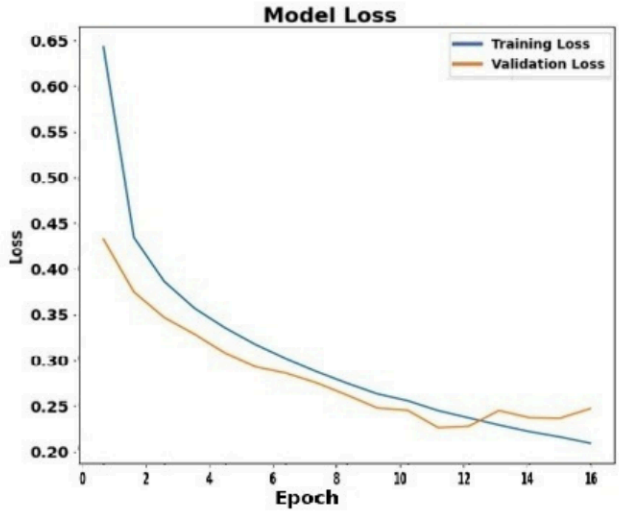

Fig. 5. BiLSTM training and validation loss on FastText achieving maximum performance.

Table IV compares the results of the current study with previous research that used one or more subsets of the same dataset. In a few studies researchers reported results for each subset individually rather than as a combined dataset. We report average results for these instances to make direct comparisons with our findings. In addition, if a study used multilingual data, only the results related to the English dataset were considered. This outcome aligns with findings from related works in healthcare, IoT, WoT, and industrial domains that address large-scale classification challenges in heterogeneous data environments, where robust model architectures and hybrid learning strategies have been demonstrated to enhance generalization and performance [43-45]. These results highlight the effectiveness of the proposed model as a global framework capable of leveraging large diverse datasets to enhance the accuracy of hate speech detection.

TABLE IV. COMPARISON OF THE PROPOSED STUDY WITH STATE-OF-THE-ART TECHNIQUES

Study	Dataset	Accuracy	F1-Score
[6]	[34-36, 31, 32]	0.89	0.87
[9]	[23, 25-28]	0.83	0.75
[11]	[28-30]	0.84	0.84
[31]	[23]	0.79	0.82
[32]	[33]	0.88	0.88
[34]	[33]		0.80
[14]	[25, 28]	0.84	0.84
[15]	[25, 30, 33, 35, 36]		0.73
[17]	[25, 37]	0.77	0.74
[18]	[24, 30, 38, 39, 40]		0.88
[20]	[29, 30, 41, 42]		0.84
Proposed	Curated dataset	0.94	0.92

IV. CONCLUSION

Digital platforms face a significant challenge in identifying harmful content requiring advanced computational approaches to analyze text-based communications effectively. These are complex patterns of speech that deep learning techniques are now efficient methods to process. The limitation of the size of the dataset and context are issues that deep learning techniques can address. This study tackled the problem of detecting hate speech across multiple digital platforms and proposed a deep learning model based on BiLSTM with a self-attention mechanism trained on a large and diverse dataset of over 0.842 million samples from multiple public sources representing multiple regions and platforms. The study used various data embedding techniques, preprocessing steps, and resampling strategies to improve performance, and the model achieved strong results with a maximum accuracy of 93% and an F1-score of 0.92, which outperformed several baseline and state-of-the-art models.

REFERENCES

[1] U. Nations. "What is hate speech?" https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech.

[2] "Global social media statistics research summary" https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/.

[3] E. A. Vogels. "The State of Online Harassment," *Pew Research Center*, Jan. 13, 2021. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment.

[4] J. H. Tien, M. C. Eisenberg, S. T. Cherng, and M. A. Porter, "Online reactions to the 2017 'Unite the right' rally in Charlottesville: measuring polarization in Twitter networks using media followership," *Applied*

Network Science, vol. 5, no. 1, pp. 1–27, Dec. 2020, https://doi.org/10.1007/s41109-019-0223-3.

[5] W. Sharif, S. Abdullah, S. Iftikhar, D. Al-Madani, and S. Mumtaz, "Enhancing Hate Speech Detection in the Digital Age: A Novel Model Fusion Approach Leveraging a Comprehensive Dataset," *IEEE Access*, vol. 12, pp. 27225–27236, 2024, https://doi.org/10.1109/ACCESS.2024.3367281.

[6] K. A. Qureshi and M. Sabih, "Un-Compromised Credibility: Social Media Based Multi-Class Hate Speech Classification for Text," *IEEE Access*, vol. 9, pp. 109465–109477, 2021, https://doi.org/10.1109/ACCESS.2021.3101977.

[7] C. Baydogan and B. Alatas, "Metaheuristic Ant Lion and Moth Flame Optimization-Based Novel Approach for Automatic Detection of Hate Speech in Online Social Networks," *IEEE Access*, vol. 9, pp. 110047–110062, 2021, https://doi.org/10.1109/ACCESS.2021.3102277.

[8] N. D. Gitari, Z. Zuping, D. Hanyurwimfura, and J. Long, "A lexicon-based approach for hate speech detection," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 10, no. 4, pp. 215–230, 2015, https://doi.org/10.14257/ijmue.2015.10.4.21.

[9] N. Vashistha and A. Zubiaga, "Online Multilingual Hate Speech Detection: Experimenting with Hindi and English Social Media," *Information*, vol. 12, no. 1, Dec. 2020, Art. no. 5, https://doi.org/10.3390/info12010005.

[10] G. L. De la Peña Sarracén and P. Rosso, "Unsupervised embeddings with graph auto-encoders for multi-domain and multilingual hate speech detection," in *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France, 2022, pp. 2196–2204.

[11] C. Duong, L. Zhang, and C.-T. Lu, "HateNet: A Graph Convolutional Network Approach to Hate Speech Detection," in *2022 IEEE International Conference on Big Data (Big Data)*, Dec. 2022, pp. 5698–5707, https://doi.org/10.1109/BigData55660.2022.10020510.

[12] Y. Zhou, Y. Yang, H. Liu, X. Liu, and N. Savage, "Deep Learning Based Fusion Approach for Hate Speech Detection," *IEEE Access*, vol. 8, pp. 128923–128929, 2020, https://doi.org/10.1109/ACCESS.2020.3009244.

[13] M. Gaikwad, S. Ahirrao, K. Kotecha, and A. Abraham, "Multi-Ideology Multi-Class Extremism Classification Using Deep Learning Techniques," *IEEE Access*, vol. 10, pp. 104829–104843, 2022, https://doi.org/10.1109/ACCESS.2022.3205744.

[14] J. Lu *et al.*, "Hate Speech Detection via Dual Contrastive Learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2787–2795, 2023, https://doi.org/10.1109/TASLP.2023.3294715.

[15] M. R. Awal, R. K.-W. Lee, E. Tanwar, T. Garg, and T. Chakraborty, "Model-Agnostic Meta-Learning for Multilingual Hate Speech Detection," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 1, pp. 1086–1095, Feb. 2024, https://doi.org/10.1109/TCSS.2023.3252401.

[16] A. R. Jafari, G. Li, P. Rajapaksha, R. Farahbakhsh, and N. Crespi, "Fine-Grained Emotions Influence on Implicit Hate Speech Detection," *IEEE Access*, vol. 11, pp. 105330–105343, 2023, https://doi.org/10.1109/ACCESS.2023.3318863.

[17] X. Fan, J. Liu, J. Liu, P. Tuerxun, W. Deng, and W. Li, "Identifying Hate Speech Through Syntax Dependency Graph Convolution and Sentiment Knowledge Transfer," *IEEE Access*, vol. 12, pp. 2730–2741, 2024, https://doi.org/10.1109/ACCESS.2023.3347591.

[18] A. Kamal, T. Anwar, V. K. Sejwal, and M. Fazil, "BiCapsHate: Attention to the Linguistic Context of Hate via Bidirectional Capsules and Hatebase," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 2, pp. 1781–1792, Apr. 2024, https://doi.org/10.1109/TCSS.2023.3236527.

[19] A. Toktarova *et al.*, "Hate Speech Detection in Social Networks using Machine Learning and Deep Learning Methods," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 14, no. 5, pp. 396-406, May 2023, https://doi.org/10.14569/IJACSA.2023.0140542.

[20] R. Raut and F. Spezzano, "Enhancing hate speech detection with user characteristics," *International Journal of Data Science and Analytics*, vol. 18, no. 4, pp. 445–455, Oct. 2024, https://doi.org/10.1007/s41060-023-00437-1.

[21] G. Ansari, P. Kaur, and C. Saxena, "Data Augmentation for Improving Explainability of Hate Speech Detection," *Arabian Journal for Science and Engineering*, vol. 49, no. 3, pp. 3609–3621, Mar. 2024, https://doi.org/10.1007/s13369-023-08100-4.

[22] D. Mody, Y. Huang, and T. E. Alves de Oliveira, "A curated dataset for hate speech detection on social media text," *Data in Brief*, vol. 46, Feb. 2023, Art. no. 108832, https://doi.org/10.1016/j.dib.2022.108832.

[23] T. Mandl *et al.*, "Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages," in *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE '19)*, Kolkata, India, 2019, pp. 14–17, https://doi.org/10.1145/3368567.3368584.

[24] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*, Montreal, QC, Canada, 2017, pp. 512–515, https://doi.org/10.1609/icwsm.v11i1.14955.

[25] V. Basile *et al.*, "SemEval-2019 Task 5: Multilingual detection of hate speech against immigrants and women in Twitter," in *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*, Minneapolis, MN, USA, 2019, pp. 54–63, https://doi.org/10.18653/v1/S19-2007.

[26] M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, and E. Belding, "Peer to peer hate: Hate speech instigators and their targets," in *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)*, Stanford, CA, USA, 2018, pp. 52-61.

[27] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung, "Multilingual and multi-aspect hate speech analysis," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 4675–4684, https://doi.org/10.18653/v1/D19-1474.

[28] P. Mathur, R. Sawhney, M. Ayyar, and R. Shah, "Did you offend me? Classification of offensive tweets in Hinglish language," in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Brussels, Belgium, 2018, pp. 138–148.

[29] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in *Proceedings of the NAACL Student Research Workshop*, San Diego, CA, USA, 2016, pp. 88–93.

[30] A. Founta *et al.*, "Large scale crowdsourcing and characterization of Twitter abusive behavior," in *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)*, Stanford, CA, USA, 2018, pp. 491–500, https://doi.org/10.1609/icwsm.v12i1.14991.

[31] N. Bölücü and P. Canbay, "Hate speech and offensive content identification with graph convolutional networks," in *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE '21)*, India, 2021, pp. 44-51.

[32] S. Dowlagar and R. Mamidi, "HASOCOne@FIRE-HASOC2020: Using BERT and Multilingual BERT models for Hate Speech Detection." arXiv, Jan. 22, 2021, https://doi.org/10.48550/arXiv.2101.09007.

[33] T. Mandl, S. Modha, A. Kumar M, and B. R. Chakravarthi, "Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German," in *Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation*, New York, NY, USA, Jan. 2021, pp. 29–32, https://doi.org/10.1145/3441501.3441517.

[34] W. Yu, B. Boenninghoff, and D. Kolossa, "Hybrid representation fusion for Twitter hate speech identification," in *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE '21)*, India, 2021, pp. 319-329.

[35] M. Zampieri *et al.*, "SemEval-2020 Task 12: Multilingual offensive language identification in social media (OffensEval 2020)," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*

(SemEval-2020), Barcelona, Spain (online), 2020, pp. 1425–1447, https://doi.org/10.18653/v1/2020.semeval-1.188.

[36] C. Bosco, F. Dell'Orletta, F. Poletto, M. Sanguinetti, and M. Tesconi, "Overview of the EVALITA 2018 hate speech detection task," in *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*, Turin, Italy, 2018, https://doi.org/10.4000/books.aaccademia.4503.

[37] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the Type and Target of Offensive Posts in Social Media," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, Mar. 2019, pp. 1415–1420, https://doi.org/10.18653/v1/N19-1144.

[38] R. Agarwal. "Twitter hate speech." https://www.kaggle.com/datasets/vkrahul/twitter-hate-speech.

[39] L. Gao and R. Huang, "Detecting Online Hate Speech Using Context Aware Models," in *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, Varna, Bulgaria, June 2017, pp. 260–266, https://doi.org/10.26615/978-954-452-049-6_036.

[40] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, pp. 14867–14875, May 2021, https://doi.org/10.1609/aaai.v35i17.17745.

[41] M. Xia, A. Field, and Y. Tsvetkov, "Demoting Racial Bias in Hate Speech Detection," in *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, Online, Apr. 2020, pp. 7–14, https://doi.org/10.18653/v1/2020.socialnlp-1.2.

[42] B. He, C. Ziems, S. Soni, N. Ramakrishnan, D. Yang, and S. Kumar, "Racism is a virus: anti-asian hate and counterspeech in social media during the COVID-19 crisis," in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, New York, NY, USA, Jan. 2022, pp. 90–94, https://doi.org/10.1145/3487351.3488324.

[43] J. Malik, A. Akhunzada, A. S. Al-Shamayleh, S. Zeadally, and A. Almogren, "Hybrid deep learning based threat intelligence framework for Industrial IoT systems," *Journal of Industrial Information Integration*, vol. 45, May 2025, Art. no. 100846, https://doi.org/10.1016/j.jii.2025.100846.

[44] T. M. Ghazal *et al.*, "Federated Learning With Small and Large Models With Privacy-Preserving Data Space for Holographic Internet of Things in Consumer Electronics," *IEEE Transactions on Consumer Electronics*, vol. 71, no. 2, pp. 5259–5274, Feb. 2025, https://doi.org/10.1109/TCE.2025.3573033.F

[45] M. Maaz, G. Ahmed, A. Sami Al-Shamayleh, A. Akhunzada, S. Siddiqui, and A. Hussein Al-Ghushami, "Empowering IoT Resilience: Hybrid Deep Learning Techniques for Enhanced Security," *IEEE Access*, vol. 12, pp. 180597–180618, 2024, https://doi.org/10.1109/ACCESS.2024.3482005.