Hate Speech Detection in Social Networks using Machine Learning and Deep Learning Methods

Aigerim Toktarova[1], Dariga Syrlybay[2], Bayan Myrzakhmetova[3], Gulzat Anuarbekova[4], Gulbarshin Rakhimbayeva[5], Balkiya Zhylanbaeva[6], Nabat Suieuova[7], Mukhtar Kerimbekov[8]

Khoja Akhmet Yassawi International Kazakh, Turkish University, Turkistan, Kazakhstan[1]
Bachelor Student, Khoja Akhmet Yassawi International Kazakh, Turkish University, Turkistan, Kazakhstan[2]
M. Auezov South Kazakhstan Universtiy, Shymkent, Kazakhstan[1]
South Kazakhstan State Pedagogical University, Shymkent, Kazakhstan[3]
Abai Kazakh National Pedagogical University, Almaty, Kazakhstan[4]
Asfendiyarov Kazakh National Medical University, Almaty, Kazakhstan[5, 6]
Yessenov University, Aktau, Kazakhstan[7]
University of Friendship of People's Academician, A. Kuatbekov, Shymkent, Kazakhstan[8]

***Abstract*—Hate speech on social media platforms like Twitter is a growing concern that poses challenges to maintaining a healthy online environment and fostering constructive communication. Effective detection and monitoring of hate speech are crucial for mitigating its adverse impact on individuals and communities. In this paper, we propose a comprehensive approach for hate speech detection on Twitter using both traditional machine learning and deep learning techniques. Our research encompasses a thorough comparison of these techniques to determine their effectiveness in identifying hate speech on Twitter. We construct a robust dataset, gathered from diverse sources and annotated by experts, to ensure the reliability of our models. The dataset consists of tweets labeled as hate speech, offensive language, or neutral, providing a more nuanced representation of online discourse. We evaluate the performance of LSTM, BiLSTM, and CNN models against traditional shallow learning methods to establish a baseline for comparison. Our findings reveal that deep learning techniques outperform shallow learning methods, with BiLSTM emerging as the most accurate model for hate speech detection. The BiLSTM model demonstrates improved sensitivity to context, semantic nuances, and sequential patterns in tweets, making it adept at capturing the intricate nature of hate speech. Furthermore, we explore the integration of word embeddings, such as Word2Vec and GloVe, to enhance the performance of our models. The incorporation of these embeddings significantly improves the models' ability to discern between hate speech and other forms of online communication. This paper presents a comprehensive analysis of various machine learning methods for hate speech detection on Twitter, ultimately demonstrating the superiority of deep learning techniques, particularly BiLSTM, in addressing this critical issue. Our findings pave the way for further research into advanced methods of tackling hate speech and facilitating healthier online interactions.**

***Keywords*—Machine learning; deep learning; hate speech; social network; classification**

I. Introduction

Social media platforms like Twitter have become an essential communication tool in our digital age, enabling users worldwide to share their thoughts, opinions, and experiences with a vast audience [1]. However, the rapid growth of social media has also given rise to undesirable content, including hate speech. Hate speech is a form of communication that is offensive, malicious, and discriminatory, targeting individuals or groups based on their race, ethnicity, gender, religion, or other attributes [2]. The proliferation of hate speech on social media is a critical issue, as it fosters animosity, threatens social cohesion, and undermines the principles of free expression and respectful discourse. Consequently, the need for effective hate speech detection and monitoring tools is more significant than ever.

In recent years, machine learning techniques have emerged as a promising avenue for addressing the challenge of detecting and mitigating hate speech on social media platforms [3-4]. Machine learning algorithms, both shallow and deep, have demonstrated potential in tackling various natural language processing (NLP) tasks, such as sentiment analysis, text classification, and named entity recognition [5]. This paper aims to investigate and compare the performance of various shallow and deep learning methods in detecting hate speech on Twitter.

Machine learning methods have shown effectiveness in various applications, including spam detection, sentiment analysis, and topic modeling [6]. However, shallow learning algorithms have limitations in capturing the complex semantics and context of natural language, which may hinder their ability to identify hate speech accurately.

Deep learning techniques, on the other hand, have exhibited promising results in multiple NLP tasks due to their capacity for modeling high-level abstractions and capturing intricate language patterns [7]. LSTM and BiLSTM are recurrent neural networks (RNNs) that excel at processing sequential data, making them suitable for analyzing the temporal structure of text. CNNs, originally designed for image classification, have also demonstrated their applicability in text classification tasks by identifying local and global patterns in text through convolutional filters.

To investigate the effectiveness of shallow and deep learning methods for hate speech detection on Twitter, we first compile a diverse and representative dataset of tweets, ensuring

that the dataset encompasses a broad spectrum of online discourse [8]. The dataset is annotated by experts, who label the tweets as hate speech, offensive language, or neutral, thus providing a nuanced classification of the content. By adopting a multi-class labeling approach, we aim to capture the complexity and subtlety of hate speech more accurately.

We then apply a range of shallow learning techniques to the dataset, evaluating their performance in identifying hate speech. We also explore the integration of feature selection techniques. Establishing a baseline performance for these methods allows us to gauge the potential advantages of deep learning techniques.

Next, we implement LSTM, BiLSTM, and CNN models and evaluate their performance against the established baseline [9]. By comparing the performance of deep learning techniques with that of shallow learning methods, we aim to identify the most effective approach for hate speech detection on Twitter. In addition to comparing the overall performance of the models, we also assess their ability to handle various challenges associated with the detection of hate speech, such as understanding context, sarcasm, and semantic nuances.

To further enhance the performance of the deep learning models, we incorporate word embeddings, such as Word2Vec and GloVe, which facilitate the representation of words in a continuous vector space [10]. These embeddings capture semantic and syntactic relationships between words, thus enriching the input features for our models. By leveraging word embeddings, we aim to improve the models' ability to discern between hate speech and other forms of online communication, thereby increasing their accuracy and reducing false positives.

Our results reveal that deep learning techniques, particularly BiLSTM, outperform the shallow learning methods in detecting hate speech on Twitter [11]. BiLSTM demonstrates a superior ability to capture the intricate nature of hate speech by understanding context, semantic nuances, and sequential patterns in tweets [12]. This finding underscores the potential of deep learning techniques in addressing the challenge of hate speech detection and monitoring on social media platforms.

Thus, this paper presents a comprehensive analysis of various machine learning methods for hate speech detection on Twitter. Our findings suggest that deep learning techniques, specifically BiLSTM, hold promise for tackling this critical issue more effectively than their shallow learning counterparts. By identifying the most accurate models for hate speech detection, we contribute to the ongoing effort to develop advanced tools and strategies to combat hate speech on social media and foster healthier online interactions.

Future research directions may include the exploration of additional deep learning architectures, such as transformers, to further enhance hate speech detection performance. Moreover, investigating the impact of transfer learning and pre-trained language models, like BERT or GPT, on the performance of the models may provide valuable insights. Lastly, the development of explainable AI techniques to provide interpretable and transparent predictions in hate speech detection can improve user trust and facilitate better decision-making in content moderation.

II. Related Works

The problem of detecting hate speech on social media platforms has been extensively studied in recent years due to its increasing prevalence and the potential harm it can inflict on individuals and communities. In this section, we provide an overview of the related works in the field of hate speech detection, focusing on both shallow and deep learning approaches.

A. Shallow Learning Approaches

Several studies have utilized logistic regression, random forest, and decision tree algorithms for hate speech detection on social media. For instance, [13] employed logistic regression for hate speech detection in online communities, using bag-of-words and paragraph2vec features. Similarly, [14] proposed a multi-class classifier using logistic regression and random forest, which demonstrated improved performance over single classifiers. Study [15] employed decision trees to detect hate speech on Twitter, highlighting the importance of feature engineering in improving model performance.

Naïve Bayes and K-NN classifiers have also been employed for hate speech detection. Like [16] used a naïve bayes classifier to detect cyber hate on Twitter, while [17] proposed a K-NN-based approach for the same task. Both studies indicated the effectiveness of these classifiers in detecting hate speech when combined with appropriate feature extraction techniques, such as bag-of-words and TF-IDF.

SVMs have been widely used for hate speech detection, with several studies demonstrating their effectiveness. For example, [18] used SVM to detect cyberbullying and hate speech on Twitter, leveraging features such as character n-grams, sentiment scores, and syntactic patterns. Similarly, [19] employed SVM for hate speech detection, demonstrating that the inclusion of linguistic and semantic features improved the model's performance.

B. Deep Learning Approaches

LSTM and BiLSTM models have been increasingly employed for hate speech detection due to their ability to capture long-range dependencies in text. The authors in [20] proposed a deep learning approach using LSTM for detecting hate speech on Twitter, demonstrating superior performance compared to shallow learning techniques. On the other hand, [21] used BiLSTM models for the same task, illustrating the effectiveness of bidirectional RNNs in capturing the context and semantics of text. Additionally, [22] used both LSTM and BiLSTM models to detect hate speech on Twitter, finding that the BiLSTM model outperformed its unidirectional counterpart.

CNNs have also been applied for hate speech detection on social media platforms. As [23] proposed a CNN-based model for detecting hate speech on Twitter, leveraging character n-grams as input features. Their approach demonstrated improved performance compared to traditional shallow learning techniques. Similarly, [24] employed a CNN-based model for hate speech detection on Twitter, illustrating the

benefits of incorporating pre-trained word embeddings such as Word2Vec and GloVe.

C. Hybrid Approaches and Ensemble Models

Some studies have explored hybrid approaches and ensemble models for hate speech detection, combining both shallow and deep learning techniques to enhance model performance. The research [25] proposed a hybrid approach that combined CNN with LSTM for detecting abusive language on Twitter, demonstrating that the integrated model outperformed standalone CNN and LSTM models. Similarly, [26] developed an ensemble model combining SVM and LSTM for hate speech detection, which achieved better performance compared to individual models.

Recent studies have highlighted the importance of using word embeddings and pre-trained language models for improving hate speech detection performance. For instance, [27] investigated the impact of using different word embeddings. Their findings revealed that the choice of word embeddings could significantly impact model performance.

The use of pre-trained language models has also been explored for hate speech detection. Like [28] proposed a BERT-based model for detecting hate speech on social media platforms, demonstrating superior performance compared to traditional machine learning techniques. Similarly, [29] employed BERT for detecting hate speech on Twitter, highlighting the model's ability to capture the complex semantics of text and adapt to various linguistic contexts.

III. Problem Statement

It is possible that the problem of early identification of cyberbullying on social networking sites is separate from the difficulty of classifying different types of cyberbullying. In the circumstances presented here, there is a group of social media sessions that we will refer to collectively as "S." As a result, there is the chance that some of them are instances of cyberbullying. A sequence of sessions on a social network may be described using the equation (1), which is as follows:

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

Where S refers to the total number of sessions, "i" indicates the current session.

The sequence in which submissions are made during a specific session is subject to change at different points in time and is governed by a variety of factors

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

where the tuple P represents the kth post for the social network session and s is the timestamp of when post P was published.

At the same time, a vector of features is utilized to identify each post in a manner that is completely unique:

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

Therefore, the objective is to acquire the knowledge necessary to develop a function f that can classify whether or not a text is related to hate speech.

IV. Materials and Methods

A. The Proposed Framework

A representation of the model that has been built for the purpose of identifying instances of cyberbullying may be shown in Fig. 1. The following are the steps that make up this model: the preprocessing stage, the feature extraction stage, the classification stage, and the assessment stage. In this part, a significant amount of focus is placed on doing a more in-depth analysis of each stage.

B. Feature Extraction

Term frequency-inverse document frequency: In the context of the paper on hate speech detection using shallow and deep learning methods, Term Frequency-Inverse Document Frequency (TF-IDF) plays an essential role as a feature extraction technique [30].

The product of TF and IDF yields the TF-IDF score, which reflects the importance of a term within a document and across the entire corpus. A high TF-IDF score suggests that the term is significant within the document and infrequent across the corpus, making it a valuable feature for classification tasks [31].

In the context of hate speech detection, TF-IDF can be employed to transform raw text data into a structured representation that captures the relative importance of words or terms [32]. The resulting feature vectors can be used as input for various shallow learning algorithms, to develop hate speech detection models. By incorporating TF-IDF, the models can effectively distinguish between hate speech and other types of communication based on the discriminative power of specific terms.

It is important to note that, while TF-IDF has been proven effective in various text classification tasks, it may not always capture the complex semantics and context inherent in natural language [33]. In such cases, advanced feature extraction techniques, such as word embeddings or pre-trained language models, may be employed to complement or replace TF-IDF in the development of more sophisticated hate speech detection models.

Word2Vec: In the context of the paper on hate speech detection using shallow and deep learning methods, Word2Vec is a significant technique for generating word embeddings. Word2Vec is an unsupervised learning algorithm that converts words into continuous vector representations, capturing semantic and syntactic relationships between words. The technique was introduced by [34] and has since become a widely used method in natural language processing (NLP) tasks, including text classification, sentiment analysis, and machine translation.

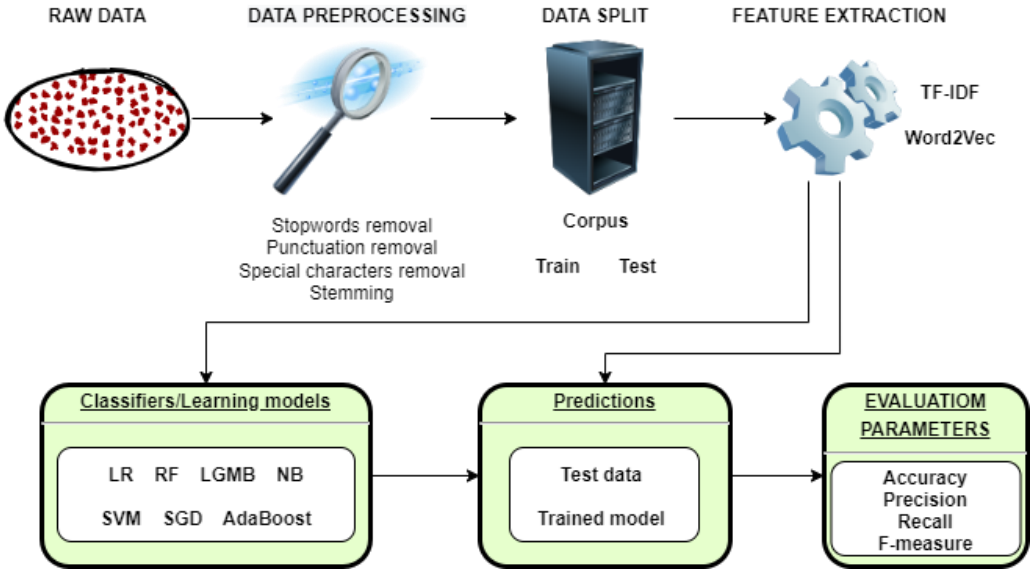

Fig. 1. Proposed framework.

In hate speech detection, Word2Vec embeddings can be employed to enrich the input features for both shallow and deep learning models [35]. By leveraging the semantic information captured in these embeddings, models can better discern between hate speech and other types of communication, resulting in improved classification performance. Word2Vec embeddings can be used in combination with other feature extraction techniques, such as TF-IDF or pre-trained language models, to further enhance the models' understanding of the complex semantics and context inherent in natural language.

In this specific piece of study, the weighting method that we make use of is the tf-idf system. For the purpose of calculating the tf-idf weight that corresponds to the ith word in the jth text, the following formula is used:

$$w_{i,j} = TF_{i,j} \times \log\left(\frac{N}{DF_i}\right) \quad (4)$$

Bag of Words: In the context of the paper on hate speech detection using shallow and deep learning methods, the Bag of Words (BoW) model serves as a fundamental text representation technique [36]. BoW is a widely used method in natural language processing (NLP) tasks, such as text classification, information retrieval, and sentiment analysis, as it provides a simple and efficient way to represent text data in a structured format.

In hate speech detection, BoW can be employed to transform raw text data into a structured representation that serves as input for various shallow learning algorithms. However, it is important to note that the BoW model lacks the ability to capture context, semantics, and word order, which may limit its effectiveness in some classification tasks. To address these limitations, more advanced feature extraction techniques, such as word embeddings (e.g., Word2Vec) or pre-trained language models, can be used in combination with or as a replacement for the BoW model. The goal is to increase the likelihood that, given the following circumstances:

$$\arg\max_{\theta} \prod_{w \in T} \left[\prod_{c \in C} p(c \mid w; \theta) \right] \quad (5)$$

C. Machine Learning Methods

Decision Tree is a supervised learning algorithm that recursively splits the input space into regions based on feature values, forming a tree-like structure [37]. It is interpretable and handles non-linear relationships well. In hate speech detection, Decision Trees can be employed to make decisions based on extracted text features, such as word frequencies or presence of specific terms.

Naïve Bayes is a probabilistic classifier based on Bayes' theorem, which assumes feature independence [38]. Despite this simplifying assumption, it often performs well in text classification tasks. In hate speech detection, Naïve Bayes can be used to classify tweets by estimating the likelihood of a tweet being hate speech given the occurrence of certain words or phrases.

K-Nearest Neighbors is a non-parametric, instance-based learning algorithm that classifies instances based on the majority class of their K-nearest neighbors in the feature space [39]. In hate speech detection, K-NN can be employed to classify tweets by considering the similarity between their feature representations, such as word embeddings or TF-IDF vectors.

Support Vector Machine (SVM) is a supervised learning algorithm that aims to find the optimal hyperplane separating different classes in the feature space [40]. It is effective in handling high-dimensional data and can be used with various kernel functions. In the context of hate speech detection, SVM can be employed to classify tweets by learning the decision boundary based on the extracted features, such as word frequencies, n-grams, or sentiment scores.

D. Deep Learning Methods

1) LSTM (Long Short-Term Memory): LSTM is a type of recurrent neural network (RNN) specifically designed to address the vanishing gradient problem common in standard RNNs [41]. LSTM networks have memory cells that can store information over long sequences, allowing them to capture long-range dependencies and context within text data. In the context of hate speech detection, LSTM models can be employed to process tweets as sequences of words or characters, enabling them to capture temporal patterns and dependencies that are crucial for understanding the semantics and intent of the text. By using LSTM networks, classification models can better distinguish between hate speech and non-hate speech based on the contextual information present in the tweets. Fig. 2 demonstrates architecture of LSTM network.

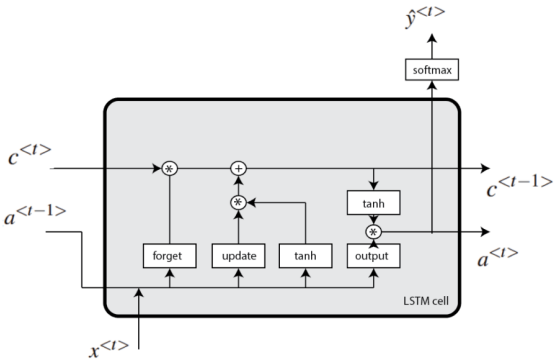

Fig. 2. LSTM network.

2) BiLSTM (Bidirectional Long Short-Term Memory): BiLSTM is an extension of LSTM that processes the input data in both forward and backward directions, enabling it to capture both past and future context [42]. In the context of hate speech detection, BiLSTM models can process tweets in a bidirectional manner, capturing the context and dependencies present in the text more effectively. This improved contextual understanding leads to better classification performance compared to unidirectional LSTM models. BiLSTM networks can be combined with other deep learning architectures, such as convolutional neural networks (CNN), to further enhance the model's ability to capture both local and global contextual information in the text. Fig. 3 demonstrates architecture of BiLSTM network.

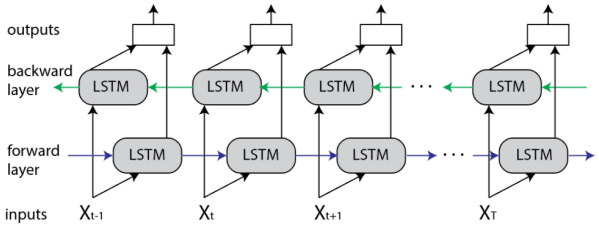

Fig. 3. BiLSTM network.

3) CNN (Convolutional Neural Network): CNN is a deep learning architecture traditionally used for image processing tasks but has also demonstrated effectiveness in various NLP tasks, including text classification [43]. CNNs employ convolutional layers to learn local patterns within input data through the application of filters or kernels. In the context of hate speech detection, CNN models can be used to process tweets by treating them as one-dimensional sequences of words or characters. These models can learn local patterns, such as n-grams or specific phrases that are indicative of hate speech. By combining CNNs with other deep learning architectures, such as LSTM or BiLSTM, models can capture both local patterns and long-range dependencies, leading to improved classification performance in hate speech detection tasks. Fig. 4 demonstrates architecture of the convolutional neural network.

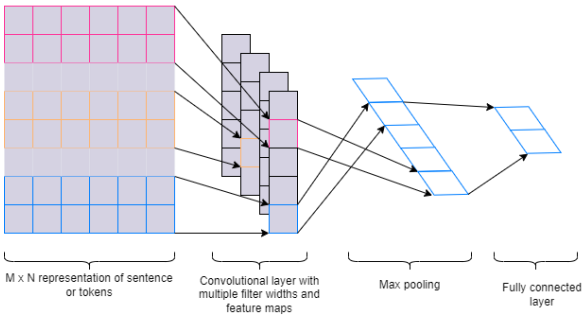

Fig. 4. CNN architecture.

V. Experimental Setup

A. Evaluation Parameters

1) Accuracy: Accuracy is a common metric used to evaluate the performance of classification models. It is calculated as the ratio of the number of correct predictions to the total number of predictions [44]. Although accuracy provides a general overview of a model's performance, it may not be suitable for imbalanced datasets, where one class dominates the other(s), as it can yield misleading results.

$$accuracy = \frac{TP + TN}{P + N} \quad (6)$$

2) Precision: Precision is a metric that evaluates the proportion of true positive predictions among all positive predictions made by a classification model [45]. It is particularly useful for assessing the performance of models when the cost of false positives is high, such as in spam detection or medical diagnosis.

$$preision = \frac{TP}{TP + FP} \quad (7)$$

3) Recall: Recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions among all actual positive instances in the dataset [46]. Recall is crucial in situations where the cost of false negatives is high, such as in fraud detection or cancer diagnosis.

$$recall = \frac{TP}{TP + FN} \quad (8)$$

4) F-score: F-score, or F1-score, is the harmonic mean of precision and recall, and provides a balanced measure of a model's performance when both false positives and false negatives are important. The F-score ranges from 0 to 1, where a higher value indicates better performance [47]. It is particularly useful for evaluating models on imbalanced datasets, where accuracy may be misleading.

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (9)$$

5) ROC curve: The Receiver Operating Characteristic (ROC) curve is a graphical representation of a classifier's performance, plotting the true positive rate (recall) against the false positive rate for various decision thresholds. The area under the ROC curve (AUC-ROC) is a scalar measure of a model's performance, with a higher value (closer to 1) indicating better classification [48]. The ROC curve and AUC-ROC are especially useful for comparing different models and selecting the optimal decision threshold.

B. Experimental Results

Accuracy, Precision, Recall, F-measure, and Area under a Receiver Operating Characteristic (AUC-ROC) are all terms that are used in the field of cyberbullying detection research. The confusion matrices for each of the techniques used in this work and evaluated in the cyberbullying classification dataset are shown in Fig. 5. We are able to clearly show the actual amount of classification results in respect to other classes by using confusion matrices. In the research that we conducted, we found that there are three different classes: cyberbullying, which was given the score of 1, non-cyberbullying, which was given the score of 0, and neutral class, which was given the score of 2.

In Fig. 6, a comparison is made between the model that was suggested and all of the other machine learning and deep learning models that were used. The AUC performance evaluation in each classification is done by finding the area under the receiver operating characteristic curve that includes all extracted attributes. The AUC-ROC curves of all of the strategies that have been implemented as well as the suggested method are compared in Fig. 7. As has been pointed out, deep learning models have been shown to be more valuable than machine learning approaches. According to the figure, the suggested model, which consists of BiLSTM, displays the best AUC-ROC value from the very first iteration and all the way along the graph.

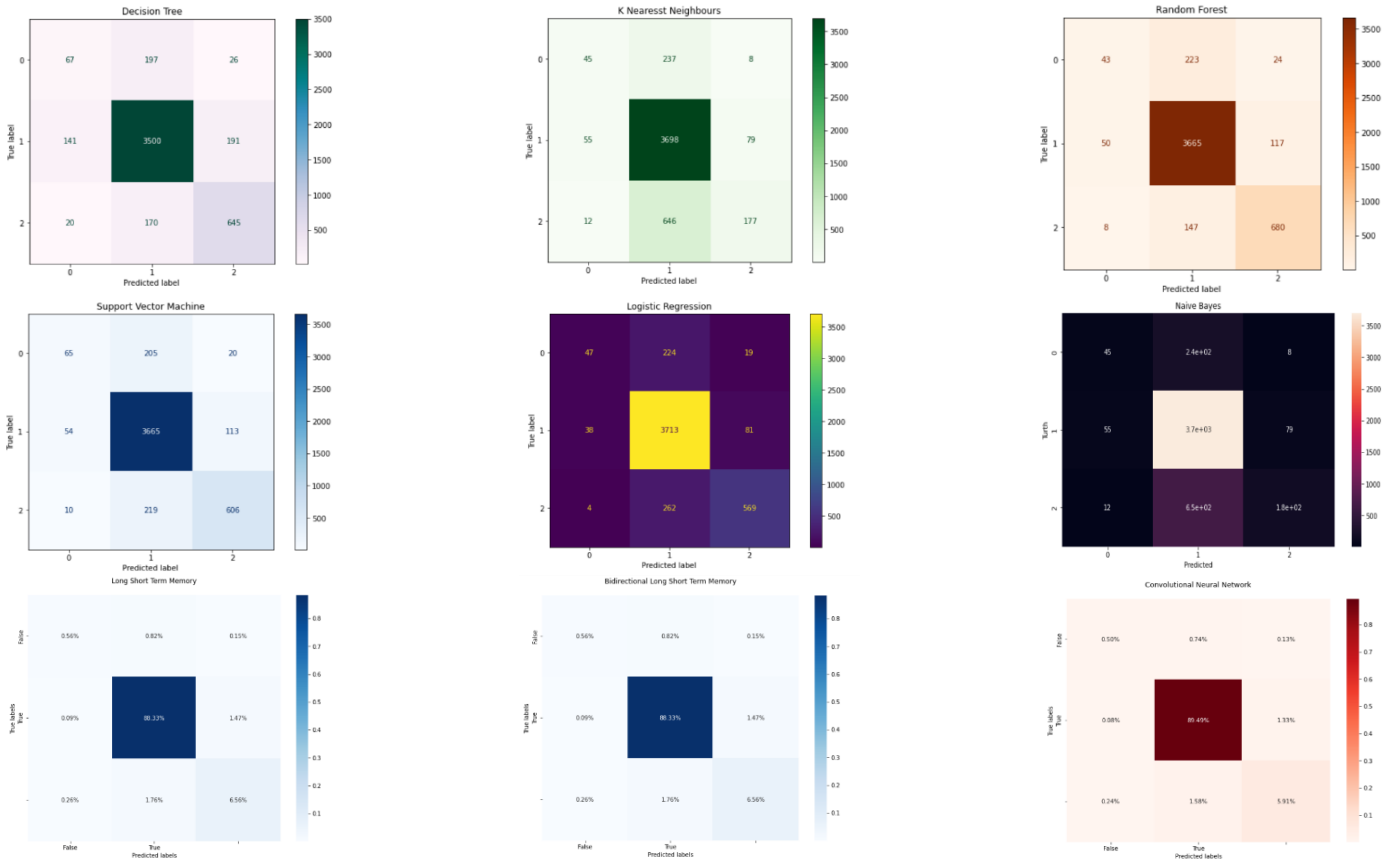

Fig. 5. Confusion matrices for hate speech detection.

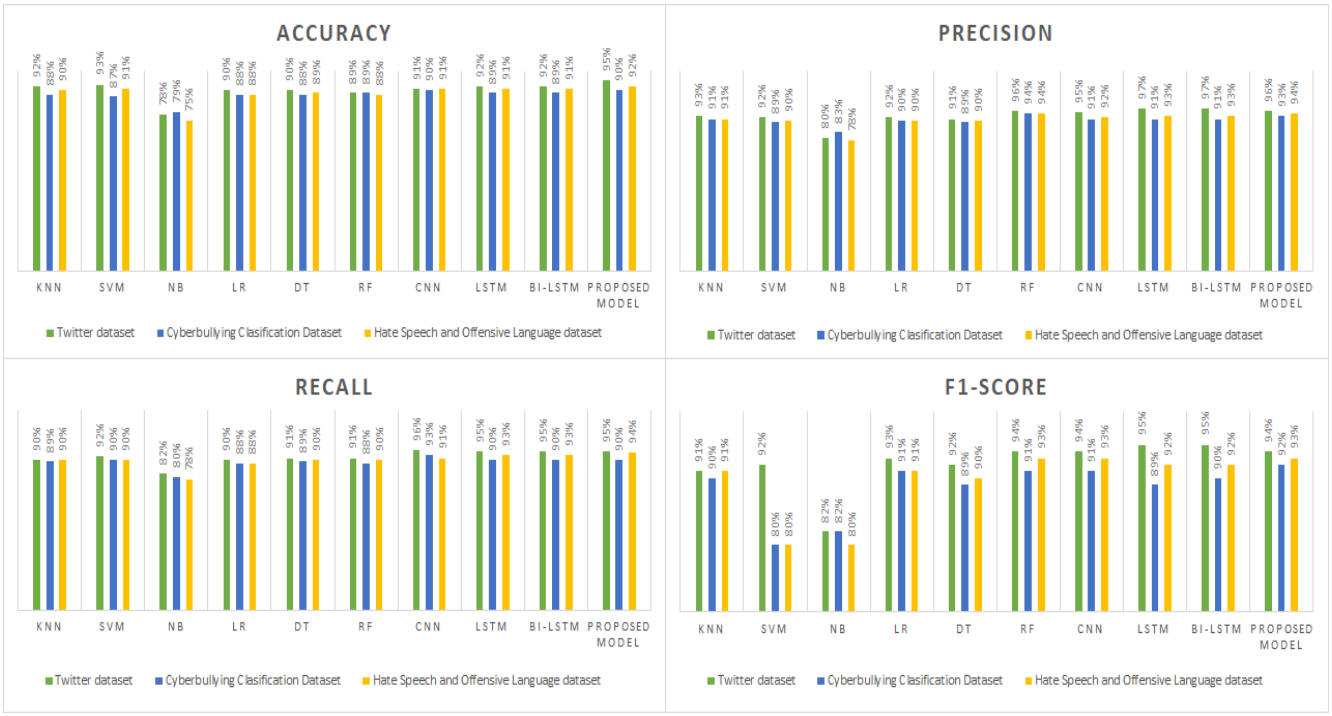

Fig. 6. Evaluation parameters for different datasets.

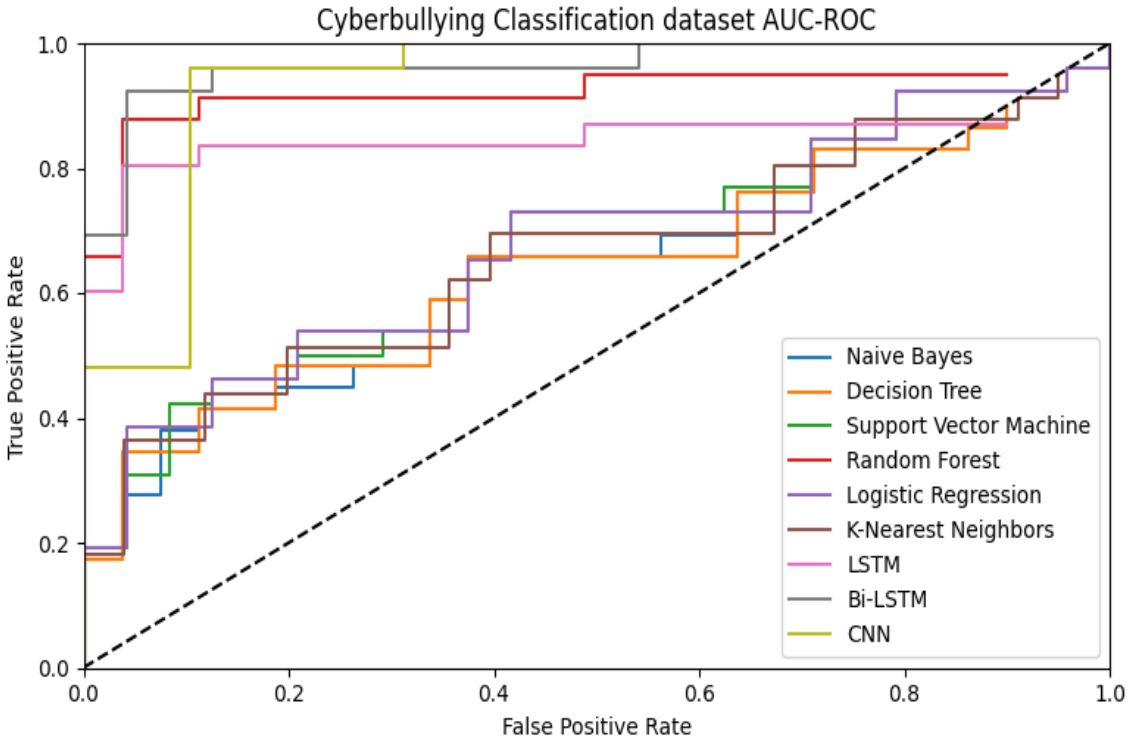

Fig. 7. ROC curve of applied machine learning and deep learrning techniques for hate speech detection.

The categorization findings of cyberbullying are shown in Table I below. These results were achieved by using machine learning and deep learning techniques to three different datasets. We employed assessment measures such as accuracy, precision, recall, and F1-score [48-51] to evaluate the approaches of machine learning and deep learning.

TABLE I. COMPARISON OF THE OBTAINED RESULTS

Dataset	Approach	Model	Accuracy	Precision	Recall	F-score	ROC
Hate Speech and Offensive Language	Machine Learning Models	SVM	0.873	0.852	0.862	0.851	0.78
		KNN	0.856	0.839	0.831	0.837	0.92
		NB	0.874	0.832	0.863	0.851	0.80
		DT	0.602	0.524	0.585	0.642	0.65
		RF	0.851	0.854	0.822	0.856	0.77
		LR	0.862	0.853	0.837	0.858	0.78
	Deep Learning Models	CNN	0.892	0.895	0.898	0.896	0.93
		LSTM	0.901	0.896	0.91	0.898	0.93
		BiLSTM	0.902	0.916	0.904	0.899	0.94
Twitter Hate Speech	Machine Learning Models	SVM	0.873	0.852	0.862	0.851	0.75
		KNN	0.856	0.839	0.831	0.837	0.90
		NB	0.874	0.832	0.863	0.851	0.76
		DT	0.602	0.524	0.585	0.642	0.68
		RF	0.851	0.854	0.822	0.856	0.77
		LR	0.862	0.853	0.837	0.858	0.78
	Deep Learning Models	CNN	0.892	0.895	0.898	0.896	0.92
		LSTM	0.901	0.896	0.91	0.898	0.92
		BiLSTM	0.902	0.916	0.904	0.899	0.93
Cyberbullying	Machine Learning Models	SVM	0.873	0.852	0.862	0.851	0.75
		KNN	0.856	0.839	0.831	0.837	0.80
		NB	0.874	0.832	0.863	0.851	0.79
		DT	0.602	0.524	0.585	0.642	0.67
		RF	0.851	0.854	0.822	0.856	0.78
		LR	0.862	0.853	0.837	0.858	0.78
	Deep Learning Models	CNN	0.892	0.895	0.898	0.896	0.91
		LSTM	0.901	0.896	0.92	0.898	0.91
		BiLSTM	0.902	0.916	0.904	0.899	0.93

As a consequence of this, taking into consideration the success rates it has achieved, the suggested strategy may be accepted as a possible method for identifying instances of cyberbullying inside social networking sites. In addition, taking into account all of the criteria used for assessment, the deep neural network that was presented had the greatest performance when it comes to identifying cases of cyberbullying. The usage of the suggested deep neural network for modifying the weights and biases, in addition to a reduction in the amount of time spent training, resulted in favorable outcomes, which can be ascribed to the employment of the proposed technique. The results indicate that the suggested technique using deep neural networks may easily be modified to handle both short and lengthy texts as they are currently used.

VI. DISCUSSION

In this section, we will discuss the advantages, disadvantages, open issues, challenges, and future perspectives of the methods explored in this paper for hate speech detection in Twitter.

A. Advantages of Computational Intelligence in Hate Speech Detection

Shallow learning methods, such as logistic regression, random forest, decision tree, naïve bayes, K-NN, and SVM, offer several benefits, including simplicity, interpretability, and computational efficiency. These algorithms can perform well on relatively small datasets and are less prone to overfitting compared to deep learning methods.

Deep learning methods, such as LSTM, BiLSTM, and CNN, have the ability to capture complex patterns and long-range dependencies in text data. These methods can learn hierarchical representations of the data, leading to improved classification performance in many NLP tasks, including hate speech detection.

Feature extraction techniques, such as Bag of Words, TF-IDF, and Word2Vec, allow for the transformation of raw text data into structured representations suitable for input to various classifiers. These techniques can capture different aspects of text data, such as word frequencies, term importance, and semantic relationships, providing valuable information for classification tasks.

B. Disadvantages of Computational Intelligence in Hate Speech Detection

Shallow learning methods may struggle to capture complex patterns and long-range dependencies in text data, which can lead to suboptimal classification performance in some cases.

Deep learning methods, despite their ability to capture complex patterns, may suffer from overfitting and require large amounts of labeled data for effective training. Additionally,

these models can be computationally expensive and less interpretable than shallow learning methods.

Feature extraction techniques, while providing valuable information for classification tasks, may not always capture the nuanced semantics and context present in natural language. This limitation can lead to misclassifications, particularly in complex tasks such as hate speech detection.

C. Open Issues and Challenges of Computational Intelligence in Hate Speech Detection

The development of robust and accurate classifiers for hate speech detection remains an open issue, as the nature of hate speech is constantly evolving. New forms of hate speech, including code words, slang, or non-textual elements (e.g., images or emojis), may not be effectively captured by existing models and feature extraction techniques.

The presence of imbalanced datasets, where the number of instances of one class significantly outweighs the other(s), is a common challenge in hate speech detection. Traditional performance metrics, such as accuracy, may be misleading in these situations, and alternative metrics or approaches (e.g., F-score, oversampling, or undersampling) may be necessary for effective model evaluation.

The issue of false positives and false negatives in hate speech detection presents a significant challenge, as the consequences of these misclassifications can be severe, leading to the suppression of free speech or the perpetuation of harmful content. Developing models that strike a balance between precision and recall remains a critical task.

D. Future Perspectives of Computational Intelligence in Hate Speech Detection

Investigating the integration of other deep learning architectures, such as transformers or attention mechanisms, may further enhance the models' ability to capture complex semantics and context, leading to improved classification performance.

The use of pre-trained language models, such as BERT or GPT, can be explored for their potential to leverage large-scale, pre-existing knowledge of language structure and semantics, leading to more accurate and robust hate speech detection systems.

Developing methods for effectively handling imbalanced datasets, such as advanced sampling techniques, cost-sensitive learning, or ensemble methods, may lead to improved model performance and more accurate classification of hate speech.

Exploring techniques for incorporating non-textual elements, such as images or emojis, into hate speech detection models can help address the evolving nature of hate speech and improve the overall effectiveness of classification systems.

Investigating methods for enhancing the interpretability of deep learning models, such as attention mechanisms or explainable AI techniques, can provide valuable insights into the decision-making process of these models, improving trust and adoption in real-world applications.

Collaborating with domain experts, such as sociologists or psychologists, can help in developing a more comprehensive understanding of the complex and evolving nature of hate speech. This interdisciplinary approach can lead to the development of more effective and contextually-aware classification models.

Exploring the potential of transfer learning and domain adaptation techniques can help in developing models that can be effectively applied to different languages, regions, or platforms, broadening the impact and applicability of hate speech detection systems.

In conclusion, the methods and techniques presented in this paper provide a foundation for the development of advanced, robust, and accurate hate speech detection systems. By addressing the open issues and challenges, and considering future perspectives, researchers can contribute to the ongoing effort to create a safer and more inclusive online environment on platforms like Twitter. The lessons learned from these investigations can also be applied to other social media platforms and domains, where hate speech and harmful content pose significant challenges to users and society at large.

VII. Conclusion

In conclusion, this paper has presented a comprehensive study of various shallow and deep learning methods for detecting hate speech on Twitter. Shallow learning algorithms, including logistic regression, random forest, decision tree, naïve bayes, K-NN, and SVM, have been explored as effective classifiers for identifying hate speech based on features extracted from text data, such as Bag of Words, TF-IDF, or word embeddings. Additionally, deep learning methods, such as LSTM, BiLSTM, and CNN, have been investigated for their ability to capture complex patterns and long-range dependencies in text, resulting in improved classification performance.

The paper has also discussed the importance of feature extraction techniques in transforming raw text data into structured representations that can be used as input for various classifiers. Techniques like Bag of Words, TF-IDF, and Word2Vec have been highlighted for their ability to capture different aspects of text data, including word frequencies, term importance, and semantic relationships.

In evaluating the performance of the various classifiers, metrics such as accuracy, precision, recall, F-score, and ROC curve have been employed to provide a comprehensive understanding of the models' effectiveness in detecting hate speech. These metrics are crucial in assessing the trade-offs between different models and selecting the most suitable approach for a particular task.

Future research in hate speech detection can explore the integration of other deep learning architectures, such as transformers or attention mechanisms, to further enhance the models' ability to capture complex semantics and context. Moreover, the use of pre-trained language models, such as BERT or GPT, can be investigated for their potential to improve classification performance by leveraging large-scale, pre-existing knowledge of language structure and semantics.

Ultimately, the detection of hate speech on social media platforms like Twitter is of paramount importance in fostering a safe and inclusive online environment. The methods and techniques presented in this paper provide valuable insights and serve as a foundation for the development of advanced, robust, and accurate hate speech detection systems.

REFERENCES

[1] T. Alsubait and D. Alfageh, "Comparison of machine learning techniques for cyberbullying detection on youtube arabic comments," International Journal of Computer Science and Network Security, vol. 21, no. 1, pp. 1–5, 2021.

[2] A. Dewani, M. Memon and S Bhatti, "Cyberbullying detection: Advanced preprocessing techniques & deep learning architecture for roman urdu data," Journal of Big Data, vol. 8, no. 1, pp. 1–20, 2021.

[3] D. Hall, Y. Silva, Y. Wheeler, L. Cheng and K. Baumel, "Harnessing the power of interdisciplinary research with psychology-informed cyberbullying detection models," International Journal of Bullying Prevention, vol. 4, no.1, pp. 47–54, 2021.

[4] K. Arce-Ruelas, "Automatic cyberbullying detection: A Mexican case in high school and Higher Education Students," IEEE Latin America Transactions, vol. 20, no. 5, pp. 770–779, 2022.

[5] T. Ahmed, M. Rahman, S. Nur, A. Islam and D. Das, "Natural language processing and machine learning based cyberbullying detection for Bangla and romanized bangla texts," TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 20, no. 1 pp. 89–97, 2021.

[6] B. Omarov, A. Tursynova, O. Postolache, K. Gamry, A. Batyrbekov et al., "Modified UNet model for brain stroke lesion segmentation on computed tomography images," CMC-Computers, Materials & Continua, vol. 71, no. 3, pp. 4701–4717, 2022.

[7] A. Al-Marghilani, "Artificial intelligence-enabled cyberbullying-free online social networks in smart cities," International Journal of Computational Intelligence Systems, vol. 15, no. 1, pp. 1–13, 2022.

[8] C. Theng, N. Othman, R. Abdullah, S. Anawar, Z. Ayop et al., "Cyberbullying detection in twitter using sentiment analysis," International Journal of Computer Science & Network Security, vol. 21, no. 11, pp. 1-10, 2021.

[9] S. Sadiq, A. Mehmood, S. Ullah, M. Ahmad, G. Choi et al., "Aggression detection through deep neural model on twitter," Future Generation Computer Systems, vol. 114, no. 1, pp. 120–129, 2021.

[10] E. Sarac Essiz and M. Oturakci, "Artificial bee colony–based feature selection algorithm for cyberbullying," The Computer Journal, vol. 64, no. 3, pp. 305–313, 2021.

[11] C. E. Gomez, M. O. Sztainberg and R. E. Trana, "Curating cyberbullying datasets: a human-AI collaborative approach," International journal of bullying prevention, vol. 4, no. 1, pp. 35-46, 2022.

[12] S. Salawu, J. Lumsden and Y. He, "A mobile-based system for preventing online abuse and cyberbullying," International Journal of Bullying Prevention, vol. 4, no. 1, pp. 66–88, 2022.

[13] M. Mladenović, V. Ošmjanski and S. V. Stanković, "Cyber-aggression, cyberbullying, and cyber-grooming: a survey and research challenges," ACM Computing Surveys (CSUR), vol. 54, no.1, pp. 1–42, 2021.

[14] S. R. Sangwan and M. P. S. Bhatia, "Denigrate comment detection in low-resource Hindi language using attention-based residual networks," Transactions on Asian and Low-Resource Language Information Processing, vol. 21, no. 1, pp. 1–14, 2021.

[15] T. T. Aurpa, R. Sadik and M. S. Ahmed, "Abusive Bangla comments detection on Facebook using transformer-based deep learning models," Social Network Analysis and Mining, vol. 12, no.1, pp. 1–14, 2022.

[16] R. Yan, Y. Li, D. Li, Y. Wang, Y. Zhu et al., "A Stochastic Algorithm Based on Reverse Sampling Technique to Fight Against the Cyberbullying," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 15, no. 4, pp. 1–22, 2021.

[17] C. J. Yin, Z. Ayop, S. Anawar, N. F. Othman and N. M. Zainudin, "Slangs and Short forms of Malay Twitter Sentiment Analysis using Supervised Machine Learning," International Journal of Computer Science & Network Security, vol. 21, no. 11, pp. 294–300, 2021.

[18] G. Jacobs, C. Van Hee and V. Hoste, "Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text?," Natural Language Engineering, vol. 28, no. 2, pp. 141–166, 2022.

[19] A. Jevremovic, M. Veinovic, M. Cabarkapa, M. Krstic, I. Chorbev et al., "Keeping Children Safe Online With Limited Resources: Analyzing What is Seen and Heard," IEEE Access, vol. 9, no. 1, pp. 132723–132732, 2021.

[20] K. Kumari, J. P. Singh, Y. K. Dwivedi and N. P. Rana, "Multi-modal aggression identification using convolutional neural network and binary particle swarm optimization," Future Generation Computer Systems, vol. 118, no. 1, pp. 187–197, 2021.

[21] A. M. Abbas, "Social network analysis using deep learning: applications and schemes," Social Network Analysis and Mining, vol.11, no. 1, pp. 1–21, 2021.

[22] S. Gupta, N. Mohan, P. Nayak, K. C. Nagaraju and M. Karanam, "Deep vision-based surveillance system to prevent train–elephant collisions," Soft Computing, vol. 26, no. 8, pp. 4005–4018, 2022.

[23] S. Mohammed, W. C. Fang, A. E. Hassanien and T. H. Kim, "Advanced Data Mining Tools and Methods for Social Computing," The Computer Journal, vol. 64, no. 3, pp. 281–285, 2021.

[24] B. Thuraisingham, "Trustworthy Machine Learning," IEEE Intelligent Systems, vol. 37, no.1, pp. 21–24, 2022.

[25] V. Rupapara, F. Rustam, H. Shahzad, A. Mehmood, I. Ashraf et al., "Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model," IEEE Access, vol. 9, no. 1, pp. 78621–78634, 2021.

[26] O. Sharif and M. M. Hoque, "Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers," Neurocomputing, vol. 490, no. 1, pp. 462–481, 2022.

[27] K. Kumari, J. P. Singh, Y. K. Dwivedi and N. P. Rana, "Bilingual Cyber-aggression detection on social media using LSTM autoencoder," Soft Computing, vol. 25, no. 14, pp. 8999–9012, 2021.

[28] A. Mohamed, E. Amer, N. Eldin, M. Hossam, N. Elmasry et al., "The Impact of Data processing and Ensemble on Breast Cancer Detection Using Deep Learning,". Journal of Computing and Communication, vol. 1, no.1, pp. 27–37, 2022.

[29] A. Sheth, V. L. Shalin and U. Kursuncu, "Defining and detecting toxicity on social media: context and knowledge are key," Neurocomputing, vol. 490, no. 1, pp. 312–318, 2022.

[30] U. Kursuncu, H. Purohit, N. Agarwal and A. Sheth, "When the bad is good and the good is bad: Understanding cyber social health through online behavioral change," IEEE Internet Computing, vol. 25, no.1, pp. 6–11, 2021.

[31] A. M. Veiga Simão, P. Costa Ferreira, N. Pereira, S. Oliveira, P. Paulino et al., "Prosociality in cyberspace: Developing emotion and behavioral regulation to decrease aggressive communication," Cognitive Computation, vol. 13, no. 3, pp. 736–750, 2021.

[32] G. Isaza, F. Muñoz, L. Castillo and F. Buitrago, "Classifying cybergrooming for child online protection using hybrid machine learning model," Neurocomputing, vol. 484, no. 1, pp. 250–259, 2022.

[33] L. Cuoghi and L. Konopelko, "Cyberbullying Classification," [Online]. Available https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification (accessed on 25 June 2022). 2022.

[34] D. Bruwaene, Q. Huang and D. Inkpen, "A multi-platform dataset for detecting cyberbullying in social media," [Online]. Available https://dl.acm.org/doi/abs/10.1007/s10579-020-09488-3 (accessed on 25 June 2022). 2022.

[35] A. Samoshyn, "Hate Speech and Offensive Language Dataset," [Online]. Available https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset (accessed on 25 June 2022). 2020.

[36] G. Perasso, N. Carone and L. Barone. "Written and visual cyberbullying victimization in adolescence: Shared and unique associated factors,"

European Journal of Developmental Psychology, vol. 18, no. 5, pp. 658–677, 2021.

[37] M. Amjad, N. Ashraf, A. Zhila, G. Sidorov, A. Zubiaga et al., "Threatening Language Detection and Target Identification in Urdu Tweets," IEEE Access, vol. 9, no. 1, pp. 128302–128313, 2021.

[38] Ö. Çoban, S. A. Özel and A. İnan, "Deep Learning-based Sentiment Analysis of Facebook Data: The Case of Turkish Users," The Computer Journal, vol. 64, no. 3, pp. 473–499, 2021.

[39] B. Omarov, N. Saparkhojayev, S. Shekerbekova, O. Akhmetova, M. Sakypbekova et al., "Artificial intelligence in medicine: Real time electronic stethoscope for heart diseases detection," CMC-Computers, Materials & Continua, vol. 70, no. 2, pp. 2815–2833, 2022.

[40] P. Parikh, H. Abburi, N. Chhaya, M. Gupta and V. Varma, "Categorizing Sexism and Misogyny through Neural Approaches," ACM Transactions on the Web (TWEB), vol. 15, no.4, pp. 1–31, 2021.

[41] S. Kiritchenko, I. Nejadgholi and K. C. Fraser, "Confronting abusive language online: A survey from the ethical and human rights perspective," Journal of Artificial Intelligence Research, vol. 71, no. 1, pp. 431–478, 2021.

[42] J. A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios and R. Valencia-García, "Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings," Future Generation Computer Systems, vol. 114, no. 1, pp. 506–518, 2021.

[43] A. Tontodimamma, E. Nissi, A. Sarra and L. Fontanella, "Thirty years of research into hate speech: topics of interest and their evolution," Scientometrics, vol. 126, no.1, pp. 157–179, 2021.

[44] X. Chen, H. Xie, G. Cheng and Z. Li, "A decade of sentic computing: topic modeling and bibliometric analysis," Cognitive Computation, vol. 14, no.1, pp. 24–47, 2022.

[45] A. S. Srinath, , H. Johnson, G. G. Dagher and M. Long, "BullyNet: Unmasking Cyberbullies on Social Networks," IEEE Transactions on Computational Social Systems, vol. 8, no.2, pp. 332–344, 2021.

[46] C. Kumar, T. S. Bharati and S. Prakash, "Online social network security: A comparative review using machine learning and deep learning," Neural Processing Letters, vol. 53, no. 1, pp. 843–861, 2021.

[47] M. Zhu, A. H. Anwar, Z. Wan, J. H. Cho, C. A. Kamhoua et al., "A survey of defensive deception: Approaches using game theory and machine learning," IEEE Communications Surveys & Tutorials, vol. 23, no. 4, pp. 2460–2493, 2021.

[48] H. Sun and R. Grishman, "Employing lexicalized dependency paths for active learning of relation extraction," Intelligent Automation & Soft Computing, vol. 34, no. 3, pp. 1415–1423, 2022.

[49] Altayeva, A., Omarov, B., & Im Cho, Y. (2018, January). Towards smart city platform intelligence: PI decoupling math model for temperature and humidity control. In 2018 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 693-696). IEEE.

[50] F. Bozyiğit, O. Doğan and D. Kilinç, "Categorization of Customer Complaints in Food Industry Using Machine Learning Approaches," Journal of Intelligent Systems: Theory and Applications, vol. 5, no. 1, pp. 85–91, 2022.

[51] B. Omarov, S. Narynov, Z. Zhumanov, A. Gumar and M. Khassanova, "A skeleton-based approach for campus violence detection," Computers, Materials & Continua, vol. 72, no.1, pp. 315–331, 2022.