International Journal of Data Informatics and Intelligent Computing (IJDIIC)
Vol. 3, No. 4, December 2024, pp. 26~39
DOI: 10.59461/ijdiic.v3i4.143

Deep Learning-Based Identification of Immoral Posts on Social Media Using Fine-tuned Bert Model

Bibi Saqia[1], Khairullah Khan[1], Atta Ur Rahman[2], Wahab Khan[1]
[1] Department of Computer Science, University of Science and Technology, Bannu, 28100, Pakistan
[2] IRC for Finance and Digital Economy, KFUPM Business School, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia

Article Info

Article history:

Received September 12, 2024
Revised November 02, 2024
Accepted November 11, 2024

Keywords:

Immoral Post
Social Media
Deep Learning
Bert model
Word Embeddings

ABSTRACT

The propagation of immoral content on social media poses substantial worries to online societal well-being and communication standards. While beneficial, traditional machine learning (ML) methods fall short of capturing the difficulty of textual and sequential data. This work reports this gap by suggesting a deep learning-based technique for detecting immoral posts on social media. The proposed model presents a fine-tuned Bidirectional Encoder representation from Transformers (BERT) with word embedding methods. Word2Vec and Global Vectors for Word Representation (GloVe) are employed to improve the identification of immoral posts on social media platforms to advance detection accuracy and strength. The incentive behind this study stems from the increasing demand for more sophisticated methods to struggle with damaging content. The proposed model is considered to capture the complicated patterns and semantic nuances in immoral posts by decreasing the dependence on manual feature engineering. The model is trained and assessed using benchmark datasets containing SARC and HatEval, which deliver a detailed set of labelled user-generated posts. The proposed model shows the best performance compared to traditional ML approaches. The fine-tuned Bert-based Word2Vec embeddings achieved a precision of 95.68%, recall of 96.85 %, and F1 scores of 96.26% on the SARC dataset. Fine-tuned Bert-based GloVe on the HatEval dataset achieved superior precision of 96.65, recall of 97.75, and F1-score of 97.20. The proposed results highlight the potential of the deep learning (DL) approach and fine-tuned BERT models, considerably refining the detection of unethical content on social networks.

Corresponding Author: Bibi Saqia (e-mail: saqiaktk@ustb.edu.pk)

1. INTRODUCTION

Social media has become an essential element of modern life, facilitating platforms for messages, views, and data sharing [1]. Social networks such as Facebook, Twitter, and Instagram permit people to communicate on a global scale. While these platforms offer different welfare, they also offer increases to different challenges, mainly the spreading of immoral, dangerous, and misleading content [2]. Immoral posts on social networks containing hate speech, offensive language, misinformation, and insulting remarks hurt society [3]. This risky digital environment is a warning to public safety, mental health, and societal norms[4]. Identifying and eliminating unethical content on social media is a serious matter of maintaining moral standards and developing a positive online atmosphere [5].

Despite the necessity for consistent systems to identify immoral content, current approaches to content moderation often rely on traditional ML techniques [6]. Whereas operational, these methods face considerable limitations when employed to the complex and variety of content. Traditional approaches such as Naive Bayes, Decision Trees and Support Vector Machines (SVMs) [7] are utilized in offensive language detection. These studies work on the nuances of natural language in social media content, which usually include slang, informal language, ambiguity, and sarcasm. These techniques are inadequate in their capability to process the sequential and contextual nature of textual data in large-scale, dynamic social networks [8].

A huge amount of user-generated content has been made as a consequence of social media's volatility [9]. Thousands of comments are made on social networks each second, making it nearly hard for human moderators to manually sort and evaluate content for moral crimes. Automated schemes have become vital to control the scale and variety of online remarks [10]. The consistency and accuracy of these schemes are essential, particularly in differentiating between unethical content and genuine dialogue. There is an increasing demand for further progressive models capable of appreciating and processing linguistics in a method that imitates its inherent complication. Earlier studies have endeavoured to address these matters using the application of traditional ML methods. These methods concentrate on binary classification works such as sentiment analysis, text classification, spam identification, and toxic remarks [11]. Though it's efficient in skilful atmospheres, these methods usually lack the ability to capture the depth of sense essential to correctly perceive immoral content. Some studies using the Term Frequency-Inverse Document Frequency (TF-IDF) [12] and the Bag of Words (BoW) model [13] provide good results in text classification but suffer from the incapability to capture the semantic meaning of words. These approaches decrease text to insufficient mathematical representations without seeing proper sequential and contextual dependencies between terms or words. Consequently, traditional ML models often yield high false positives or negatives, specifically in control cases relating to implicit hate speech and sarcasm, whereas context performs an important role.

The quantity of social networks has brought forth unparalleled occasions to share views, concepts, and information [14]. But, it has also generated a new avenue for immoral attitudes, such as offensive language, hate speech, cyberbullying, misinformation, and the spread of activism. In reaction, academic and industrial societies have discovered different approaches to automatically identify and moderate the damage produced by such content. The development of these techniques, from traditional ML to advanced DL methods, shows important progressions in both the accuracy and complexity of immoral post-identification [15]. However, this study is inspired by the latent DL to fill the gap left by traditional ML approaches. BiLSTM networks are employed to identify the challenges of immoral content identification on social media. BiLSTM networks are a category of Recurrent Neural Networks (RNN) that perform best in processing sequential data by finding long-term associations between words and sentences [16]. Contrasting traditional techniques that process data in a unidirectional way, BiLSTM reflects both the past (previous words) and the upcoming (next words) in its evaluation, making it compatible with complex textual data where context is mandatory. The BERT-based models have noticed an important progression by presenting contextualized embeddings, yet BERT alone can sometimes slip domain-based language nuances and could need maximum computational resources.

The novelty of this study lies in addressing these limitations through improving fine-tuned BERT with word embeddings like Word2Vec and GloVe. The proposed work makes a forceful, dual-layered semantic understanding structure. This blend influences the contextual power of BERT with the semantic specificity presented via traditional word embeddings. This allows a more inclusive understanding of social media linguistics containing abbreviations, slang, and unspoken content. The suggested model expressively progresses on traditional approaches by catching both the general context and domain-base refinements. The benchmark datasets SARC and HatEval are used to certify the generalizability and reliability of the model across varied contexts. These datasets are extensively known in Natural Language Processing (NLP) for work such as abusive language and hate speech detection. The proposed model is tested on different, real-world data, delivering a complete assessment of its performance.

1.1. Contributions

This work provides a deep learning-based technique for revealing immoral posts on social media to improve the growth of English NLP research. This paper considers the challenges of immoral post-detection, which has gotten slight attention from NLP researchers. This competency is essential for identifying immoral content, as it allows the model to capture the nuanced patterns and associations within the text that may sign hurtful intent. The key contributions of this study are given below:

1. The proposed study employed a fine-tuned BERT model to capture intricate dialectal nuances, with indirect linguistic and sarcasm, which are important in detecting unethical posts. The assimilation of pre-trained word embeddings, Word2Vec or GloVe, with fine-tuned BERT embeddings extract both domain-base meanings and rich contextual data.
2. The integration of progressive word embedding techniques, like GloVe and Word2Vec, into the model. These embeddings capture the semantic associations between words, permitting the model to understand the verbatim meaning of words and their contextual utilization. By incorporating word embeddings, the suggested model improves its capability to perceive immoral content, even when the linguistic employed contains slang and is informal. This is significant for social networks, where non-standard language is common.
3. The proposed study applied two benchmark datasets, SARC and HatEval, to evaluate the accuracy and reliability of the model. These datasets deliver different instances of social media remarks labelled

as moral or immoral, permitting the model to generalize well across different scenarios. We also then integrated a dataset to detect offences. We gather data from the two benchmark datasets that report different kinds of dissolution, such as cyberbullying, hate speech, and aggressive online script, to make an integrated dataset for identifying immorality on social media. This combined dataset might be utilized to notice different sorts of immoral content on the web.

4. The proposed work improves the identification of immoral content on social networks by leveraging a dual-embedding strategy, increasing both semantic productivity and model accuracy in a complex, real-world situation.

1.2. Paper Organization

The rest of this paper is organized as follows: Section 2 describes an assessment of the related work along with limitations of earlier work. Section 3 shows the suggested methodology. Section 4 covers the findings of the experiments to assess the proposed model. Section 5 presents the results and comparisons with other state-of-the-art techniques. Section 6 defines the conclusion along with their future work.

2. LITERATURE REVIEW

2.1. Early Approaches in Unethical Post Detection

Previous work in immoral post-detection mainly relied on keyword filtering methods [17] and rule-based schemes [18]. These approaches involved generating lexicons of abusive or unsuitable words and manually making guidelines to flag posts [19]. However, as social network language improved with the abbreviations, the inclusion of slang [20]. The more subtle forms of immoral text the limitations of these methods became misleading. These rule-based methods have some limitations to generalize well across diverse contexts and dialects. This frequently leads to false negatives or false positives [21]. Work performed in [22] highlighted the faults of keyword-based methods in controlling the context-dependent and dynamic nature of social media posts. To address these challenges, academics employ traditional ML algorithms for immoral post-recognition [23]. Methods like Support Vector Machines (SVM), Decision Trees, K Nearest Neighbor (KNN), and Naive Bayes were usually applied for offensive language detection on SM [24]. For instance, The study published by [25] employed SVM classifiers to identify hate speech in social media content, and [26] employed Logistic Regression to detect hate speech on Twitter. These ML methods relied on statistical attributes mined from text, like term frequency-inverse document frequency (TF-IDF), n-grams, and sentiment analysis. While traditional ML algorithms achieved enhancements over rule-based approaches, they were still limited in their capability to comprehend context and semantics. The dependence on attributes, such as word frequency and tokenization, considers these techniques to reflect the integral complication of web content. When dealing with emerging indirect language, slang and metaphors [27] painted this issue. Traditional ML models need wide attribute engineering and cannot extract the consecutive nature of the text. Consequently, these models were less operational in recognizing subtle or contextually dependent immoral posts.

2.2. Deep Learning and its Emergence in Text Classification

There was a paradigm move in how text classification and text mining studies were approached with the origination of DL [28]. DL models containing neural networks have the aptitude to reduce the want for manual feature engineering and automatically learn features from data. This flexibility joins with their capability to process enormous datasets and consider DL models appropriate for the dynamic and complex nature of social media content [29].

One of the initial uses of DL in text classification includes the Convolutional Neural Networks (CNNs). The work conducted in [30] presents that CNNs, initially established for image processing studies, could also be employed to text by handling sentences as orders of word vectors. CNNs were capable of extracting local dependencies between words, making them efficient for studies such as sentiment evaluation and abusive language identification [31]. Still, CNNs were inadequate in their capacity to capture long-term dependencies in text, as they mainly concentrated on local features.

Recurrent Neural Networks (RNNs) were presented to report the confines of CNNs. RNNs are intended to process consecutive data, making them more appropriate for works containing text, where the sequence of words performs an essential role in determining word meaning. The study proposed in [32] employed an RNNs-based model for hate speech identification and found that they outperformed traditional ML approaches. Therefore, RNNs are disposed toward racism or sexism problems such as vanishing gradients, which hamper their aptitude to capture long-term dependencies in lengthy text orders.

2.3. Long Short-Term Memory (LSTM) Networks and Their Application

To consider RNN limitations suggested Long Short-Term Memory (LSTM) networks [33], intended to retain data over prolonged series. LSTMs influence memory cells to update and store data across time stages,

showing operational outputs in sequential tasks such as text classification, translation, and speech recognition. LSTMs to perceive abusive linguistics on Twitter show that LSTMs overtake traditional techniques through context-dependent nature and capturing text's sequential. The study done by [34] employed CNN and BiLSTM, which surpassed cyberbullying identification on social networks.

2.4. Bidirectional LSTM (BiLSTM) Networks

LSTMs enhanced text classification due to their unidirectional handling of incomplete context capture. BiLSTMs were presented to report this through handling text in both directions and apprehending dependences from prior and subsequent words. This is particularly beneficial in tasks such as sarcasm identification and context-sensitive immoral post-organization. Current studies, like [35] on hate speech identification on Twitter, show BiLSTM's improved context awareness. Similarly, [36] use of BiLSTM outperformed traditional models in the identification of abusive language and undesirable content on social networks.

2.5. Word Embeddings and BERT model

Word embeddings such as Word2Vec and GloVe capture semantic associations. It assists models in understanding informal language and slang in web content [37]. Contextual embeddings such as BERT improve linguistic understanding by dynamically regulating word meaning based on context, which benefits from identifying nuanced linguistics, such as social media spam identification and sarcasm [38]. The study conducted by [39] employed BERT with different techniques to recognize aggressive language and improve understanding of the model's performance.

3. METHODOLOGY

In the proposed methodology, we enhance semantic understanding and identify immoral content on social media through an advanced BERT model improved with conventional word embeddings (Word2Vec and GloVe). This hybrid embedding method integrates the contextual complexity of BERT with the semantic features. Word embeddings are employed to improve and capture nuanced linguistics like slang, sarcasm, and indirect expressions often used in social media posts.

3.1. Dataset

In the proposed study, the Self-Annotated Reddit Corpus (SARC) [40] and HatEval datasets [41] are employed for benchmarking sarcasm and hate speech identification models. SARC, based on Reddit remarks with user-tagged sarcasm, permits context-based sarcasm detection by structured remark associations. HatEval is a Twitter-based dataset for hate speech identification with labels differentiating general and targeted hate. Both datasets are split into 80% data for training and 20% data for testing the proposed model. Table 1 reflects the statistics of the datasets used in this study.

3.2. Preprocessing

Data preprocessing is an essential step in making raw text data for ML models, as it normalizes and standardizes the input to certify reliability and relevancy. Each dataset endures an organized preprocessing pipeline to make the text data for model training. These steps comprise tokenization, removal of stop words, lowercasing, punctuation, and special characters. After preprocessing, every tweet or post is tokenized through the BERT tokenizer. Which converts text into tokens and is further vectorized through pre-trained Word2Vec and GloVe embeddings for extremely semantic representations. Let the total number of tokens be represented by T in different social media posts indicated by P. While every token t_i such as i=1, 2……. T handled it with the help of the BERT tokenizer. Moreover, every token's word embedding E_{t_i} is extracted from the pre-trained Word2Vec or GloVe embeddings. Table 2 shows an instance of each preprocessing step employed on both datasets. Figure 1 indicates the basic structure of the proposed model.

3.3. BERT Embedding Layer

The BERT embedding layer is the primary layer in the BERT model construction that modifies input text tokens into dense vector representations, often called embeddings. These embeddings capture both the meaning of individual words and their contextual associations with other words in a sentence. This creates the The BERT embedding layer is mostly powerful for natural language processing tasks such as text classification, sentiment analysis, and question-answering [42]. Let T, indicate the total number of social media posts P, whereas every token t_i (for i=1,2,..., T) is handled by the BERT tokenizer. Furthermore, every token's term embedding E_{t_i} is captured from the pre-trained Word2Vec or GloVe embeddings.

Table 1. Statistics of datasets used in this study

Dataset	Total Size	Training Set Size	Test Set Size	Purpose
SARC	1,300,000 remarks	1,040,000 remarks	260,000 remarks	Sarcasm Detection
HatEval	13,000 tweets	10,400 tweets	2,600 tweets	Hate Speech Detection

Table 2. Preprocessing steps performed in this study

Step	Description	Instance (Original Text)	Sample (After Processing)
Original Post	Initial raw text data	"@user Women deserve respect, not hate. #RespectWomen"	"@user Women deserve respect, not hate. #RespectWomen"
Tokenization	Splits text into tokens	["@user", "Women", "deserve", "respect", "not", "hate", "#RespectWomen"]	["@user", "Women", "deserve", "respect", "not", "hate", "#RespectWomen"]
Lowercasing	Converts all text to lowercase	["@user", "women", "deserve", "respect", "not", "hate", "#respectwomen"]	["@user", "women", "deserve", "respect", "not", "hate", "#respectwomen"]
Stop Words Removal	Removes common words	["@user", "women", "deserve", "respect", "hate", "#respectwomen"]	["women", "deserve", "respect", "hate", "respectwomen"]
Punctuation Removal	Strips punctuation and special characters	["women", "deserve", "respect", "hate", "respectwomen"]	["women", "deserve", "respect", "hate", "respectwomen"]
BERT Tokenizer & Vectorization	Converts tokens to BERT embeddings or GloVe vectors	Input to BERT Tokenizer / Word2Vec or GloVe embeddings applied	Ready for model input

The BERT produces a contextualized embedding B_{t_i} capturing both previous and next words in the text for each input token t_i. The BERT-encoded illustration of the post is indicated by B(p), which is represented in Eq.1.

$$B(P) = BERT(P) \quad (1)$$

Where, $t_1, t_2, \ldots, t_T$ are tokens while $B(P) = [B_{t_1}, B_{t_2}, \ldots B_{t_r}]$ are the contextual embeddings.

3.4. Word Embedding Integration

Word embedding is a popular NLP technique that aims to transfer a word's semantic meaning [43]. Depending on its context, it deals with a useful numerical explanation of the term. An N-dimensional dense vector signifying the words can be used to evaluate how words of the same features are in a given language. Word embedding has been extensively employed in numerous current NLP studies because of its effectiveness, containing document clustering, part of speech tagging, named entity recognition, text classification, sentiment analysis, and many other issues. The above-mentioned subcategories contain explanations of the two most prevalent pre-trained word embedding models: Stanford GloVe and Google Word2Vec.

Word embeddings convert web contents into dense vector representations for DL models. By leveraging BERT embeddings, which capture contextual nuances in linguistics, the model can evaluate the semantic and syntactic features of unethical content efficiently. Incorporating with BERT improves the model's capability to understand difficult language patterns. For instance, implicit aggressive expressions and sarcasm creation are well-matched for identifying immoral posts across different social media settings.

To improve BERT's embeddings with domain-specific semantic knowledge, we assimilate supplementary embeddings $E(p) = [E_{t_1}, E_{t_2}, \ldots, E_{t_r}]$ for each token from Word2Vec or GloVe. Each token embedding is combined, as shown in Eq.2.

$$F_{t_i} = \alpha \,.\, Bt_i \; + \; (1 - \alpha).\, E_{t_i} \quad (2)$$

Where the weighting factor is denoted by α (ranging from 0 to 1) that balances the involvement of BERT and additional embeddings, and for each token ti, the fused representation is denoted by F_{t_i}.

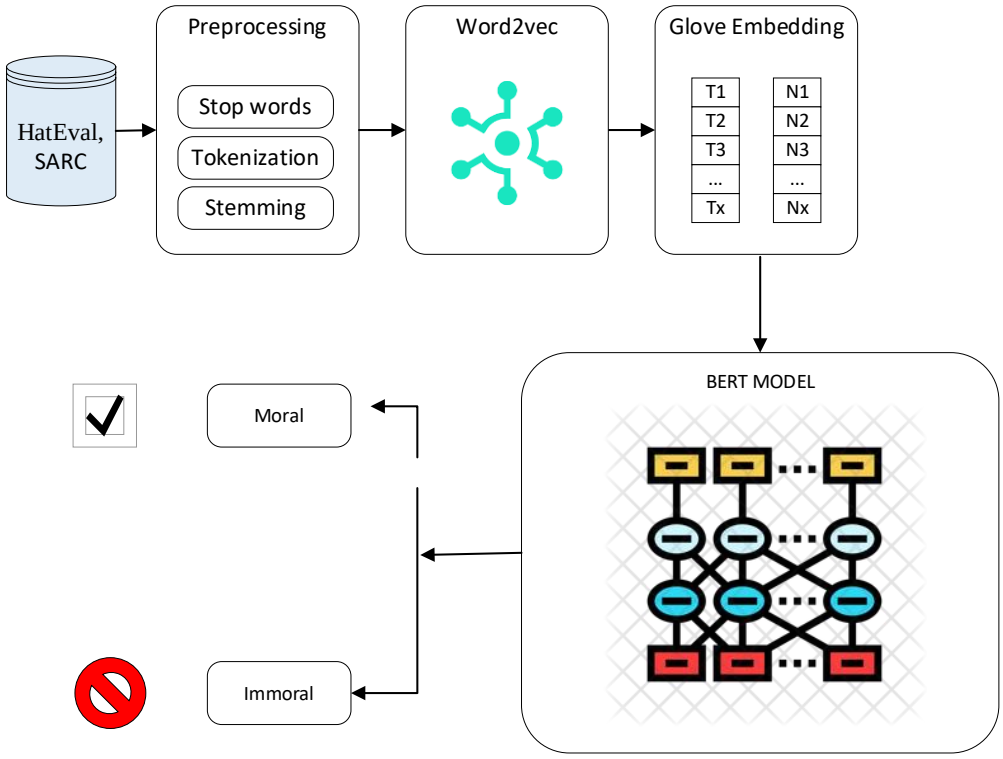

Figure 1. Steps performed to complete the proposed study

3.4.1. Word2vec

Word2Vec is one of the most prominent word embedding techniques offered by the Google research team [44]. Word2Vec uses a massive corpus of data to allocate a vector to each word depending on its nearby context. Two types of training events are employed to extract the word vector. One type of word vector is the continuous bag of words model, which calculates the target word based on its circumstance. The second type is the Skip-Gram model, which employs a word to estimate the target context. The word's feature vector is changed and updated based on every corpus context in which it happens. Google Word2Vec, a vector model that was trained on a huge quantity of more than 100 billion words, was made public by Google.

3.4.2. GloVe

The GloVe is also a dominant word embedding technique [45]. Using an unsupervised learning algorithm trained on a corpus, GloVe generates the distributional feature vectors and learns embeddings. A statistics-based matrix representing the corpus's word-to-word co-occurrence is constructed throughout the learning process. Word2Vec is a prediction-based model, whereas Glove is a count-based model. This is the primary distinction between GloVe and Word2Vec in the learning process. The GloVe features models with various vector dimensions, which are learned from Wikipedia, Twitter, and web content [42].

3.5. BiLSTM Layer

In the proposed study, a Bidirectional Long Short-Term Memory (BiLSTM) layer improves the model's aptitude to capture contextual dependencies in social media content. BiLSTM processes data in both forward and backward directions [33], permitting it to reserve the sequence data critical for understanding nuanced linguistics, like irony or indirect crimes. Assimilating BiLSTM with BERT embeddings further polishes the model's representation of difficult linguistic attributes, assisting it in detecting unethical posts by leveraging both sequential patterns and contextual depth in text [46].

The fusion embeddings $F(P) = [F_{t_1}, F_{t_2}, \ldots, F_{t_r}]$] are served into a BiLSTM network to capture both onward and backward dependences. The BiLSTM production is signified by H(P) as presented in Eq. 3.

$$H(P) = BiLSTM(F(P)) \quad (3)$$

Where the hidden state output from the BiLSTM for token t_i is represented by $H(P) = [H_{t_1}, H_{t_2}, \ldots, H_{t_r}]$ with each H_{t_i}.

3.6. Attention Mechanism

An attention layer is employed in the BiLSTM output to improve the model's emphasis on tokens that donate more to detecting unethical content. The attention mechanism allocates maximum weights to important tokens, following the BiLSTM layer, which assists in capturing subtle nuances and context in text more efficiently. This mechanism points the model's attention to impactful words, inspiring the semantic representation and thereby refining the entire performance in categorizing posts as ethical or unethical. The attention score αt_i for each token t_i is calculated as shown in Eq. 4.

$$\alpha t_i = \frac{exp\ (u^T . tanh(W_a\ .H_{t_i} + b_a))}{\sum_{j=1}^{T} exp\ (u^T . tanh\ (\ W_a\ .H_{t_j} + b_a))} \tag{4}$$

Where the trainable attention weights are denoted by W_a and b_a, the weight vector is indicated by u. The attended representation $H_a(P)$ of the post is then calculated as a weighted sum of BiLSTM outputs as given in Eq. 5.

$$H_a(P) = \sum_{i=1}^{T} \alpha t_i\ .\ H_{t_i} \tag{5}$$

3.7. Classification Layer

A fully connected dense layer that employs a softmax activation function obtains the outcomes from the BiLSTM layer in the classification layer. This step creates probability scores for every class, permitting the model to classify posts as ethical or unethical. By applying the refined representations from preceding layers, the classification layer completes the model's decision-making process with a perfect difference between the targeted classes.

The final post demonstration $H_a(P)$is approved over a fully connected layer monitored through a softmax activation function to categorize each post as either immoral or moral. P denoted the probability of each post being categorized as immoral $\hat{y}$ is computed as indicated in Eq. (6).

$$\hat{y} = softmax(W_c\ .H_a(P) + b_c) \tag{6}$$

Where the trainable parameters of the classification layer are represented by W_c and b_c.

3.8. Loss Function

The loss function performs an important role in enhancing the model during training. In the proposed work, cross-entropy loss is applied, which is mostly well-organized for classification tasks. It computes the difference between the actual and predicted class probabilities, controlling the model in decreasing faults through adjusting weights.

We employed cross-entropy loss to train the proposed model. For an assumed target label y, the loss L is clear in Eq. 7:

$$L = -\sum_{c} y_c\ .log(\hat{y}_c) \tag{7}$$

Where the actual class label is denoted by y_c and the predicted probability of class c is represented by $\hat{y}_c$.

4. EXPERIMENTS

In this section, the entire experiments with the training process, experimental setup, and evaluation criteria are discussed.

4.1. Training Process

We perform data preprocessing to make the textual data for efficient analysis of linguistic patterns, such as hate speech and sarcasm. This involves text cleaning, tokenization, and transformation of every post into vectors using BERT embeddings. These embeddings extract contextual nuances in language. To improve semantic complexity, we combine BERT embeddings with Word2Vec or GloVe by a weighted fusion method. This incorporation balances contextual and lexical data, refining the model's understanding of complex

language attributes. This permits the model to capture both contextual and semantic associations, which is beneficial for social media content, which is often informal and full of slang.

After embedding combination, the mutual vectors are delivered over a BiLSTM layer that processes the sequence bidirectionally. It captures dependencies between words from both the previous and upcoming within the text. This bidirectional processing allows a more detailed understanding of the text's sequential nature. It's important to identify subtle signs of immorality, like indirect language and scorn.

The attention mechanism is employed, conveying higher weights to noteworthy tokens in the posts to further improve interpretability. This relieves the model's attention on keywords or terms that give expressively to immoral or moral classification. The output from the attention layer is then served into a fully associated layer, which yields the final classification of every post as either ethical or unethical.

We applied the Adam optimizer for optimization, primarily setting a learning rate of 0.001 and using a cross-entropy loss function (Eq. 7) to decrease classification mistakes. Training is accomplished across numerous epochs with initial stopping to evade overfitting based on validation loss. This iterative training process certifies vigorous model learning, balancing accuracy with generalizability.

4.2. Experimental Setup

This section describes the simulation tools applied for the experimental work. Table 3 indicates the model parameters applied in the suggested model of immoral contents. The experiment's model settings are intended to enhance performance on tasks, including the classification of social media content. Initially, the embedding dimensions contain 768 for BERT and 300 for Word2Vec, balancing traditional word features with deep contextual embeddings. The BiLSTM layer covers 128 hidden units per direction, permitting the model to extract dependencies both before and after each term. A lenient attention mechanism is combined, serving the model to allocate higher weights to important tokens within each comment or post.

The batch size is set at 32, meaning that 32 samples are managed concurrently throughout each training repetition. The Adam optimizer is employed to dynamically adapt learning, and a learning rate of 0.001 that reduces over epochs permits stable and slow optimization. A dropout rate of 0.5 is employed to evade overfitting. An extreme sequence length of 128 tokens is set for each post, which decreases calculating strain while still giving suitable context. The model is trained through 10 to 20 epochs, with initial stopping to halt training when outcomes stabilize. This certifies excellent accuracy while utilizing resources proficiently.

4.3. Evaluation Criteria

The subsequent metrics are applied to assess the suggested model's performance. Eq. 8-11 is the evaluation metrics of the proposed model. Meanwhile, TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively. The TP are cases where the model correctly categorizes an unethical post as unethical. TN are cases where an ethical post is properly categorized as ethical. The FP are cases where an ethical post is wrongly categorized as unethical. FN are cases where an unethical post is erroneously categorized as ethical.

Table 3. Model parameters used in this study

Parameter	Value	Description
Embedding Dimension	768 (BERT), 300 (Word2Vec)	Dimensionality of BERT and Word2Vec embeddings
BiLSTM Hidden Units	128	Number of units in each direction of BiLSTM
Attention Mechanism	Soft attention	Applies weights to focus on key tokens
Batch Size	32	Number of samples processed at once
Learning Rate	0.001 (decayed over epochs)	Optimizer learning rate
Optimizer	Adam	Adaptive moment estimation optimizer
Dropout Rate	0.5	Dropout applied to prevent overfitting
Max Sequence Length	128 tokens	Maximum token length for each post
Epochs	10–20 (with early stopping)	Number of training iterations

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = \frac{TP}{TP + FP} \tag{10}$$

$$F1 - score = 2 . \frac{Precision \times Recall}{Precision + Recall} \tag{11}$$

5. RESULTS

This section indicates a detailed discussion of different studies on the proposed model's significance.

5.1. Results and Discussion

Table 4 represents the results of proposed fine-tuned BERT-based models on two benchmark datasets: SARC, which targets sarcasm identification, and HatEval, which emphasizes hate speech identification. For the standard BERT fine-tuned model, we perceive reasonable results with precision, recall, and F1-score values near the mid-70s for both datasets, replicating its capability to seize some nuances in hate speech and sarcasm but with some boundaries. When integrating further embeddings, like Word2Vec and GloVe, the model's outcomes expressively advance. The BERT-based Word2Vec model attains great precision, recall, and F1 scores on the SARC dataset (95.68%, 96.85%, and 96.26%, correspondingly), representing improved sarcasm identification abilities. In the same way, the BERT-based GloVe model on HatEval produces brilliant outcomes, with scores above 96%, presenting robust hate speech recognition. These consequences highlight the usefulness of assimilating traditional embeddings with BERT in refining the model's capability to recognize unethical posts. Figure 2 indicates the accuracy of training and testing for the SARC dataset. Figure 3 shows the training and testing accuracy of the HatEval dataset.

5.2. Comparison with baseline approaches

Table 5 shows the comparison results of the proposed model with baseline techniques. The performance of different models across diverse datasets reported the efficiency of different methods for tasks such as social media evaluation and cyberbullying discovery. The MHA-BiLSTM model [46] accomplishes precision, recall, and an F1-score on the SARC dataset of 60.26%, 53.71%, and 56.79%, respectively, showing a balanced but modest outcome. The RSGNN model [47], assessed on the HatEval dataset, validates a greater level of accuracy with precision, recall, and F1-score, all nearly 74%, signifying its ability in this field. The BanglaBERT model [48], concentrated on cyberbullying recognition, displays superior usefulness, with precision at 85.80%, recall at 90.0%, and an F1-score of 87.85%.

Table 4. Results of proposed models

Models	Dataset	Precision (%)	Recall (%)	F1-Score (%)
Fine-tuned Bert Models	SARC	76.12	78.45	77.27
Fine-tuned Bert Models	HatEval	73.62	78.24	75.86
Fine-tuned Bert-based Word2Vec	SARC	95.68	96.85	96.26
Fine-tuned Bert-based GloVe	HatEval	96.65	97.75	97.20

Table 5. Comparison with baseline studies

Study	Method	Dataset	Precision (%)	Recall (%)	F1- Score (%)
[47]	MHA-BiLSTM	SARC	60.26	53.71	56.79
[48]	RSGNN	HatEval	74.29	74.14	74.04
[49]	BanglaBERT	Cyberbullying	85.80	90.0	87.85
[50]	LSTM-SSA model	HatEval	0.920	0.955	0.937
[51]	Hybrid Auto-Encoder-Based Model	SARC	0.83	0.85	0.84
Proposed	Fine-tuned Bert-based Word2Vec	SARC	82.36	83.23	82.79
		HatEval	78.75	90.13	84.06
	Fine-tuned Bert-based GloVe	SARC	95.68	96.85	96.26
		HatEval	96.65	97.75	97.20

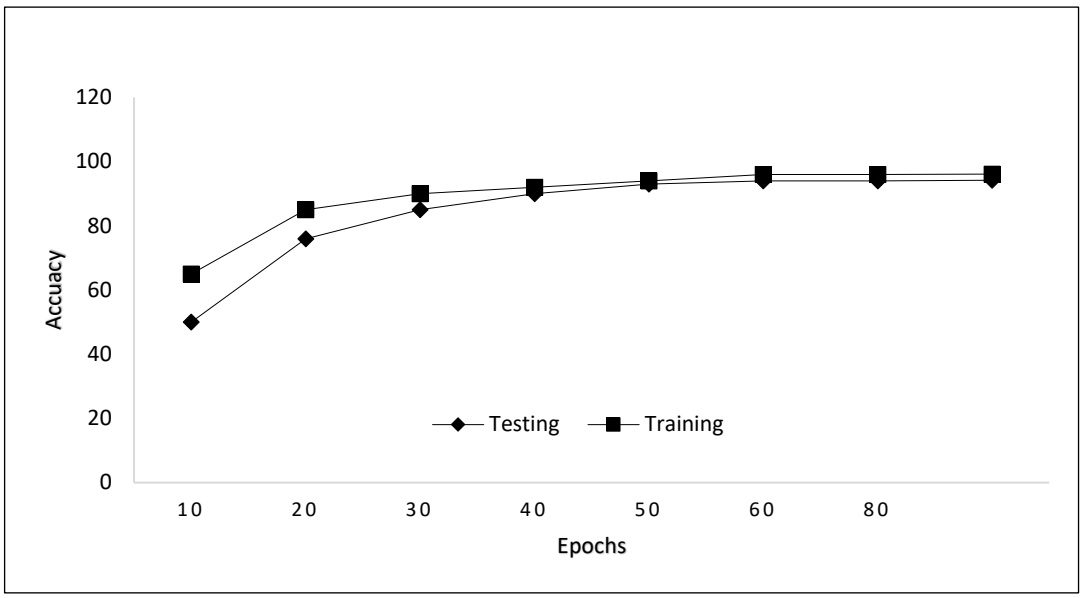

Figure 2. Training and testing accuracy for the SARC dataset

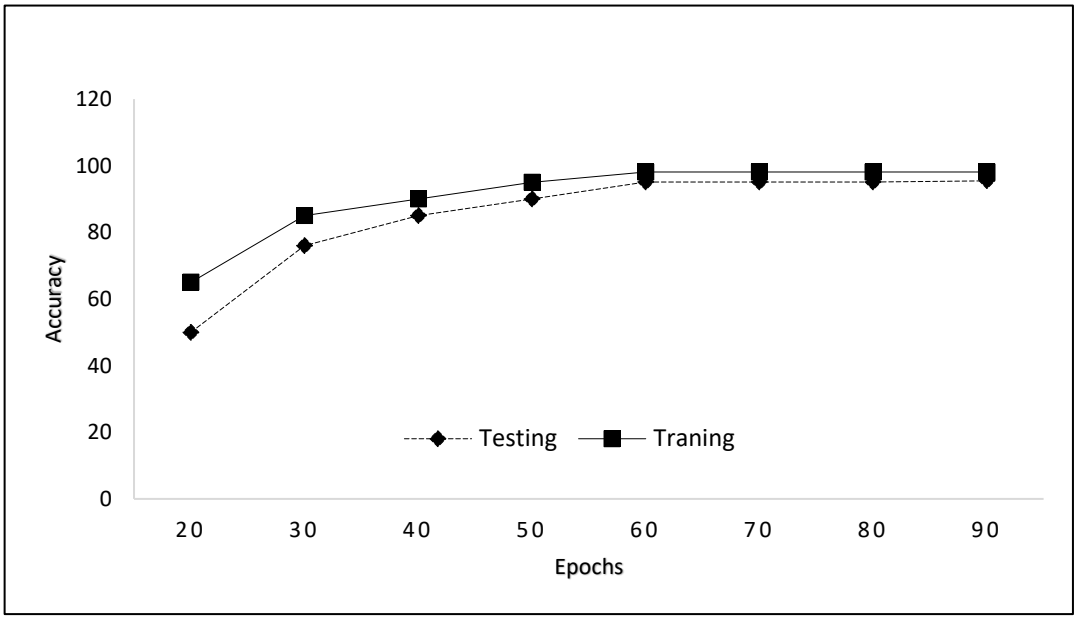

Figure 3. Training and testing accuracy for the HatEval dataset

The LSTM-SSA model [49] develops on these outcomes, particularly for the HatEval dataset, where it reaches an inspiring precision of 92%, recall of 95.5%, and F1-score of 93.7%, representing strength for unwanted data on social networks. The tasks on social media networks. Likewise, a Hybrid Auto-Encoder-Based Model [50] employed in the SARC dataset reports precision and recall values of 83% and 85%, correspondingly, with an F1-score of 84%, presenting consistent performance. The proposed technique, using fine-tuned BERT-based embeddings, outperforms prior models across several metrics. When merged with Word2Vec, it achieves a precision of 82.36%, recall of 83.23%, and F1-score of 82.79% on SARC, whereas on HatEval, it produces precision, recall, and F1 scores of 78.75%, 90.13%, and 84.06%, correspondingly, representing its flexibility. Employing GloVe embeddings, the fine-tuned BERT model considerably increases performance, reaching outstanding scores on both datasets. It records precision at 95.68%, recall at 96.85%, and an F1-score of 96.26% for SAR. It shines more with precision, recall, and F1 scores at 96.65%, 97.75%, and 97.20% on HatEval, underlining the model's improved ability to handle intricate linguistic nuances efficiently.`

Figure 4 represents a comparison of the proposed mode against baseline methods. The graph compares the outcomes of different models tested on benchmark datasets like SARC and HatEval. Each model's efficiency is exposed in identifying unethical, ironic, or detestable posts. The suggested Fine-tuned Bert-based

Word2Vec and Fine-tuned Bert-based GloVe are assessed on the SARC dataset. The proposed model shows maximum scores, signifying robust mockery recognition abilities, emphasizing the benefits of merging contextual embeddings and traditional. Consequently, the graph proves that the proposed model effectively perceives difficult social media terms, highlighting the position of embedding integration and attention mechanisms.

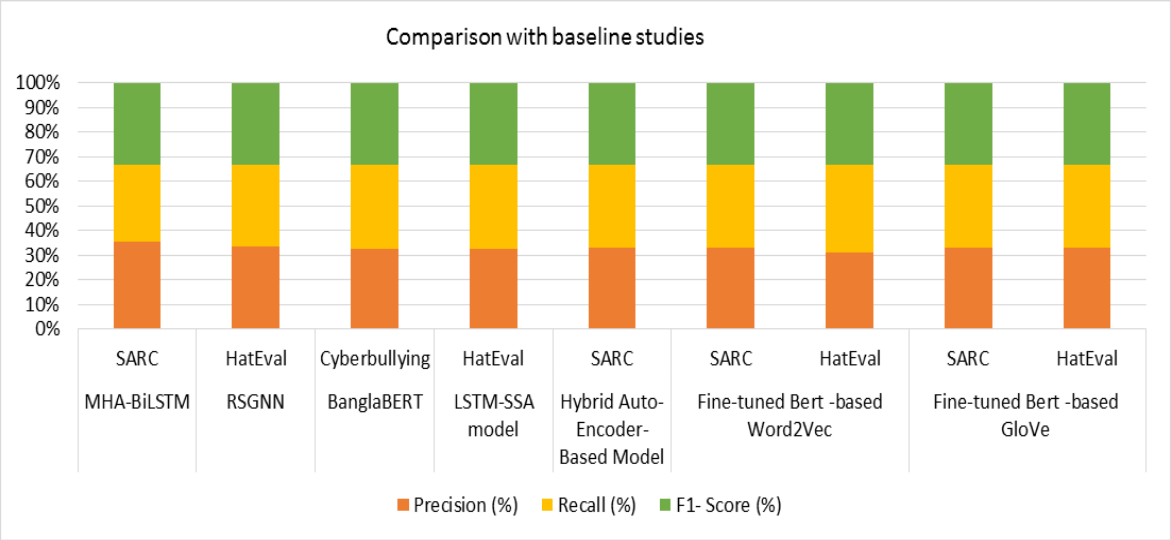

Figure 4. Comparison with closely related works

Table 6. Complexity of the proposed mode

Aspect	Details
Model Architecture	Fine-tuned BERT
Parameter Count	110 million
Memory Requirements	High
Training Duration	Moderate to high
Inference Speed	Moderate
Batch Size	16-64
GPU Utilization	High
CPU Utilization	Moderate
Scalability	Efficient with larger datasets
Real-time Applicability	Feasible with high-performance GPUs
Resource Efficiency	Demands optimized hardware for best results
Data Preprocessing Time	Low to moderate
Model Fine-tuning Cost	Moderate to high

Table 6 summarizes the scalability features and essential resource requirements, showing the trade-offs involved in using the BERT model for the proposed work. The complexity of the proposed DL-based technique for detecting depraved posts on social media is carefully inspected, covering computational, scalability, and resource requirements. This technique employs fine-tuned BERT models, demanding considerable resources due to the model's complicated architecture and the requirement for great processing influence. Despite the model's great requirements, it displays potential in scaling to larger datasets without a visible reduction in performance. The technique is a feasible solution for real-time unethical post recognition on web content due to its balance between resource efficiency and performance.

6. CONCLUSIONS AND FUTURE DIRECTIONS

In the proposed study, we offered an advanced technique for detecting unethical content on social networks. We assimilated fine-tuned BERT embeddings with word embeddings and a Bidirectional Long Short-Term Memory (BiLSTM) network. The model is evaluated on the SARC and HatEval datasets, capturing nuances in hate speech, sarcasm, and immoral posts efficiently. The suggested model validated greater

outcomes in categorizing moral and immoral content compared to traditional ML approaches and state-of-the-art DL techniques. The integration of contextual embeddings empowered the model to efficiently capture the nuances and difficulties of social media content. The consequences of the proposed model highlight the latent of the DL model in addressing the key problems in digital spaces. By leveraging advanced models BERT, we have developed a system that achieves maximum F1 score.

Despite the promising results, numerous avenues for future study remain unexplored. Emerging real-time schemes for identifying unsuitable content can improve instant interference policies on the web. Furthermore, future work could discover the incorporation of multimodal data, containing images and videos, alongside text to deliver a more inclusive understanding of immoral content. Future research should focus on evolving models that can modify to new slang, cultural references, and abbreviations. Future research can expand on the results of this study and help create more complex, flexible, and efficient systems for controlling objectionable information in digital spaces by pursuing these avenues.

DATA AVAILABILITY STATEMENT

The original data presented in the study are openly available in Kaggle at https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset
https://www.kaggle.com/datasets/danofer/sarcasm

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest in this work.

REFERENCES

[1] P. M. Valkenburg, "Social media use and well-being: What we know and what we need to know," *Curr. Opin. Psychol.*, vol. 45, p. 101294, Jun. 2022, doi: 10.1016/j.copsyc.2021.12.006.

[2] B. N. R. B Narasimha Rao, "A Study on Positive and Negative Effects of Social Media on Society," *J. Sci. Technol.*, vol. 7, no. 10, pp. 46–54, Dec. 2022, doi: 10.46243/jst.2022.v7.i10.pp46-54.

[3] V. Mercan, A. Jamil, A. A. Hameed, I. A. Magsi, S. Bazai, and S. A. Shah, "Hate Speech and Offensive Language Detection from Social Media," in *2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*, IEEE, Oct. 2021, pp. 1–5. doi: 10.1109/ICECube53880.2021.9628255.

[4] L. Charmaraman, O. Sode, and D. Bickham, "Adolescent Mental Health Challenges in the Digital World," in *Technology and Adolescent Health*, Elsevier, 2020, pp. 283–304. doi: 10.1016/B978-0-12-817319-0.00012-8.

[5] J. Lee and S. Kim, "Social media advertising: The role of personal and societal norms in page like ads on Facebook," *J. Mark. Commun.*, vol. 28, no. 3, pp. 329–342, Apr. 2022, doi: 10.1080/13527266.2019.1658466.

[6] V. U. Gongane, M. V. Munot, and A. D. Anuse, "Detection and moderation of detrimental content on social media platforms: current status and future directions," *Soc. Netw. Anal. Min.*, vol. 12, no. 1, p. 129, Dec. 2022, doi: 10.1007/s13278-022-00951-3.

[7] K. M. Hana, Adiwijaya, S. Al Faraby, and A. Bramantoro, "Multi-label Classification of Indonesian Hate Speech on Twitter Using Support Vector Machines," in *2020 International Conference on Data Science and Its Applications (ICoDSA)*, IEEE, Aug. 2020, pp. 1–7. doi: 10.1109/ICoDSA50139.2020.9212992.

[8] A. H. Razavi, D. Inkpen, S. Uritsky, and S. Matwin, "Offensive Language Detection Using Multi-level Classification," 2010, pp. 16–27. doi: 10.1007/978-3-642-13059-5_5.

[9] M. van Dieijen, A. Borah, G. J. Tellis, and P. H. Franses, "Big Data Analysis of Volatility Spillovers of Brands across Social Media and Stock Markets," *Ind. Mark. Manag.*, vol. 88, pp. 465–484, Jul. 2020, doi: 10.1016/j.indmarman.2018.12.006.

[10] T. Dias Oliva, "Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression," *Hum. Rights Law Rev.*, vol. 20, no. 4, pp. 607–640, Dec. 2020, doi: 10.1093/hrlr/ngaa032.

[11] A. Baccouche, S. Ahmed, D. Sierra-Sosa, and A. Elmaghraby, "Malicious Text Identification: Deep Learning from Public Comments and Emails," *Information*, vol. 11, no. 6, p. 312, Jun. 2020, doi: 10.3390/info11060312.

[12] S. Tongman and N. Wattanakitrungroj, "Classifying Positive or Negative Text Using Features Based on Opinion Words and Term Frequency - Inverse Document Frequency," in *2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, IEEE, Aug. 2018, pp. 159–164. doi: 10.1109/ICAICTA.2018.8541274.

[13] S. Akuma, T. Lubem, and I. T. Adom, "Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets," *Int. J. Inf. Technol.*, vol. 14, no. 7, pp. 3629–3635, Dec. 2022, doi: 10.1007/s41870-022-01096-4.

[14] S. Varghese, "Dynamics of Social Media Networks in the Post-Truth Era," in *Handbook of Digital Journalism*, Singapore: Springer Nature Singapore, 2024, pp. 419–429. doi: 10.1007/978-981-99-6675-2_36.

[15] A. Toktarova *et al.*, "Hate Speech Detection in Social Networks using Machine Learning and Deep Learning Methods," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 5, 2023, doi: 10.14569/IJACSA.2023.0140542.

[16] H. Elfaik and E. H. Nfaoui, "Deep Bidirectional LSTM Network Learning-Based Sentiment Analysis for Arabic Text," *J. Intell. Syst.*, vol. 30, no. 1, pp. 395–412, Dec. 2020, doi: 10.1515/jisys-2020-0021.

[17] L. and A. P.-M. Hasimi, "Detection of disinformation and content filtering using machine learning: implications to human rights and freedom of speech," *ROMCIR@ ECIR*, 2024.

[18] A. Khan, M. Z. Asghar, H. Ahmad, F. M. Kundi, and S. Ismail, "A Rule-Based Sentiment Classification

Framework for Health Reviews on Mobile Social Media," *J. Med. Imaging Heal. Informatics*, vol. 7, no. 6, pp. 1445–1453, Oct. 2017, doi: 10.1166/jmihi.2017.2208.

[19] K. Crawford and T. Gillespie, "What is a flag for? Social media reporting tools and the vocabulary of complaint," *New Media Soc.*, vol. 18, no. 3, pp. 410–428, Mar. 2016, doi: 10.1177/1461444814543163.

[20] D. Saputra, V. S. Damayanti, Y. Mulyati, and W. Rahmat, "Expressions of the use of slang among millennial youth on social media and its impact of the extension of Indonesia in society," *BAHASTRA*, vol. 43, no. 1, pp. 21–40, Apr. 2023, doi: 10.26555/bs.v43i1.325.

[21] J. Grimmer and B. M. Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Polit. Anal.*, vol. 21, no. 3, pp. 267–297, Jan. 2013, doi: 10.1093/pan/mps028.

[22] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2017, pp. 1–10. doi: 10.18653/v1/W17-1101.

[23] N. S. Mullah and W. M. N. W. Zainon, "Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review," *IEEE Access*, vol. 9, pp. 88364–88376, 2021, doi: 10.1109/ACCESS.2021.3089515.

[24] J. Preetham and J. Anitha, "Offensive Language Detection in Social Media Using Ensemble Techniques," in *2023 International Conference on Circuit Power and Computing Technologies (ICCPCT)*, IEEE, Aug. 2023, pp. 805–808. doi: 10.1109/ICCPCT58313.2023.10245673.

[25] Warner, W. and J. Hirschberg, "Detecting hate speech on the world wide web," *Proc. Second Work. Lang. Soc. media*, 2012.

[26] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 11, no. 1, pp. 512–515, May 2017, doi: 10.1609/icwsm.v11i1.14955.

[27] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep Learning for Hate Speech Detection in Tweets," in *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, New York, New York, USA: ACM Press, 2017, pp. 759–760. doi: 10.1145/3041021.3054223.

[28] A. Kumar, "A Study: Hate Speech and Offensive Language Detection in Textual Data by Using RNN, CNN, LSTM and BERT Model," in *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, May 2022, pp. 1–6. doi: 10.1109/ICICCS53718.2022.9788347.

[29] S. F. Ahmed *et al.*, "Deep learning modelling techniques: current progress, applications, advantages, and challenges," *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 13521–13617, Nov. 2023, doi: 10.1007/s10462-023-10466-8.

[30] Y. Kim, "Convolutional Neural Networks for Sentence Classification," Aug. 2014, doi: https://doi.org/10.48550/arXiv.1408.5882.

[31] B. Jang, I. Kim, and J. W. Kim, "Word2vec convolutional neural networks for classification of news articles and tweets," *PLoS One*, vol. 14, no. 8, p. e0220976, Aug. 2019, doi: 10.1371/journal.pone.0220976.

[32] G. K. Pitsilis, H. Ramampiaro, and H. Langseth, "Effective hate-speech detection in Twitter data using recurrent neural networks," *Appl. Intell.*, vol. 48, no. 12, pp. 4730–4742, Dec. 2018, doi: 10.1007/s10489-018-1242-y.

[33] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[34] C. Raj, A. Agarwal, G. Bharathy, B. Narayan, and M. Prasad, "Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural Language Processing Techniques," *Electronics*, vol. 10, no. 22, p. 2810, Nov. 2021, doi: 10.3390/electronics10222810.

[35] I. R. and I. A. H. Naseem, Usman, "Deep Context-Aware Embedding for Abusive and Hate Speech detection on Twitter," *Aust. J. Intell. Inf. Process. Syst*, vol. 15, no. 3, pp. 69–76, 2019.

[36] M. S. Lekshmi, A. Mariya Shaji, and S. K. Amrita, "Cyberbullying Detection Using BiLSTM Model," 2024, pp. 339–350. doi: 10.1007/978-3-031-47942-7_29.

[37] C. P. Soto, G. M. S. Nunes, J. G. R. C. Gomes, and N. Nedjah, "Application-specific word embeddings for hate and offensive language detection," *Multimed. Tools Appl.*, vol. 81, no. 19, pp. 27111–27136, Aug. 2022, doi: 10.1007/s11042-021-11880-2.

[38] S. Alshattnawi, A. Shatnawi, A. M. R. AlSobeh, and A. A. Magableh, "Beyond Word-Based Model Embeddings: Contextualized Representations for Enhanced Social Media Spam Detection," *Appl. Sci.*, vol. 14, no. 6, p. 2254, Mar. 2024, doi: 10.3390/app14062254.

[39] S. S, U. S, N. Abinaya, J. P, S. Priyanka, and D. M N, "A Comparative Exploration in Text Classification for Hate Speech and Offensive Language Detection Using BERT-Based and GloVe Embeddings," in *2024 2nd International Conference on Disruptive Technologies (ICDT)*, IEEE, Mar. 2024, pp. 1506–1509. doi: 10.1109/ICDT61202.2024.10489019.

[40] M. Khodak, N. Saunshi, and K. Vodrahalli, "A Large Self-Annotated Corpus for Sarcasm," Apr. 2017, doi: 10.48550/arXiv.1704.05579.

[41] Ò. Garibo i Orts, "Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter at SemEval-2019 Task 5: Frequency Analysis Interpolation for Hate in Speech Detection," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 460–463. doi: 10.18653/v1/S19-2081.

[42] H. Saleh, A. Alhothali, and K. Moria, "Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model," *Appl. Artif. Intell.*, vol. 37, no. 1, Dec. 2023, doi: 10.1080/08839514.2023.2166719.

[43] Y. Li and T. Yang, "Word Embedding for Understanding Natural Language: A Survey," 2018, pp. 83–104. doi: 10.1007/978-3-319-53817-4_4.

[44] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space,"

Jan. 2013, Available: http://arxiv.org/abs/1301.3781
[45] Pennington, J., R. Socher, and C.D. Manning, "Glove: Global vectors for word representation," *Proc. 2014 Conf. Empir. methods Nat. Lang. Process.*, 2014.
[46] K. Sreelakshmi, B. Premjith, B. R. Chakravarthi, and K. P. Soman, "Detection of Hate Speech and Offensive Language CodeMix Text in Dravidian Languages Using Cost-Sensitive Learning Approach," *IEEE Access*, vol. 12, pp. 20064–20090, 2024, doi: 10.1109/ACCESS.2024.3358811.
[47] A. Kumar, V. T. Narapareddy, V. Aditya Srikanth, A. Malapati, and L. B. M. Neti, "Sarcasm Detection Using Multi-Head Attention Based Bidirectional LSTM," *IEEE Access*, vol. 8, pp. 6388–6397, 2020, doi: 10.1109/ACCESS.2019.2963630.
[48] R. Song, F. Giunchiglia, Q. Shen, N. Li, and H. Xu, "Improving Abusive Language Detection with online interaction network," *Inf. Process. Manag.*, vol. 59, no. 5, p. 103009, Sep. 2022, doi: 10.1016/j.ipm.2022.103009.
[49] K. Saifullah, M. I. Khan, S. Jamal, and I. H. Sarker, "Cyberbullying Text Identification based on Deep Learning and Transformer-based Language Models," *EAI Endorsed Trans. Ind. Networks Intell. Syst.*, vol. 11, no. 1, Feb. 2024, doi: 10.4108/eetinis.v11i1.4703.
[50] Kumar, A. and S. Kumar, "Optimized Deep Neural Networks Using Sparrow Search Algorithms for Hate Speech Detection," *Int. J. Comput. Digit. Syst.*, vol. 15, no. 1, pp. 1–9, 2024.
[51] D. K. Sharma, B. Singh, S. Agarwal, H. Kim, and R. Sharma, "Sarcasm Detection over Social Media Platforms Using Hybrid Auto-Encoder-Based Model," *Electronics*, vol. 11, no. 18, p. 2844, Sep. 2022, doi: 10.3390/electronics11182844.

BIOGRAPHIES OF AUTHORS

Bibi Saqia received the (MSCS) degree in Computer Science from the University of Science and Technology Bannu in 2018. Her PhD is in progress at the University of Science and Technology Bannu. She has more than 7 publications in various reputed journals and conferences, including IEEE Transactions. Her research interests include data mining, machine learning, and artificial intelligence. She can be contacted at email: saqiaktk@ustb.edu.pk

Khairullah Khan received a PhD degree in information technology from Universiti Teknologi PETRONAS, Malaysia, in 2012, where he worked on machine learning for the automatic detection of opinion targets from the text. He is currently a Professor at the Department of Computer Science, University of Science and Technology, Bannu, Pakistan. He has more than 45 publications in various reputed journals and conferences, including IEEE Transactions. His research interests include Data Mining, Web Mining, Opinion Mining, Machine Learning and Artificial Intelligence. He can be contacted at email: khairullah@ustb.edu.pk

Atta Ur Rahman received the BS (Hons) degree in Telecommunication from USTB in 2014 and the MS degree in Computer Science from the same university. He obtained his PhD degree in Computer Science in 2022 from Ghulam Ishaq Khan (GIK) Institute of Engineering Sciences and Technology, Pakistan. Rahman's research interests include data mining, computer vision, human-computer interaction, NLP, and computational intelligence. He has various publications in reputed journals, including IEEE Transactions. He worked as an Assistant Professor at Riphah Institute of System Engineering (RISE), Riphah International University Islamabad, Pakistan. Currently, his postdoc is in progress at the Interdisciplinary Research Centers for Finance and Digital Economy, King Fahd University of Petroleum & Minerals (KFUPM), Dhahran, Saudi Arabia. He has more than 20 publications in various reputed journals and conferences, including IEEE Transactions. His research interests include Human-computer Interaction, Artificial Intelligence in healthcare, and Federated learning for privacy preservation. He can be contacted at email: attaur.rahman@kfupm.edu.sa

Wahab Khan received an M.S. degree in computer science from the University of Science and Technology, Bannu, Pakistan, in 2009. He is currently pursuing a PhD degree in computer science at the International Islamic University Islamabad, Pakistan. He has more than 20 publications in various reputed journals and conferences, including IEEE Transactions. His research interests include natural language processing, machine learning, deep learning and data mining. He can be contacted at email: wahabshri@gmail.com