**DOI: 10.37943/22SKSG8575**

**Aziza Zhidebayeva**
Candidate of Technical Sciences, Senior Lecturer, Department of Computer Science and Mathematics
aziza_68.kz@mail.ru, orcid.org/0000-0002-3768-4835
Academician A.Kuatbekov Peoples' Friendship University, Kazakhstan

**Sabira Akhmetova**
Candidate of Physical and Mathematical Sciences, Associate Professor, Department of Information system
sabdas65@mail.ru, orcid.org/0000-0001-5164-2028
Mukhtar Auezov South Kazakhstan University, Kazakhstan

**Satmyrza Mamikov**
Candidate of Pedagogical Sciences, Associate Professor, Department of Computer Science and Mathematics
satmyrza_mamikov@mail.ru, orcid.org/0009-0009-3476-2920
Academician A.Kuatbekov Peoples' Friendship University, Kazakhstan

**Mukhtar Kerimbekov**
Candidate of pedagogical Sciences, Associate Professor, Department of Computer Science and Mathematics
mukhtar_m@mail.ru, orcid.org/0009-0002-3310-1556
Academician A.Kuatbekov Peoples' Friendship University, Kazakhstan

**Sapargali Aldeshov**
Candidate of Pedagogical Sciences, Associate Professor, Department of Computer Science and Mathematics
aldeshov4@mail.ru, orcid.org/0000-0001-7735-2299
Ozbekali Zhanibekov South Kazakhstan Pedagogical University, Kazakhstan

**Guldana Shaimerdenova**
PhD, Associate Professor, Department of Information Communication Technologies
danel101kz@gmail.com, orcid.org/0000-0001-8685-7125
Mukhtar Auezov South Kazakhstan University, Kazakhstan

# DETECTION OF HATE SPEECH ON SOCIAL MEDIA UTILIZING MACHINE LEARNING

**Abstract:** This article investigates the identification of hate speech on social media using machine learning and deep learning techniques. The research uses metrics such as F-measure, AUC-ROC, precision, accuracy, and recall assessing the effectiveness of various tactics. The findings indicate that deep learning models, particularly the bidirectional long short-term memory (BiLSTM) architecture, consistently outperform other methods in categorization tasks. The research emphasizes the importance of sophisticated neural network designs in identifying the intricacies of hostile and offensive content online. The study offers insights for promoting early identification and prevention of cyberbullying, improving secure and inclusive online environments. Future research may explore real-time detection systems, hybrid approaches, or the integration of complementary components to enhance and improve innovative technology in tackling this significant social issue.

Received: 16.12.2024 Accepted: 25.06.2025 Published: 30.06.2025

A sample tweet was annotated by specialists who categorize tweets as hate speech, offensive language, or neutral. The researchers applied shallow learning methodologies and integrated word embeddings like Word2Vec and GloVe to enhance the efficacy of deep learning models. The results indicate that BiLSTM surpasses shallow learning methods in detecting hate speech on Twitter, highlighting the efficacy of deep learning approaches in recognizing and tracking hate speech on social media platforms. When comparing different deep learning and machine learning models on different datasets, the results reveal that deep learning techniques are usually more effective. A reasonably high level of accuracy is achieved by KNN and SVM among classical algorithms, whereas Naïve Bayes performs the poorest. While deep learning approaches provide better results, tree-based models such as Random Forest and Decision Trees offer more consistent accuracy. Models based on neural networks, such as LSTM, CNN, and BI-LSTM, perform well, with LSTM-based methods excelling in particular. The most successful strategy for classification problems is the model presented, which obtains the greatest accuracy, precision, recall, F1-score of 95%. The research aids in the development of advanced tools and methodologies to mitigate hate speech on social media and foster positive online interactions. Future research may investigate alternative deep learning architectures, such as transformers, to enhance hate speech detection efficacy. The advancement of interpretable AI methodologies for identifying hate speech and delivering transparent forecasts might enhance user confidence and facilitate better content moderation decisions.

**Keywords:** hate speech; machine learning; natural language processing; detection; social media.

## Introduction

The goal of this article is to examine how machine learning and deep learning may be used to detect hate speech on social media. In order to improve detection accuracy, it compares different models' performance in identifying hate speech, offensive language, and neutral material, with a focus on advanced neural network architectures.

The study's overarching goal is to help create more effective resources to combat cyberbullying and make the internet a safer place for everyone.

1. Aim of article. Assess Machine Learning and Deep Learning Models Determine how well various categorization approaches, such as deep learning architectures and shallow learning methods, identify hate speech.
2. Assess Performance Metrics: Evaluate different models' ability to classify social media material using precision, recall, accuracy, F-measure, and AUC-ROC.
3. Showcase Deep Learning's Advantages: Prove that BiLSTM and other deep learning models are more effective than more conventional method of detecting hate speech.
4. Improve Model Accuracy with Word Embeddings: Use word embedding methods such as Word2Vec and GloVe to make deep learning models more accurate.
5. Foster Research to Prevent Cyberbullying: Share results that support the creation of early detection systems to counteract cyberbullying and hate speech on the internet.
6. Investigate Possible Future Improvements: Propose Possible Areas for Future Research, Such as Real-Time Detection Systems, Hybrid Models, and Architectures Based on Transformers, to better identify Hate Speech.

This research is different from others that have used BiLSTM to find hate speech because it uses custom Word2Vec embeddings that were trained on domain-specific data. This makes contextual understanding better. Our system also has error analysis and majority-vote annotations, making it a strong path for real-world release.

Social media platforms like Twitter have become an essential communication tool in the digital age, allowing users around the world to share their thoughts, opinions and experiences with a wide audience. However, the rapid expansion of social media has also given rise to negative content, such as hate speech. Hate speech is a communicative act that insults, harms and biases people or groups based on their race, ethnicity, gender, religion or other characteristics. The rise of hate speech on social media is of great concern because it fosters hostility, threatens social cohesion and undermines the foundations of free expression and civil discourse. Therefore, the need for effective systems to detect and monitor hate speech is paramount.

In recent years, machine learning approaches have emerged as a viable way to address the problem of hate speech detection and moderation on social media platforms [1]. Both shallow and deep machine learning methods have proven their potential in solving many natural language processing (NLP) tasks such as sentiment analysis, text classification, and named object recognition. This study aims to investigate and compare the performance of several shallow and deep learning algorithms in detecting hate speech on Twitter. Machine learning algorithms have proven effective in several areas such as spam detection, sentiment analysis, and topic modeling [2]. However, shallow learning algorithms have limitations in understanding the complex semantics and context of natural language, which hinders their ability to effectively detect hate speech. In contrast, deep learning techniques have shown promising results in several NLP tasks due to their ability to model high-level abstractions and distinguish complex language patterns.

LSTMs and BiLSTMs are recurrent neural networks (RNNs) that excel at processing sequential input, making them ideal for learning the temporal structure of text. Originally developed for image classification, convolutional neural networks (CNNs) have shown to be effective in text classification by finding local and global patterns in text data using convolutional filters.

To evaluate the effectiveness of shallow and deep learning methods for detecting hate speech on Twitter, we first collect a diverse and representative tweet dataset, ensuring that it covers a wide range of online communication. The dataset is annotated by experts who classify tweets as hate speech, offensive language or neutral, thus providing a detailed description of the content. We want to better capture the richness and nuance of hate speech by introducing a multi-level tagging approach.

It implements various shallow learning approaches, evaluating their effectiveness in detecting hate speech on the dataset. We further explore the addition of feature selection techniques. Establishing baseline performance for these methods allows us to assess the potential benefits of deep learning methods.

We then run the LSTM, BiLSTM, and CNN models and evaluate their performance against a predefined baseline [3]. We aim to determine the most effective method for detecting hate speech on Twitter by comparing the effectiveness of deep learning methods with shallow learning approaches. In addition to evaluating the overall performance of the models, we also examine their ability to address several issues related to hate speech detection, including context understanding, sarcasm interpretation, and semantic subtleties.

To improve the performance of deep learning models, we incorporate word embeddings such as Word2Vec and GloVe, which allow representing words in a continuous vector space. These embeddings incorporate semantic and syntactic correlations between words, thereby improving the embedding capabilities of our models. By embedding words, we try to improve the models' ability to distinguish between hate speech and other forms of online communication, thereby increasing accuracy and reducing false positives.

Our results show that deep learning techniques, namely BiLSTM, outperform shallow learning techniques in identifying hate speech on Twitter. BiLSTM demonstrates an improved ability to understand the complex characteristics of hate speech by analyzing context, semantic subtleties, and consistent patterns in tweets [4]. This finding demonstrates the effectiveness of deep learning techniques in identifying and monitoring hate speech on social media platforms.

This study provides a thorough analysis of various machine learning techniques for detecting hate speech on Twitter. Our results show that deep learning techniques, in particular BiLSTM, are more promising than shallow learning alternatives for solving this important problem. By building the most accurate hate speech detection models, we contribute to the ongoing effort to create sophisticated tools and techniques to prevent hate speech on social media and promote healthy online relationships. Future research may include exploring additional deep learning architectures such as transformers to improve hate speech detection performance. Additionally, investigating the effects of transfer learning and pre-trained language models such as BERT or GPT on model performance may provide important insights.

This research's innovation extends beyond the traditional application of BiLSTM and pre-trained word embeddings, as it incorporates domain-specific Word2Vec vectors developed from a carefully selected corpus of social media content pertaining to cyberbullying and hate speech. This customized embedding method improves semantic representation, especially for informal and abusive verbal patterns that are inadequately captured in general-purpose embeddings.

Additionally, the annotation pipeline incorporates majority-vote consensus and inter-annotator agreement validation, hence enhancing label quality. The model integrates interpretability factors by examining misclassification patterns, hence offering insights into its decision boundaries. These modifications collectively yield a better context-aware and practically implementable hate speech detection system.

Ultimately, the development of interpretable AI techniques for detecting hate speech and providing transparent predictions can increase user trust and contribute to improved content moderation decision-making.

### *Related works*

Cyberbullying occurs continuously and leaves people feeling unsafe; Messages and comments can come out of nowhere, which has profound psychological consequences for teens. Moreover, the anonymity of the Internet can prevent a teenager from identifying the perpetrator, thus increasing their fear. Unlike physical abuse, the consequences of emotional abuse ultimately affect psychological well-being. It is difficult to identify a victim of emotional abuse. Rapid automatic detection of cyberbullying facilitates its prevention [5]. Consequently, social media companies and politicians have implemented measures against the spread of hate speech.

Machine learning techniques use natural language processing (NLP) technologies to autonomously identify and classify the content of hate speech [6]. Importantly, they do not rely solely on traditional keyword-based approaches, which sometimes fail to adequately identify subtle forms of hate speech [7].

An important advantage of using machine learning (ML) and deep learning (DL) techniques in hate speech detection is their adaptability. Hate speech evolves over time to include new derogatory terminology, symbols, and phrases that may not be accurately recognized by inflexible rule-based systems. Machine learning (ML) and deep learning (DL) models have the ability to continuously learn and adapt to new patterns.

This study provides a thorough examination of several approaches, methodologies, and datasets used in hate speech research, highlighting their individual strengths and weaknesses [8]. By carefully analyzing the intricacies of identifying hate speech, we aim to make a meaningful contribution to the ongoing discourse on this important issue. Furthermore, we aim to provide important insights that will inform future research and development in this particular area [9]. Additionally, we explore methodologies and empirical results, offering a thorough analysis of the effectiveness of machine learning (ML) and deep learning (DL) techniques in identifying hate speech on social media platforms.

The barrier to early detection of cyberbullying on social networking platforms can be fundamentally different from the difficulty of categorizing its various forms. In the setting described here, we define a group of social media interactions called "S". As such, some of these exchanges may constitute instances of cyberbullying. The development of these relations in a certain social network can be briefly described by the following equation

$$S = \{s_1, s_2, \ldots, s_{|S|}\} \tag{1}$$

In this inquiry, the variable "S" represents the total number of sessions, while the variable "i" indicates the current session being analyzed. The sequence of submissions in a session can change at various times due to multiple complex factors.

$$P_s = \left( \langle P_1^S, t_1^S \rangle, \langle P_2^S, t_2^S \rangle, \ldots, \langle P_n^S, t_n^S \rangle \right) \tag{2}$$

In this study, the tuple "P" represents the kth post within the social network session, while "s" is the timestamp marking the exact instant post P was disseminated. A unique set of qualities utilized for the definitive identification of each individual post.

$$P_k^S = \left[ f_{k_1}^S, f_{k_2}^S, \ldots, f_{k_n}^S \right], k \in [1, n] \tag{3}$$

The major objective of this attempt is to gather the necessary knowledge to develop a function called "f," which can identify the correlation between a specific text and the occurrence of hate speech.

A lot of the early work on finding hate speech used simple machine learning methods like support vector machines (SVMs) and naive Bayes models [9], [12], [13]. These methods did a good job of catching direct hate speech, but they had a hard time with more subtle and hidden types of hate speech, like snark, irony, and meanings that change depending on the situation. New deep learning techniques, especially those based on bidirectional long short-term memory (BiLSTM) networks, have made big steps forward in this field by making it easier to predict how text depends on its surroundings and what comes after it [10], [11]. These models are better at understanding complicated language patterns, which leads to more accurate spotting. Recently, researchers have been focusing on mixed designs that take the best parts of several transformer-based models and methods, including BERT, RoBERTa, and their variations.

These transformer models use self-attention to pick up on long-term relationships and semantic details in text. This makes it easier to tell the difference between hate speech and other content. More recently, studies have added external knowledge bases, multidimensional cues, and domain-specific embeddings to make recognition even better. However, it is still hard to stop hate speech that is different in languages, countries, and platforms. This shows how much we need new and flexible modeling methods.

*Existing Research Challenges and Deficits*

Although prior research has made substantial contributions, there are still numerous obstacles that require resolution:

a) Contextual Understanding: Misclassifications are a common result of the difficulty that existing models have in detecting implicit hate speech, sarcasm, and code-switching.
b) Computational Efficiency: Real-time detection is a difficult task due to the computationally costly nature of deep learning models, particularly BiLSTMs.

c) Biased models that underperform in detecting nuanced forms of objectionable content are frequently the result of data imbalances in hate speech datasets.
d) Models that have been trained on specific datasets are unable to generalize effectively across various social media platforms as a result of differences in language usage.
e) Interpretability is a problem due to the fact that many deep learning models operate as black-box systems, which makes it challenging to elucidate the reasons why specific content is classified as hate speech.

*The objective of our investigation is to rectify these deficiencies by:*

a) Presenting an improved BiLSTM-based model that enhances the accuracy of classification and contextual comprehension.
b) Incorporating Word2Vec embeddings to more effectively capture implicit hate speech and semantic relationships.
c) Optimizing computational efficacy to facilitate real-time detection on social media platforms.
d) Assessing the proposed model's generalization capacity across numerous datasets to enhance its robustness.
e) Investigating techniques for explainable AI to improve the interpretability and transparency of hate speech classification.

**Materials and Methods**

This study utilizes a publicly accessible dataset comprising 24,783 tweets, assembled for hate speech recognition research, as cited in [10]. This corpus exhibits significant class disparity, comprising 5% categorized as hate speech, 77% as offensive language, and 18% as neutral content. This imbalance presents intrinsic difficulties for supervised classification problems, frequently skewing models in favor of the majority class.

To guarantee annotation trustworthiness, each tweet was independently categorized by three domain-expert annotators acquainted with the linguistic and cultural background of the data. The conclusive class labels were established by a majority voting mechanism. The inter-annotator agreement was statistically evaluated using Fleiss' Kappa, resulting in a value of $\mathrm{k} = 0.71$. This degree of concordance signifies considerable uniformity among the annotators, hence affirming the reliability and quality of the labeling process. This comprehensive annotation structure guarantees the dataset's appropriateness for the development and assessment of hate speech detection algorithms.

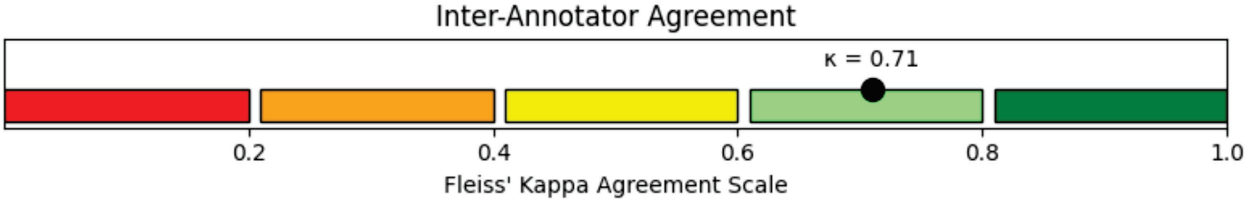

Figure 1. Visualization of Fleiss' Kappa Inter-Annotator Agreement

Fig. 2 provides a visual representation of the established model of hate speech classification. The model consists of separate phases: preprocessing, feature extraction, classification and evaluation. This section provides a thorough examination of each phase, with an emphasis on critical issues.

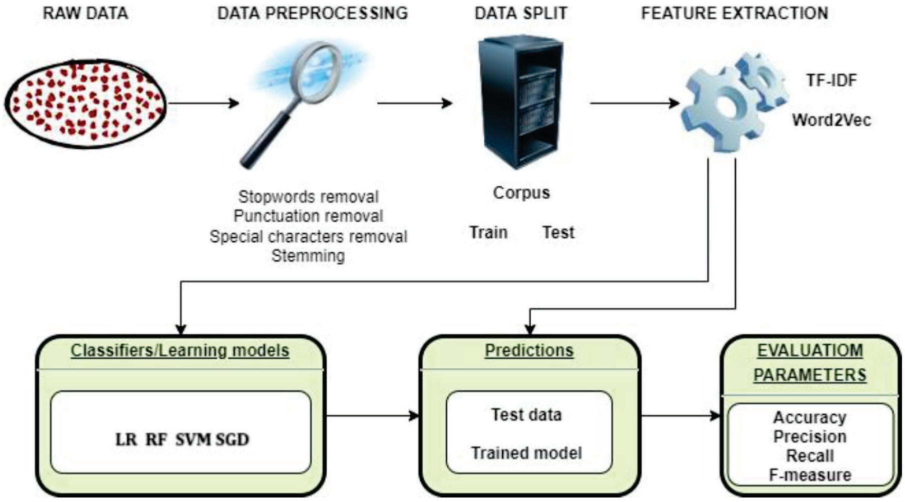

Figure 2. Proposed framework

Word2Vec is a method for displaying common features in natural language processing. It is part of a family of word embedding techniques that transform words into continuous vector representations in a high-dimensional space. Word2Vec interprets the semantic and contextual relationships between words by analyzing a large text corpus [10].

This method allocates a vector to each word and places words with similar meanings close in the vector space [15]. Word2Vec improves NLP tasks by enabling models to understand the context and semantics of words, which is particularly beneficial for applications such as sentiment analysis, document clustering, and information retrieval [11]. Word2Vec improves text analysis and natural language understanding by converting words into vectors.

Bag of Words (BoW) The Bag of Words (BoW) model is a fundamental technique in natural language processing (NLP) and text processing that transforms textual information into numerical data to enable computational algorithms to analyze language. This technique works by creating a dictionary of different words from the corpus and then converting the text documents into vectors, each vector component representing the frequency of a particular word in the document [12]. Despite its simplicity, the BoW model has played an important role in various NLP applications such as document categorization, sentiment analysis, and topic modeling. However, it has its drawbacks; in particular, the model's neglect of word order and context can lead to loss of semantic meaning. In addition, the high dimensionality of the result vectors, especially those with a large vocabulary, hinders the computational efficiency [13]. However, the ease of implementation and clarity of interpretation of the BoW model make it an asset in the early stages of text analysis efforts.

### *Utilizing Machine Learning for the Detection of Hate Speech*

In the field of hate speech detection on social networks, various machine learning models have been used to solve the difficult problem of distinguishing between abusive language and non-abuse information. Each model offers unique advantages and trade-offs that suit different aspects of the problem [14].

Decision trees: Decision tree models offer a systematic representation of decision-making processes. They are interpretable and can be important in identifying clear patterns and characteristics that characterize hate speech [15]. However, they may have difficulty picking up on environmental cues.

Logistic regression facilitates the estimation of probabilities and predictions in scenarios where the outcome is categorical, such as spam email detection or medical diagnosis. The simplicity and interpretability of logistic regression make it a useful tool in a variety of domains, including data analysis, healthcare, and marketing.

$$P\left(y = \frac{1}{x}\right) = \frac{1}{(1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)})} \tag{4}$$

The sigmoid (logistic) function is employed in logistic regression to characterize the likelihood that a given input is a member of a specific class. This function maps any real value to the range (0, 1).

Naive Bayes models are based on probabilistic principles. They are good at handling textual data due to independence assumptions. Naive Bayes models can efficiently process significant amounts of text and adeptly adjust to the high ambiguity of social media information. K-Nearest Neighbors [16] can effectively identify similar messages containing relative hate content.

$$P(C|x) = \frac{P(x|C) * P(C)}{P(x)} \tag{5}$$

Naive Bayes models are founded on probabilistic principles. They are adept at managing textual data as a result of their independence assumptions. Naive Bayes models are capable of processing substantial volumes of text with ease and can effectively adapt to the high degree of ambiguity present in social media information.

Support vector machines (SVMs) can effectively distinguish between complex decision boundaries in hate speech detection. The choice of a machine learning model should take into account the unique attributes of the hate word detection problem, including nuanced hate word occurrence, dimensionality of the textual data, and interpretability requirements.

$$f(x)=\mathrm{sign}(w \cdot x+b) \tag{6}$$

Often, the synthesis of these models through ensemble methods or hybrid techniques is used to exploit their obvious advantages and mitigate their disadvantages, thus increasing the overall efficiency of hate speech detection systems [17].

*Deep Learning for the Detection of Hate Speech*

Deep learning models have become an effective tool for detecting hate speech in social networks due to their ability to understand complex linguistic subtleties and contextual relationships in textual data. Three important deep learning architectures, convolutional neural networks (CNNs), long short-term memory networks (LSTMs), and bi-directional LSTMs (BiLSTMs) have been widely used to address the challenges associated with this task [18].

Convolutional Neural Networks (CNN): Originally developed for image processing, Convolutional Neural Networks (CNN) have been repurposed for text analysis (Figure 3). They use convolutional layers to identify local patterns and feature hierarchies within text. In hate speech detection, CNN is good at distinguishing important textual structures and recognizing short-term dependencies, including n-grams and patterns typical of hate speech [19].

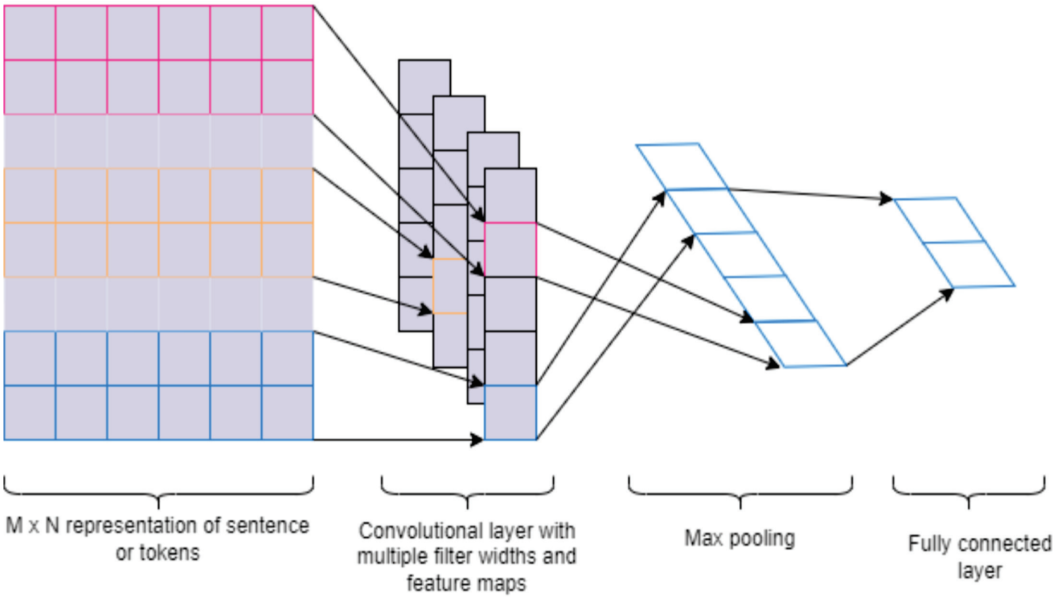
M x N representation of sentence or tokens
Convolutional layer with multiple filter widths and feature maps
Max pooling
Fully connected layer

Figure 3. CNN for Hate Speech Detection

Long Short-Term Memory Networks (LSTMs): LSTMs are recurrent neural networks (RNNs) designed to store sequential information over extended intervals (Figure 3). They are good at modeling temporal dependencies and have proven useful in understanding the evolution of hate speech over time. LSTM can identify contextually relevant information and offer easy text comprehension.

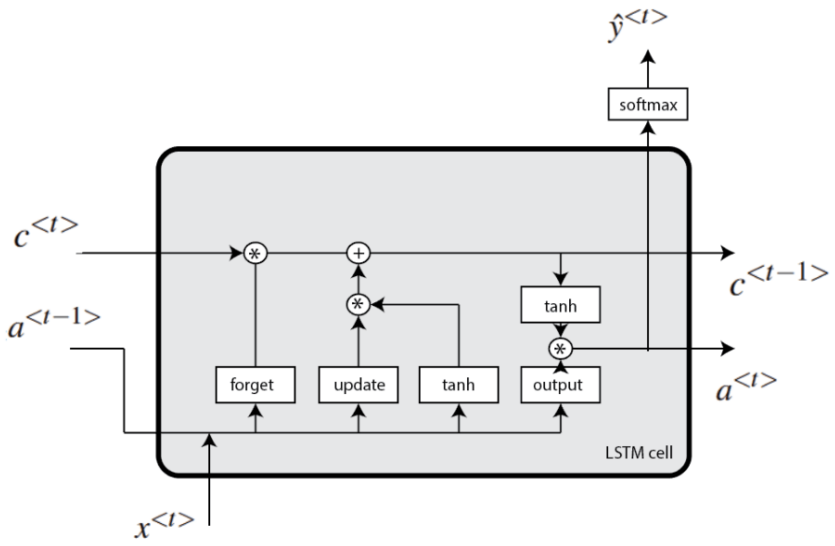

Figure. 4. LSTM for Hate Speech Detection

Bi-directional LSTM (BiLSTM) improves the LSTM architecture by analyzing sequences in both forward and reverse directions, thus allowing the capture of bi-directional dependencies (Fig. 5). BiLSTM is particularly adept at understanding the subtleties of context and identifying the relationships between words in preceding and following contexts in detecting single-speech words [20].

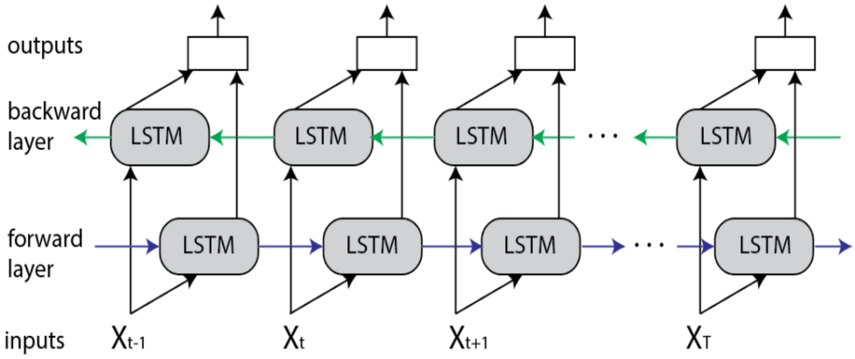

Figure. 5. BiLSTM for Hate Speech Detection

*Evaluation Parameters*

In the realm of hate speech identification on social networks, it is essential to evaluate the performance of machine learning and deep learning models to determine their efficacy in curbing the dissemination of objectionable content. Various evaluation metrics are typically utilized to assess the performance of these models completely.

$$accuracy = \frac{TP + TN}{P + N} \tag{7}$$

$TP$ = True Positives (cases correctly predicted as positive).

$TN$ = True Negatives (cases correctly predicted as negative).

$P$ = Total Positive cases (TP + FN).

$N$ = Total Negative cases (TN + FP).

Accuracy quantifies the ratio of correctly classified occurrences (both positive and negative) to the total instances. Although beneficial for balanced datasets, it can be deceptive in the context of imbalanced datasets, since it fails to distinguish between mistake kinds (e.g., false positives and false negatives).

$$preision = \frac{TP}{TP + FP} \tag{8}$$

$FP$ = False Positives (cases incorrectly predicted as positive).

Precision, also known as Positive Predictive Value, assesses the ratio of real positive predictions to the total instances anticipated as positive. It emphasizes the model's efficacy in minimizing false positives. Precision is essential in situations where false positives entail significant consequences (e.g., spam detection or medical diagnosis).

$$recall = \frac{TP}{TP + FN} \tag{9}$$

$FN$ = False Negatives (cases incorrectly predicted as negative).

Recall quantifies the ratio of accurately recognized positive cases to the total real positive cases by the model. It is crucial in scenarios where overlooking a positive instance (false negative) is vital, such as in the detection of diseases or fraud.

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{10}$$

The F1-score is the harmonic mean of precision and recall, serving as a singular statistic for assessing a model's performance. It is especially advantageous in cases with unequal class distribution, as it reconciles the trade-off between precision and recall. A high F1-score signifies that the model excels in both dimensions.

The BiLSTM model was created using the TensorFlow deep learning framework. A carefully selected collection of hyperparameters was employed to enhance model performance and prevent overfitting.

To strike a decent balance between quick computation and consistent training, 64 mini-batch size was used. This figure guarantees sufficient accuracy of the gradient estimations while nevertheless consuming a fair bit of memory.

The model was trained by running through the training sample ten times. Examining the early halting criteria employed in pilot trials to prevent overfitting and guarantee convergence revealed this figure.

This Adam planner picked up at a pace of 0.001. Especially when combined with techniques for adaptive optimization, this is a common default setting that performs well for steady convergence. The Adam optimizer was chosen as it offers a technique for changing the learning rate and has been demonstrated to perform well when training recurrent neural networks such BiLSTM.

Training was done with a loss rate of 0.5 after the BiLSTM layers. This regularization technique prevents the model from being too proficient at what it does by randomly shutting off half of the neurons in each cycle. This causes it to pick up more robust characteristics.

The work calls for categorizing items into several categories including neutral speech, provocative speech, and hate speech. Categorical cross-entropy served as the loss function to identify the gap between the expected and actual class distributions.

The balance between accuracy and recall is important in hate speech detection, as misidentifying non-hate speech as hate speech (false positives) or failing to identify actual hate speech (false negatives) can lead to significant true effects. Researchers and practitioners may additionally consider domain-specific assessment measures and adjust thresholds according to preferred trade-offs between precision and recall. Comprehensive assessment approaches are critical to the development and implementation of successful hate speech detection systems that promote safe and inclusive online communities.

The BiLSTM model was executed with the TensorFlow framework and trained with a mini-batch size of 64. The training epochs were established at 10, based on preliminary experiments utilizing early stopping conditions to mitigate overfitting. The learning procedure employed the Adam optimizer with a learning rate of 0.001, a configuration esteemed for its stability and convergence efficacy in training recurrent neural networks.

A dropout rate of 0.5 was implemented post-BiLSTM layers to improve generalization by randomly deactivating 50% of neurons in each cycle. The model's objective function was categorical cross-entropy, indicating the multi-class classification characteristic of the task. This setup offered an optimal compromise between learning capacity and processing efficiency, facilitating effective generalization across classes despite the dataset's class imbalance.

## Results

Evaluation metrics are essential to measure the effectiveness of algorithms in categorizing events in cyber threat classification datasets.  The confusion matrices shown in Figure 6 are essential to illustrate the results of these classification methods. They provide a transparent picture of the true distribution of categorization results among different classes.

Using confusion matrices allows researchers to identify true positives, true negatives, false positives, and false negatives, thereby helping to better understand the effectiveness of the model in differentiating between cyberbullying and non-cyberbullying situations. These assessments are critical to improving and updating cyberbullying detection algorithms to increase their accuracy and reliability in addressing the critical issue of online harassment and bullying.

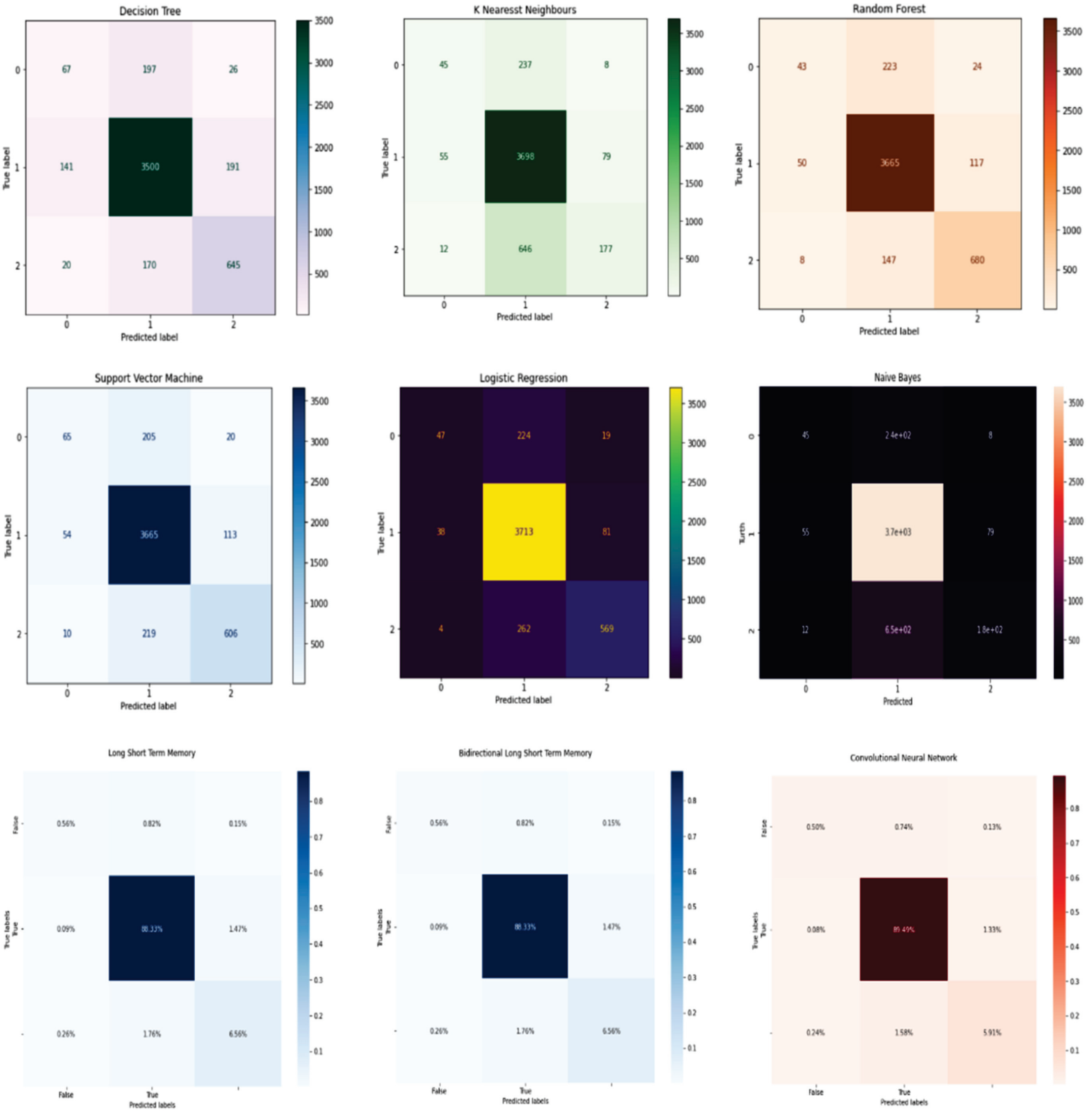

Figure 6. Confusion matrices results in hate speech detection

Fig. 7 provides a comparison of the proposed model with several machine learning and deep learning models used in this work. Performance evaluation in each classification scenario is performed by calculating the area under the receiver operating characteristic curve (AUC-ROC) using all acquired information. This method facilitates careful evaluation of the model's discriminative ability and performance relative to other approaches, providing important insights into its performance on a wide range of classification tasks.

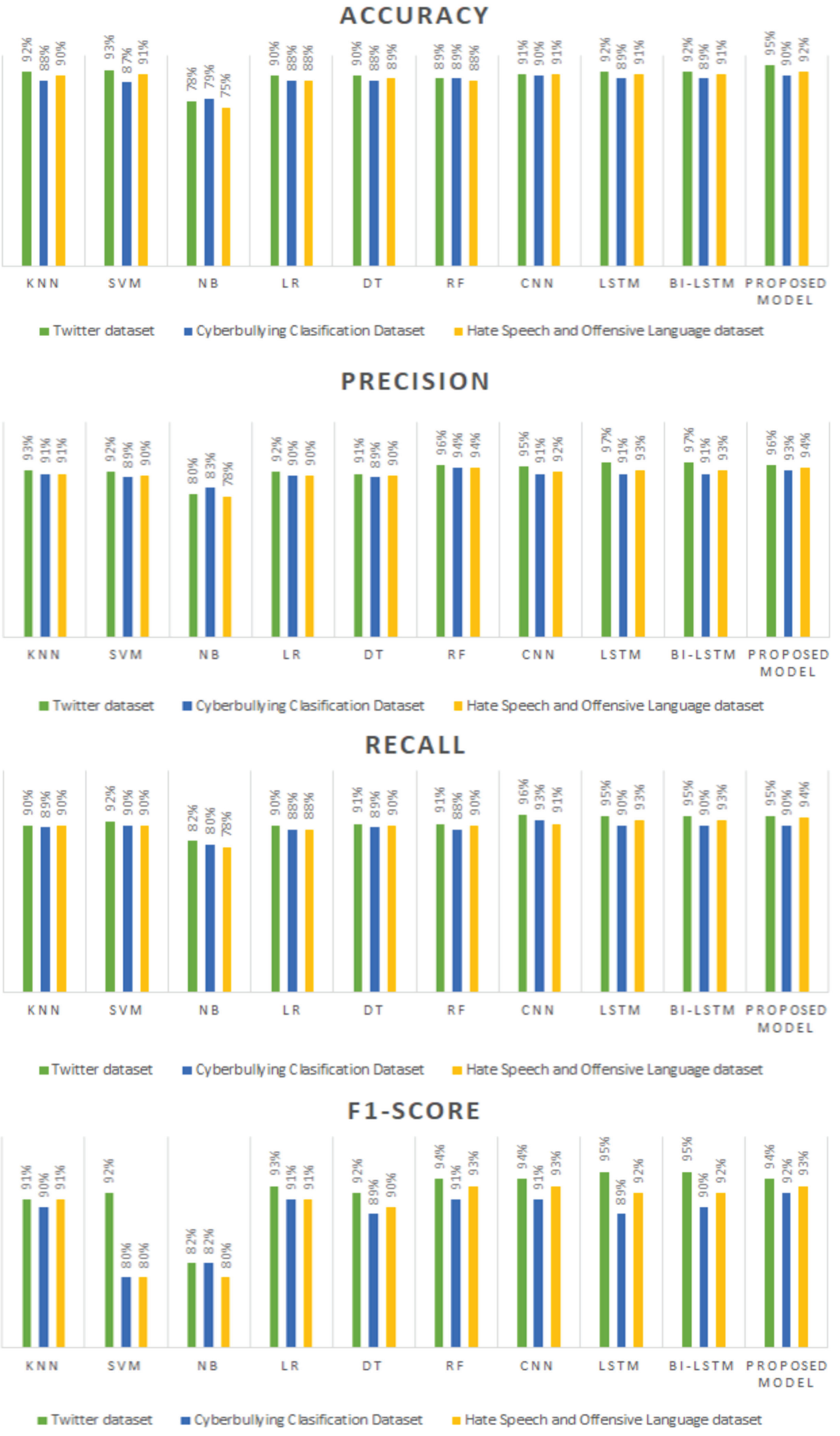

Figure 7. Results in Hate Speech Detection

The findings highlight the effectiveness and reliability of the BiLSTM-based model in accurately distinguishing and categorizing the target classes, hence reinforcing the use of deep learning frameworks in this study.

A study of errors identified distinct patterns in the model's misclassifications. The algorithm notably struggled to accurately identify sarcastic or ironic statements, frequently categorizing them as hate speech due to their superficial lexical resemblance to overtly abusive language. Tweets utilizing satirical commentary on racial injustice were often inaccurately categorized as hostile content.

Likewise, slang and colloquial expressions employed in non-hostile contexts – particularly among peer groups – were often misidentified. The findings underscore the shortcomings of depending exclusively on textual semantics and stress the necessity for multimodal or context-sensitive modeling methods that may more effectively address pragmatic subtleties, cultural allusions, and tone.

## Discussion

The introduction underscores the efficacy of deep learning, specifically BiLSTM, in the identification of hate speech on platforms such as Twitter. Word embeddings, including GloVe and Word2Vec, improve model accuracy by encoding semantic meaning. In addition to the exploration of transformer architectures such as BERT for context-sensitive hate speech identification, real-time detection systems and hybrid models present promising opportunities for future enhancements. Enabling informed content moderation, the development of interpretable AI is essential for fostering transparency and trust. It is essential to address challenges such as dataset biases and ethical considerations in order to construct fair and reliable systems, which will ensure a more inclusive and secure online environment.

In our related work part were shown critical obstacles in the detection of hate speech, such as the interpretability of models, computational efficiency, and contextual comprehension. The research enhances the detection of implicit hate speech and semantic relationships by proposing an improved BiLSTM-based model with Word2Vec embeddings. Real-time detection is facilitated by optimizing computational efficiency, and robustness across platforms is guaranteed by evaluating generalization across datasets. Furthermore, the investigation of AI techniques that are elucidable contributes to the improvement of user trust and transparency in hate speech classification systems. The objective of these developments is to surmount the constraints of existing models, thereby facilitating the development of more adaptable and effective solutions for the prevention of cyberbullying and hate speech on social media.

The materials and methods part highlights the efficacy of machine learning and deep learning methods in detecting hate speech. Our research work demonstrates enhanced contextual comprehension and semantic representation via the use of models like BiLSTM and embedding techniques such as Word2Vec. A survey of more than 200,000 tweets reveals the widespread occurrence of cyberbullying, especially among teenagers. Evaluation measures including precision, recall, and F1-score provide essential insights into model performance, highlighting the need of balancing accuracy and recall. The study's approach tackles critical difficulties such as computing efficiency, interpretability, and flexibility, providing a solid basis for the development of more effective and scalable hate speech detection systems.

The result shows the significance of assessment criteria, especially confusion matrices, in measuring the efficacy of cyberbullying detection systems. These matrices elucidate model performance by detailing true positives, true negatives, false positives, and false negatives.

The suggested model has a number of caveats that need to be considered, even if it has shown encouraging results. To begin with, it can't be used to real-life social media networks that use code-switched or multilingual material because it only uses annotated data in English. Another potential issue is that the model may be too cautious when it comes to identifying hate speech due to the underlying class disparity, which means that occurrences of hate speech are underrepresented. Efforts to guarantee agreement among annotators may not be enough to eliminate the possibility of annotation-related subjectivity, which in turn might affect label consistency.

The study of errors showed that there were problems with correctly identifying ironic or sarcastic statements. Misclassification of ironic criticisms of racial actions as hate speech occurred, for instance, in certain tweets. Similarly, many didn't comprehend the context and took offensive slang statements that their friends used as jokes. The necessity for multimodal techniques that take into account cultural, linguistic, and conversational signals is emphasized by these results.

Notwithstanding its encouraging outcomes, the suggested approach has numerous drawbacks that require attention. The application is limited to English-language data, omitting multilingual and code-switched content frequently encountered in real-world social media contexts. This linguistic limitation restricts the model's applicability across various platforms and demographics.

Secondly, the class imbalance within the dataset may lead to biased predictions, resulting in the model potentially under-identifying occurrences of the minority class, such as hate speech. Furthermore, despite considerable inter-annotator agreement, the intrinsic subjectivity in distinguishing offensive language from hate speech poses a danger of label noise, which may compromise model reliability.

The system may encode hidden societal biases inherent in the training data, potentially resulting in disproportionate misclassifications of minority groups. These problems highlight the imperative for comprehensive, culturally varied datasets and ongoing assessment of equity in hate speech detection methods.

## Conclusion

This research article addressed the important area of identifying cyberbullying in social media. In addition to a thorough study of various machine learning and deep learning techniques, we conducted a thorough evaluation using metrics such as precision, accuracy, recall, F-measure, and AUC-ROC to demonstrate the effectiveness of these methods in solving the complex problem of cyberbullying. to determine.

Our findings highlight the critical importance of deep learning models, particularly the bidirectional long-short-term memory (BiLSTM) architecture, in improving the discriminative ability and accuracy of cyber threat detection systems. The consistent dominance of the BiLSTM-based model in many classification tests confirms the effectiveness of complex neural network architectures in understanding the complexity of online hate and offensive content. Furthermore, the use of confusion matrices and visualizations helped to better understand model performance. This research provides important insight into ongoing efforts to create safe and inclusive online environments where early detection and mitigation of cyberbullying is critical. Future research could explore hybrid methodologies, exploit additional features, or investigate real-time cyberbullying detection systems to further advance the state of the art in this important field.

- Introduced an enhanced BiLSTM-based model that improves classification accuracy and contextual understanding.
- Integrated Word2Vec embeddings to more efficiently capture implicit hate speech and semantic linkages.
- Enhanced computing efficiency to enable real-time detection on social media sites.

- Evaluated the proposed model's generalization ability across many datasets to improve its robustness.
- Examined methodologies for explainable AI to enhance the interpretability and transparency of hate speech categorization.

## References

[1] Alsubait, T., & Alfageh, D. (2021). Comparison of machine learning techniques for cyberbullying detection on youtube arabic comments. *International Journal of Computer Science & Network Security*, *21*(1), 1-5. https://doi.org/10.22937/IJCSNS.2021.21.1.1

[2] Dewani, A., Memon, M.A., & Bhatti, S. (2021). Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data. *Journal of big data*, *8*(1), 160. https://doi.org/10.1186/s40537-021-00550-7

[3] Hall, D.L., Silva, Y.N., Wheeler, B., Cheng, L., & Baumel, K. (2022). Harnessing the power of interdisciplinary research with psychology-informed cyberbullying detection models. *International journal of bullying prevention*, *4*(1), 47-54. https://doi.org/10.1007/s42380-021-00107-5

[4] Arce-Ruelas, K.I., Alvarez-Xochihua, O., Pellegrin, L., Cardoza-Avendaño, L., & González-Fraga, J.Á. (2022). Automatic cyberbullying detection: A Mexican case in high school and Higher Education Students. *IEEE Latin America Transactions*, *20*(5), 770-779. https://doi.org/10.1109/TLA.2022.9693561

[5] Ahmed, M.T., Rahman, M., Nur, S., Islam, A.Z.M.T., & Das, D. (2021). Natural language processing and machine learning based cyberbullying detection for Bangla and Romanized Bangla texts. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *20*(1), 89-97. http://doi.org/10.12928/telkomnika.v20i1.18630

[6] Toktarova, A., Sultan, D., & Azhibekova, Z. (2024, May). Review of Machine Learning Models in Cyberbullying Detection Problem. In *2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST)* (pp. 233-238). IEEE. http://doi.org/10.1109/SIST61555.2024.10629223

[7] Al-Marghilani, A. (2022). Artificial intelligence-enabled cyberbullying-free online social networks in smart cities. *International Journal of Computational Intelligence Systems*, *15*(1), 9. https://doi.org/10.1007/s44196-022-00063-y

[8] Theng, C.P., Othman, N.F., Abdullah, R.S., Anawar, S., Ayop, Z., & Ramli, S.N. (2021). Cyberbullying detection in twitter using sentiment analysis. *International Journal of Computer Science & Network Security*, *21*(11), 1-10. https://doi.org/10.22937/IJCSNS.2021.21.11.1

[9] Sadiq, S., Mehmood, A., Ullah, S., Ahmad, M., Choi, G.S., & On, B.W. (2021). Aggression detection through deep neural model on twitter. *Future Generation Computer Systems*, *114*, 120-129. https://doi.org/10.1016/j.future.2020.07.050

[10] Sarac Essiz, E., & Oturakci, M. (2021). Artificial bee colony–based feature selection algorithm for cyberbullying. *The Computer Journal*, *64*(3), 305-313. https://doi.org/10.1093/comjnl/bxaa066

[11] Gomez, C.E., Sztainberg, M.O., & Trana, R.E. (2022). Curating cyberbullying datasets: A human-AI collaborative approach. *International journal of bullying prevention*, *4*(1), 35-46. https://doi.org/10.1007/s42380-021-00114-6

[12] Salawu, S., Lumsden, J., & He, Y. (2022). A mobile-based system for preventing online abuse and cyberbullying. *International journal of bullying prevention*, *4*(1), 66-88. https://doi.org/10.1007/s42380-021-00115-5

[13] Mladenović, M., Ošmjanski, V., & Stanković, S.V. (2021). Cyber-aggression, cyberbullying, and cyber-grooming: A survey and research challenges. *ACM Computing Surveys (CSUR)*, *54*(1), 1-42. https://doi.org/10.1145/3424246

[14] Sangwan, S.R., & Bhatia, M.P.S. (2021). Denigrate comment detection in low-resource Hindi language using attention-based residual networks. *Transactions on Asian and Low-Resource Language Information Processing*, *21*(1), 1-14. https://doi.org/10.1145/3431729

[15] Yan, R., Li, Y., Li, D., Wang, Y., Zhu, Y., & Wu, W. (2021). A stochastic algorithm based on reverse sampling technique to fight against the cyberbullying. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *15*(4), 1-22. https://doi.org/10.1145/3441455

[16] Yin, C. J., Ayop, Z., Anawar, S., Othman, N.F., & Zainudin, N.M. (2021). Slangs and short forms of malay twitter sentiment analysis using supervised machine learning. *International Journal of Computer Science & Network Security*, *21*(11), 294-300. https://doi.org/10.22937/IJCSNS.2021.21.11.40

[17] Jacobs, G., Van Hee, C., & Hoste, V. (2022). Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text?. *Natural Language Engineering*, *28*(2), 141-166. https://doi.org/10.1017/S135132492000056X

[18] Kumari, K., Singh, J.P., Dwivedi, Y.K., & Rana, N.P. (2021). Multi-modal aggression identification using convolutional neural network and binary particle swarm optimization. *Future Generation Computer Systems*, *118*, 187-197. https://doi.org/10.1016/j.future.2021.01.014

[19] Abbas, A.M. (2021). Social network analysis using deep learning: applications and schemes. *Social Network Analysis and Mining*, *11*(1), 106. https://doi.org/10.1007/s13278-021-00799-z

[20] Toktarova, A., Beissenova, G., Kozhabekova, P., Makhanova, Z., Tulegenova, B., Rakhymbek, N., ... & Azhibekova, Z. (2021). Automatic offensive language detection in online user generated contents. *Journal of Theoretical and Applied Information Technology*, *99*(9), 2054-2067. https://www.elibrary.ru/item.asp?id=46818459