Offensive Language Detection on Online Social Networks using Hybrid Deep Learning Architecture

Gulnur Kazbekova[1], Zhuldyz Ismagulova[2], Zhanar Kemelbekova[3],
Sarsenkul Tileubay[4], Boranbek Baimurzayev[5], Aizhan Bazarbayeva[6]
Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan[1, 2, 5, 6]
M. Auezov South Kazakhstan University, Shymkent, Kazakhstan[3]
Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan[4]

***Abstract*—In the digital era, online social networks (OSNs) have revolutionized communication, creating spaces for vibrant public discourse. However, these platforms also harbor offensive language that can proliferates hate speech, cyberbullying, and discrimination, significantly undermining the quality of online interactions and posing severe social implications. This research paper introduces a sophisticated approach to offensive language detection on OSNs, employing a novel Hybrid Deep Learning Architecture (HDLA). The urgency of addressing offensive content is juxtaposed with the challenges inherent in accurately identifying nuanced communications, thus necessitating an advanced model that transcends the limitations of traditional natural language processing techniques. The proposed HDLA model synergistically integrates Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks, capitalizing on the strengths of both methodologies. While the CNN component excels in the hierarchical extraction of spatial features within text data, identifying offensive patterns often concealed in the structural nuances, the LSTM network, adept in processing sequential data, captures the contextual dependencies in user posts over time. This duality ensures a comprehensive analysis of complex linguistic constructs, enhancing the detection accuracy for both overt and covert offensive content. Our research meticulously evaluates the HDLA model using extensive, multi-source datasets reflective of diverse OSN environments, establishing benchmarks against prevailing deep learning models. Results indicate a substantial improvement in precision, recall, and F1-score, demonstrating the model's efficacy in identifying offensive language amidst varying degrees of subtlety and complexity. Furthermore, the model maintains high interpretability, providing insights into the intricate mechanisms of offensive content propagation. Our findings underscore the potential of HDLA in fostering healthier online communities by efficiently curating digital content, thereby upholding the integrity of digital communication spaces.**

Keywords—Offensive language; machine learning; deep learning; social media; detection; classification

I. Introduction

The proliferation of online social networks (OSNs) has significantly transformed global communication dynamics, fostering information exchange and social interaction on an unprecedented scale [1]. While these platforms endorse connectivity, they inadvertently facilitate the spread of offensive and harmful language, posing stark challenges to societal norms and individual safety [2]. Instances of hate speech, cyberbullying, and targeted offensive campaigns have been escalating, necessitating robust detection mechanisms [3].

Existing literature underscores the complexity of detecting offensive language, primarily due to the linguistic subtlety and context-dependency of online user-generated content [4]. Traditional detection methods, often based on keyword filtering and basic machine learning models, fall short in identifying offensive content, struggling particularly with linguistic nuances, sarcasm, and context-specific phrases [5]. Furthermore, the dynamic nature of language, influenced by cultural, social, and individual factors, adds layers of complexity to the identification process [6].

Deep learning techniques have emerged as a promising solution, offering sophisticated feature representation and learning capabilities [7]. Studies leveraging Convolutional Neural Networks (CNNs) have demonstrated success in text classification and offensive content detection, owing to their ability to capture hierarchical text features [8]. Separately, Long Short-Term Memory (LSTM) networks, a form of recurrent neural networks (RNNs), have proven effective in understanding sequential data, thereby interpreting the context within the text efficiently [9]. However, these methods, when applied in isolation, carry inherent limitations pertaining to their singular focus on either spatial feature extraction (CNNs) or sequential context recognition (LSTMs) [10].

Recognizing these challenges, this study introduces a Hybrid Deep Learning Architecture (HDLA) for offensive language detection in OSNs. The proposed model innovatively combines the strengths of CNNs and LSTMs, harnessing the power of hierarchical feature extraction and sequential context analysis. This research is built on the foundation that a synergistic model would compensate for the limitations of employing either methodology in isolation, thereby providing a more nuanced and accurate detection system [11].

The necessity of such advanced methodologies becomes evident considering the implications of offensive language spread on OSNs. Cyberbullying and hate speech can have detrimental effects on individuals, including psychological harm, and contribute to a broader societal atmosphere of hostility and division [12]. Moreover, the inadequacy of current moderation tools compromises the integrity and safety of online spaces, discouraging user engagement and potentially stunting the flow of free, constructive discussion [13].

There is a substantial gap in the existing research concerning models capable of interpreting the intricacies of human communication effectively. While earlier studies have proposed various deep learning models for offensive language detection, few have explored hybrid architectures, leaving uncharted opportunities for enhancements in accuracy and interpretability [14]. Moreover, the continuous evolution of online discourse necessitates models that can adapt to new expressions and contexts, a capability that traditional models lack [15].

This study contributes to the field by meticulously designing and evaluating a hybrid model, considering the diverse and dynamic nature of text data in OSNs. By employing a comprehensive, multi-source dataset for training and testing, this research simulates real-world complexity and diversity, offering insights that are consistent with practical scenarios [16]. Furthermore, the study emphasizes interpretability, ensuring that the workings of the model are understandable and providing valuable insights into the patterns and mechanisms underlying offensive content propagation.

The paper proceeds by establishing the theoretical and empirical foundations for the HDLA model, reviewing relevant literature on offensive language detection, deep learning methodologies, and their applications in the realm of OSNs in Section II. Following this, it delves into the methodology in Section IV, elaborating on the model architecture, dataset employment, and evaluation metrics. Experimental setup in Section V. Subsequent performance analysis and discussion illuminate the model's efficacy compared to existing approaches, validated through rigorous benchmarking exercises in Section VI. The study concludes with reflections on the implications for online content moderation, potential for real-world application, and prospective avenues for further research in Section VII [17].

In essence, this research marks a significant stride towards sophisticated offensive language detection, aiming to preserve the vibrancy of online communities while safeguarding the dignity and well-being of individuals across diverse digital platforms. Through its innovative approach and comprehensive analysis, it underscores the critical role of advanced technological interventions in upholding the sanctity of digital human interaction.

II. Related Works

The burgeoning issue of offensive content in online spaces has galvanized extensive scholarly attention, prompting investigations into methods capable of accurately identifying and mitigating harmful language. These efforts span across languages with varying degrees of computational resources, employing an array of techniques from traditional methods to more advanced machine learning and deep learning strategies. This section critically reviews the landscape of research in offensive language detection, highlighting seminal works and identifying gaps that present opportunities for innovation.

A. *Offensive Language Detection in High Resource Languages*

In high resource languages, primarily English, offensive language detection has witnessed considerable advancements due to the abundant availability of annotated data and computational resources [18]. Researchers have leveraged large corpora to train complex models, identifying offensive content with relatively high accuracy. Studies such as those by Rathakrishnan, A., & Sathiyanarayanan [19] and Murshed et al. [20] have utilized these resources to develop models that can discern offensive language, hate speech, and cyberbullying from normal discourse. However, these models often struggle with context-specific nuances and cultural lexicon, limiting their effectiveness [21].

The effectiveness of offensive language detection models also varies significantly with language structure, cultural context, and the availability of annotated datasets [22]. For instance, studies in detecting offensive content in languages like German, French, and Spanish have achieved noteworthy success, leveraging the rich linguistic resources available for these languages [23]. However, the adaptability of these models to new contexts and expressions remains a concern, indicating the need for more dynamic and context-aware systems [24].

B. *Offensive Language Detection in Low Resource Languages*

Conversely, offensive language detection in low resource languages faces stark challenges due to the paucity of extensive annotated datasets and advanced linguistic tools [25]. Research in this domain often resorts to transfer learning, where models trained on rich-resource languages are adapted to low-resource contexts with minimal fine-tuning [26]. Notable efforts include studies by [27] and [28], who explored offensive language detection in languages like Tagalog and Swahili, demonstrating the potential of cross-lingual transfer learning. Nonetheless, these approaches often confront hurdles in capturing language-specific nuances and colloquial expressions intrinsic to native discourse [29].

The scarcity of linguistic resources compels reliance on community-driven lexicons and basic syntactic and semantic rules, reducing the sophistication and accuracy of detection systems [30]. Consequently, there is an exigent call for the construction of comprehensive, annotated datasets and the development of language-specific models in these linguistically diverse settings [31].

C. *Traditional Methods in Offensive Language Detection*

Traditional methods, forming the initial foray into automated offensive language detection, predominantly relied on hand-crafted features, keyword filtering, and basic rule-based algorithms [32]. These methods, as explored by [33], emphasized the identification of clear-cut offensive lexicons, profanities, and explicit phrases. However, they are notoriously deficient in handling sophisticated language constructs, sarcasm, or contextually offensive content, resulting in high false-positive rates [34].

The reliance on lexical attributes and neglect of the structural and contextual aspects of language in these traditional approaches underscores their limitations [35]. Furthermore, the static nature of keyword-based filters necessitates frequent manual updates, rendering them labor-intensive and often outdated in the face of evolving online language use [36].

D. Machine Learning in Offensive Language Detection

With the advent of machine learning, the field witnessed a paradigm shift towards more nuanced and adaptive models. Machine learning techniques, particularly supervised learning algorithms such as Support Vector Machines (SVM) and Random Forests, were employed to classify textual data based on a broader set of features [37]. Vatambeti et al. [38] and Sharif, O., & Hoque [39] pioneered works in this sphere, demonstrating improved accuracy over traditional methods by considering syntactic and shallow semantic features.

Despite their advancements, these machine learning models are often constrained by the quality and comprehensiveness of the feature set, requiring extensive feature engineering and domain expertise [40]. Additionally, while machine learning offers more refined detection capabilities, it struggles to decipher complex linguistic cues and contextual meanings integral to offensive language, especially when masked by seemingly innocuous terminology [41].

E. Deep Learning in Offensive Language Detection

Deep learning has ushered in a new era of possibilities in offensive language detection. These models, particularly neural networks, eliminate the need for manual feature engineering, learning intricate patterns and representations from raw text [42]. Convolutional Neural Networks (CNNs) have been instrumental in capturing local dependencies and recognizing offensive patterns within the text data [43]. Work by [44] established the CNN’s efficacy in text classification tasks, inspiring subsequent research in offensive language detection.

Moreover, Recurrent Neural Networks (RNNs) and their advanced variant, Long Short-Term Memory networks (LSTMs), have gained prominence for their aptitude in handling sequential data, offering a deeper understanding of contextual information in sentences [45]. This attribute is crucial in deciphering offensive content embedded in conversational threads or sentences reliant on context for interpretation [46]. Next study [47] on using LSTMs for offensive language detection in Twitter data underscores the model's success.

However, deep learning models, while powerful, are not without their challenges. They demand extensive annotated data, are often perceived as "black boxes" due to their complex architectures, and can falter in the face of ambiguous or creatively disguised offensive content [48]. Recent research has started addressing these challenges by proposing hybrid models, combining the strengths of CNNs and LSTMs, or integrating attention mechanisms to enhance model interpretability and performance [49].

Despite these significant strides, the literature collectively points to persistent challenges in balancing high detection accuracy with context sensitivity, especially in linguistically diverse online environments. The nuances of language, ever-evolving use of lexicon, and cultural variations continue to complicate the landscape of offensive content detection [50]. This research gap necessitates continued exploration into advanced model architectures, like the proposed HDLA, that promise enhanced performance by synergizing various aspects of deep learning technology.

In conclusion, while substantial progress has been made, the quest for highly accurate, context-aware, and language-sensitive offensive language detection systems remains an active and exigent field of research. The current study positions itself within this ongoing discourse, aspiring to contribute a nuanced detection approach that acknowledges linguistic diversity and the sophisticated manifestations of offensive language in digital communication [51]. By introducing a hybrid deep learning approach, this research seeks to address the identified gaps and limitations evident in the current body of literature, marking a step forward in the realm of safer and more respectful online interactions [52].

III. Problem Statement

The challenge of promptly detecting cyberbullying within online social networking platforms potentially operates independently from the intricacies involved in categorizing various forms of such digital harassment. Within the context delineated for this study, we encounter a set of social media interactions, herein designated under the collective term "S." It is plausible to consider that within this aggregation, certain exchanges manifest characteristics of cyberbullying.

The dynamic interactions occurring within these social media platforms can be conceptualized and subsequently articulated through a series of networking sessions. These sessions, characterized by their sequential nature, can be mathematically represented, facilitating a systematic analysis. The representation of such a sequence within the network sessions can be encapsulated in Eq. (1), expressed below:

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

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

This formal representation serves as a foundational framework in our endeavor to identify patterns indicative of cyberbullying activities within the vast, interconnected realms of social networking sites. By establishing a mathematical basis for these interactions, we enhance the precision and objectivity of subsequent analyses, thereby refining the processes underlying the early detection and classification of cyberbullying instances.

The order of submissions within a given session exhibits variability, dynamically altering across temporal junctions. This fluidity in sequence progression is influenced by a constellation of determinants that govern the interactive patterns observable during these specific temporal frames. This inherent non-static nature of user engagement underscores the complexity of behavioral patterns within online platforms, necessitating a nuanced understanding and approach to studying interaction dynamics in digital communication environments.

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

In this context, the tuple 'P' epitomizes the kth entry within a particular social network session, while 's' designates the precise chronological marker denoting the publication instance of post 'P'. This formulation underscores the temporal dimension by associating each discrete communicative act, represented by 'P', with a specific moment, identified by 's', thereby capturing the sequential dynamics integral to interactions within the session's framework.

Concurrently, a distinctive array of attributes is employed to characterize each individual post, ensuring a representation that is unequivocally unique. This methodological approach underscores the utilization of a feature vector, articulating a multidimensional space that encapsulates the singularities of each entry within the communicative exchange. Through this, every post is afforded a distinct identification schema, enabling nuanced differentiation and detailed analysis within the collective dataset.

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

Consequently, the endeavor's focal point is the assimilation of requisite expertise for the formulation of a function, denoted as 'f', with the capability to discern the presence or absence of hate speech affiliations within a given text. This intellectual pursuit involves not only the understanding of linguistic and contextual intricacies inherent in hate speech but also the computational mechanisms necessary for the accurate operationalization of 'f'. The ultimate aspiration is to engineer a methodological apparatus that, through 'f', can reliably navigate the subtleties of language, thereby flagging content that aligns with the characteristics of hate speech, while minimizing false positives that can stem from misinterpretation or lack of contextual consideration.

IV. Materials and Methods

A. *The Proposed Framework*

A schematic illustration of the constructed model, tailored for the detection of cyberbullying instances is given in this section. Fig. 1 demonstrates a sample of LSTM network. This model is meticulously architected through several cardinal phases: the preprocessing stage, wherein the data is refined and primed for analysis; the feature extraction stage, responsible for distilling relevant attributes from the data; the classification stage, which employs certain criteria for predictive delineation; and finally, the assessment stage, which critically evaluates the outcomes.

In the ensuing discussion, each phase of the model is subjected to an exhaustive analytical scrutiny. This rigorous exploration is pivotal in elucidating the nuanced methodologies employed at each juncture, thereby highlighting their collective contribution to the model's overall precision and effectiveness. Through this, the paper seeks to accentuate the underlying complexities and the methodical considerations incumbent in developing a robust computational model capable of identifying cyberbullying with high accuracy. A schematic illustration of the constructed model, tailored for the detection of cyberbullying instances, is depicted in Fig. 2.

B. *Feature Extraction*

1) Term frequency-inverse document frequency. Within the scope of this research, the Term Frequency-Inverse Document Frequency (TF-IDF) methodology is employed as a crucial vectorization technique, instrumental in the transformation of textual data into feature vectors that can be efficaciously processed by machine learning algorithms. This subsection delves into the systematic application and theoretical underpinnings of TF-IDF in the context of identifying cyberbullying instances on online platforms.

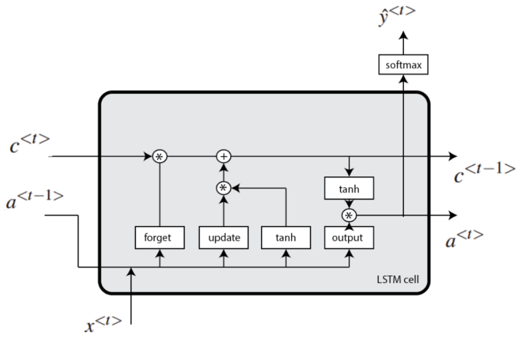

Fig. 1. LSTM network.

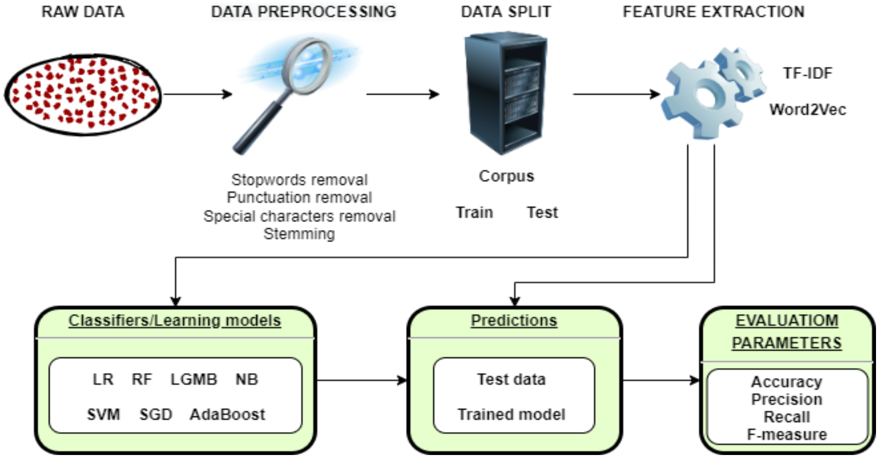

Fig. 2. Proposed framework.

TF-IDF, a renowned technique in text mining and information retrieval, quantifies the importance of specific terms within a corpus, contextualized by their frequency in individual documents and rarity across the entire dataset [53]. This technique is bifurcated into two primary components: Term Frequency (TF) and Inverse Document Frequency (IDF).

a) Term Frequency (TF): This component computes the recurrence of a term within a single document, positing that the relevance of the term augments proportionally with its frequency of occurrence [54].

b) Inverse Document Frequency (IDF): Complementing TF, IDF ascertains the scarcity of a term across the corpus, assigning more weight to terms that provide higher discriminatory power due to their infrequency. The mathematical formulation of IDF mitigates the prominence of terms that are ubiquitous across documents, thus offering a balanced view of term significance [55].

2) Word2Vec embedding. The complexity of detecting offensive language within the vast spectrum of human interaction necessitates an approach that transcends mere keyword spotting techniques, demanding a deeper understanding of contextual linguistic relationships. This research incorporates the Word2Vec model, known for its efficiency in capturing semantic relations between words, offering an advanced linguistic parsing mechanism vital for offensive language detection.

In the context of offensive language detection, Word2Vec plays a crucial role during the feature extraction phase. Here, textual data, laden with potential offensive content, is converted into vectors. This vectorization process is not arbitrary but is reflective of the words' semantic relationships within the dataset, informed by the context in which they appear.

The Skip-Gram model processes each word and its contextual neighbors, adjusting the vectors to closely represent the relational semantics in a multi-dimensional space. This approach is particularly pertinent to offensive language detection, as language nuances, euphemisms, and community-specific lexis often used in offensive content are contextually bound and not readily identifiable through conventional keyword detection methods [56].

In the present research, the selected method for weighting is the term frequency-inverse document frequency (tf-idf) system. To compute the tf-idf weight associated with the ith term within the jth document, the subsequent equation is employed:

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

3) Bag of Words. Within the scope of computational linguistics, the Bag of Words (BoW) model stands as a simplified representation, used to preprocess the text by transforming it into a set of distinguishable words, or "tokens," thereby constructing a dictionary of the language used in the entire text corpus [57].

In the context of offensive language detection, the BoW model serves a fundamental role. By disregarding the syntactic relationships between words and focusing solely on the occurrence frequency, BoW facilitates a form of "token-based" analysis. Each unique word in the text is interpreted as a feature, and the value corresponds to the frequency of that word in the document. Despite its simplicity, this model offers considerable utility in scenarios where the structural complexity of text is less consequential compared to the relevance of word occurrences [58].

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

C. Machine Learning Methods

In the realm of hate speech detection, several machine learning algorithms have gained prominence due to their efficacy in classifying and predicting offensive content. Each algorithm's unique computational approach aids in the nuanced identification of hate speech within various digital communications.

1) Decision trees: Representing a form of supervised learning, Decision Trees create a framework that categorizes input data into specific output classes based on their statistical properties, effectively forming a tree of decisions [59]. These trees offer a highly interpretable model and adeptly manage non-linear relationships. Within hate speech detection, they function by analyzing features extracted from the text, such as the frequency of particular words or the presence of certain lexical items, thereby facilitating nuanced content-based decision-making processes.

2) Naïve bayes classifiers: These are grounded in the principles of Bayes' theorem and operate under the assumption of independence among predictors [60]. Despite its inherent assumption of feature independence, which oversimplifies linguistic relationships, the Naïve Bayes algorithm often yields robust performance in text classification. In the specific context of hate speech detection, it evaluates the probability of a message being categorized as hate speech, considering the presence of indicative terms or phrases.

3) K-Nearest Neighbors (K-NN): Functioning as a non-parametric method, K-NN employs instance-based learning, classifying data points based on the characteristics of neighboring instances [61]. Within hate speech detection paradigms, K-NN leverages the comparative analysis of feature similarities, utilizing representations like word embeddings or TF-IDF vectors, to classify textual data. This method hinges on identifying the most common classification among the 'K' nearest references in the feature space.

4) Support Vector Machines (SVM): SVMs operate through supervised learning, designed to discern the optimal boundary between multiple classes, effectively managing scenarios with high-dimensional spaces [62]. They are adaptable through the use of diverse kernel functions to handle non-linear feature spaces. In hate speech detection, SVMs are instrumental in discerning boundaries in feature representation, relying on text attributes like word frequencies, n-grams, or sentiment indicators, thereby enhancing the precision of classification tasks.

Each of these algorithms, with their distinct methodological underpinnings, contributes to the more extensive framework of hate speech detection, providing comprehensive analytical capabilities essential for effectively navigating the complexities of digital communication landscapes.

D. Deep Learning Methods

1) LSTM (Long Short-Term Memory): Long Short-Term Memory (LSTM) networks, a variant of recurrent neural networks (RNNs), specifically address the challenges of learning long-term dependencies, thereby mitigating the vanishing gradient problem inherent in traditional RNNs. This is crucial for tasks such as text analysis, where understanding the sequence and context is essential [63]. Fig. 3 demonstrates architecture of LSTM network.

The unique component of LSTMs is their cell state, often conceptualized as the network's "memory," adjusted through structures called gates. The cell state C_t at time t is modified by three gates: the input gate (i_t), the forget gate (f_t), and the output gate (o_t), calculated as follows:

$$\begin{aligned} i_t &= \delta\left(W_{xi}x_t + W_{ht}h_{t-1} + b_i\right) \\ f_t &= \delta\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \\ o_t &= \delta\left(W_{xo}x_t + W_{ho}h_{t-1} + b_o\right) \end{aligned} \tag{6}$$

where x_t is the input at time t, h_{t-1} is the previous hidden state, W and b are the weight matrices and bias terms, respectively. The crucial cell state update is then performed as:

$$C_t = f_t * C_{t-1} + i_t * \tan g\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \tag{7}$$

This architecture allows LSTMs to selectively enhance or diminish the information passed along the sequence, making them particularly adept at modeling sequential data with complex dependencies.

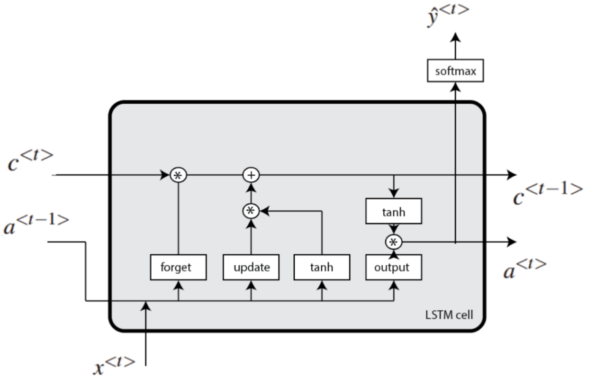

Fig. 3. Architecture of LSTM network.

2) BiLSTM (Bidirectional Long Short-Term Memory): Bidirectional Long Short-Term Memory (BiLSTM) networks augment the architecture of traditional LSTM by processing the data in both forward and backward directions, offering enhanced understanding through a two-way sequence representation [64]. This approach ensures that the information from both past and future contexts is utilized during the learning process, thus enriching the representation of each point in the sequence.

The BiLSTM model incorporates two layers of LSTMs: one processes the sequence from start to end (forward LSTM), and the other from end to start (backward LSTM). The final representation of each sequence point is the concatenation of the forward and backward information:

$$H_t = \left\lfloor \overrightarrow{H_t}; \overleftarrow{H_t} \right\rfloor \tag{7}$$

where, $\overrightarrow{H_t}$ and H_t are the hidden states of the forward and backward LSTMs at time t respectively.

For each direction, the LSTM computations are similar to the unidirectional case, with the same gating mechanisms and cell state updates:

$$\overrightarrow{H_t} = \overrightarrow{LSTM}\left\lfloor x_t, \overrightarrow{H_{t-1}} \right\rfloor \tag{8}$$

$$\overleftarrow{H_t} = \overleftarrow{LSTM}\left\lfloor x_t, \overleftarrow{H_{t-1}} \right\rfloor \tag{9}$$

By synthesizing context from both directions, BiLSTMs provide a richer, more comprehensive feature extraction, significantly improving the performance on tasks requiring an understanding of the entire data sequence, such as text classification and sentiment analysis. Fig. 4 demonstrates architecture of BiLSTM network.

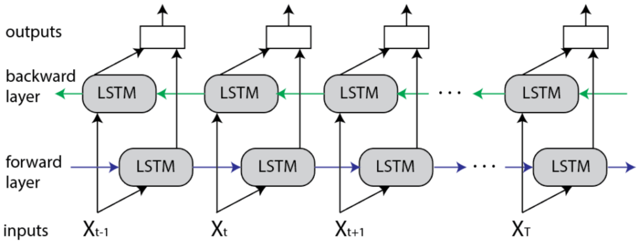

Fig. 4. BiLSTM network.

Convolutional Neural Networks (CNNs) are a class of deep neural networks highly effective in recognizing patterns directly from pixels of images, allowing hierarchical pattern recognition. They are especially powerful for tasks such as image classification, often producing superior results compared to traditional methods [65].

A key component of CNNs is the convolutional layer, which applies numerous learnable filters to the input. Each filter is used for the convolution operation, producing a feature map. Formally, for each spatial position, the convolution is computed as:

$$(F * G)(i, j) = \sum_m \sum_n F(m,n) \cdot G(i-m, j-n) \tag{10}$$

where, F is the filter matrix, G is a region of the input image, and ∗ denotes the convolution operation. The result is a feature map that undergoes a non-linear transformation (usually ReLU).

CNNs also include pooling layers, typically max pooling, which reduce the spatial dimensions (downsampling) of the input representation, decreasing the computational complexity and allowing for feature invariances. The layers of convolution and pooling operations are followed by fully connected layers that perform high-level reasoning on the features.

By learning hierarchies of features through backpropagation, CNNs construct increasingly complex representations of the input data, making them highly proficient at visual understanding. Fig. 5 demonstrates architecture of the convolutional neural network.

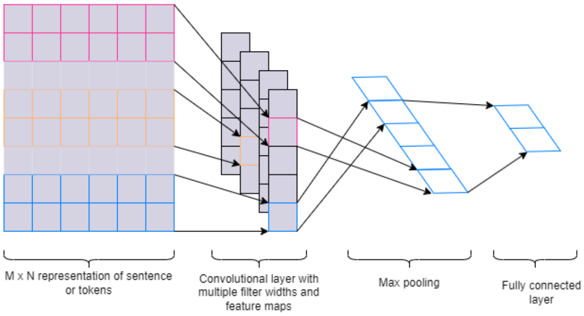
M x N representation of sentence or tokens
Convolutional layer with multiple filter widths and feature maps
Max pooling
Fully connected layer

Fig. 5. Sample of a CNN architecture for offensive language detection.

V. Experimental Setup

A. Evaluation Parameters

In the realm of offensive language detection within digital platforms, the accuracy metric serves as a fundamental gauge for evaluation. Accuracy is quantified as the ratio of correctly identified instances—both offensive and non-offensive—to the total number of instances examined. Mathematically, it is expressed as [66]:

$$accuracy = \frac{TP + TN}{P + N} \tag{6}$$

While this metric provides an initial insight into the model's performance, relying solely on accuracy can be misleading, particularly in imbalanced datasets where non-offensive classes may significantly outnumber offensive ones. Therefore, accuracy is often employed alongside other metrics to furnish a more comprehensive evaluation landscape.

1) Precision: Precision is a critical metric in the evaluation of models tasked with offensive language detection, focusing specifically on the exactness of the classification. In this context, precision is the ratio of correctly predicted offensive instances to all instances predicted as offensive, whether rightly or wrongly identified. Formally, precision (P) is defined as [67]:

$$preision = \frac{TP}{TP + FP} \tag{7}$$

This metric is paramount in scenarios where the cost of false positives is high. For instance, incorrectly classifying content as offensive could impinge on free speech, making precision a vital measure of a model's reliability in distinguishing genuinely offensive content from the non-offensive.

2) Recall: Recall, in the context of offensive language detection, is an indispensable metric that quantifies the model's capacity to identify the entirety of offensive instances within a dataset. It is defined as the ratio of correctly predicted offensive comments (True Positives) to the total amount of offensive comments actually present in the data (True Positives + False Negatives), mathematically represented as [68]:

$$recall = \frac{TP}{TP + FN} \tag{8}$$

This metric is particularly crucial in scenarios where the implications of overlooking offensive content are severe, demanding a system sensitive enough to capture as many offensive instances as possible, thereby prioritizing the minimization of false negatives.

3) F-score: In offensive language detection, the F-score (or F1-score) is a crucial metric that balances precision and recall, providing a singular measure for the effectiveness of a classifier in identifying offensive content. The F-score is the harmonic mean of precision and recall, ensuring a robust metric that accounts for both false positives and false negatives. It is defined as [69]:

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{9}$$

Given that precision underscores the avoidance of over-policing while recall emphasizes the importance of not overlooking offensive content, the F-score harmonizes these aspects, offering a comprehensive measure of a model's performance.

4) ROC curve: The Receiver Operating Characteristic (ROC) curve is a fundamental tool for diagnostic test evaluation in offensive language detection systems. It graphically portrays the trade-off between true positive rate (sensitivity) and false positive rate (1 - specificity) across various threshold settings, highlighting the model's performance in terms of its discriminatory capacity. The area under the ROC curve (AUC-ROC) quantifies the overall ability of the model to discern between offensive and non-offensive content, irrespective of threshold. A perfect model scores an AUC of 1, while a score of 0.5 suggests no discrimination capability, equivalent to random guessing. This metric's resilience against class imbalance makes it essential for unbiased model evaluation.

B. Experimental Results

In the domain of cyberbullying detection research, several critical metrics are instrumental in evaluating model performance, including Accuracy, Precision, Recall, F-measure, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The elucidation of the effectiveness of various methodologies employed within this study is visually represented through confusion matrices, as depicted in Fig. 6. These matrices provide an insightful depiction of classification outcomes, distinctly presenting the distribution of predictions across different categories.

This investigation categorizes online interactions into three distinct classes, assigning them numerical representations for clarity and analytical rigor: 'cyberbullying' (scored as 1), 'non-cyberbullying' (scored as 0), and a 'neutral' category (scored as 2). Through this classification, the research not only underscores the nuanced nature of online discourse but also enhances the precision in quantifying the instances and nature of cyberbullying, facilitating a more robust and detailed analysis.

Fig. 7 critically contrasts the proposed model against a spectrum of extant machine learning and deep learning models, adjudging their efficacies. This rigorous assessment involves computing the area under the receiver operating characteristic curve (AUC-ROC), which encapsulates the totality of attributes extracted for each classification paradigm. Subsequently, Fig. 8 presents an exhaustive comparative analysis of the AUC-ROC curves emanating from each deployed strategy alongside the advocated methodology.

A salient observation from this visual representation indicates that deep learning frameworks, particularly the BiLSTM model, consistently outperform traditional machine learning counterparts. This assertion is corroborated by the superior AUC-ROC values exhibited by the BiLSTM model, commencing from the initial iteration and sustained throughout the subsequent procedural timeline, underscoring its predictive precision and reliability.

Table I delineates the classification outcomes pertinent to cyberbullying instances, derived through the application of various machine and deep learning algorithms across three distinct datasets. The evaluative criteria employed encompassed a range of metrics, including accuracy, precision, recall, and F1-score [70], offering a comprehensive perspective on the performance benchmarks of the machine learning and deep learning methods under scrutiny. This systematic approach ensures a holistic and nuanced understanding of each model's strengths and potential areas for enhancement in the context of cyberbullying detection.

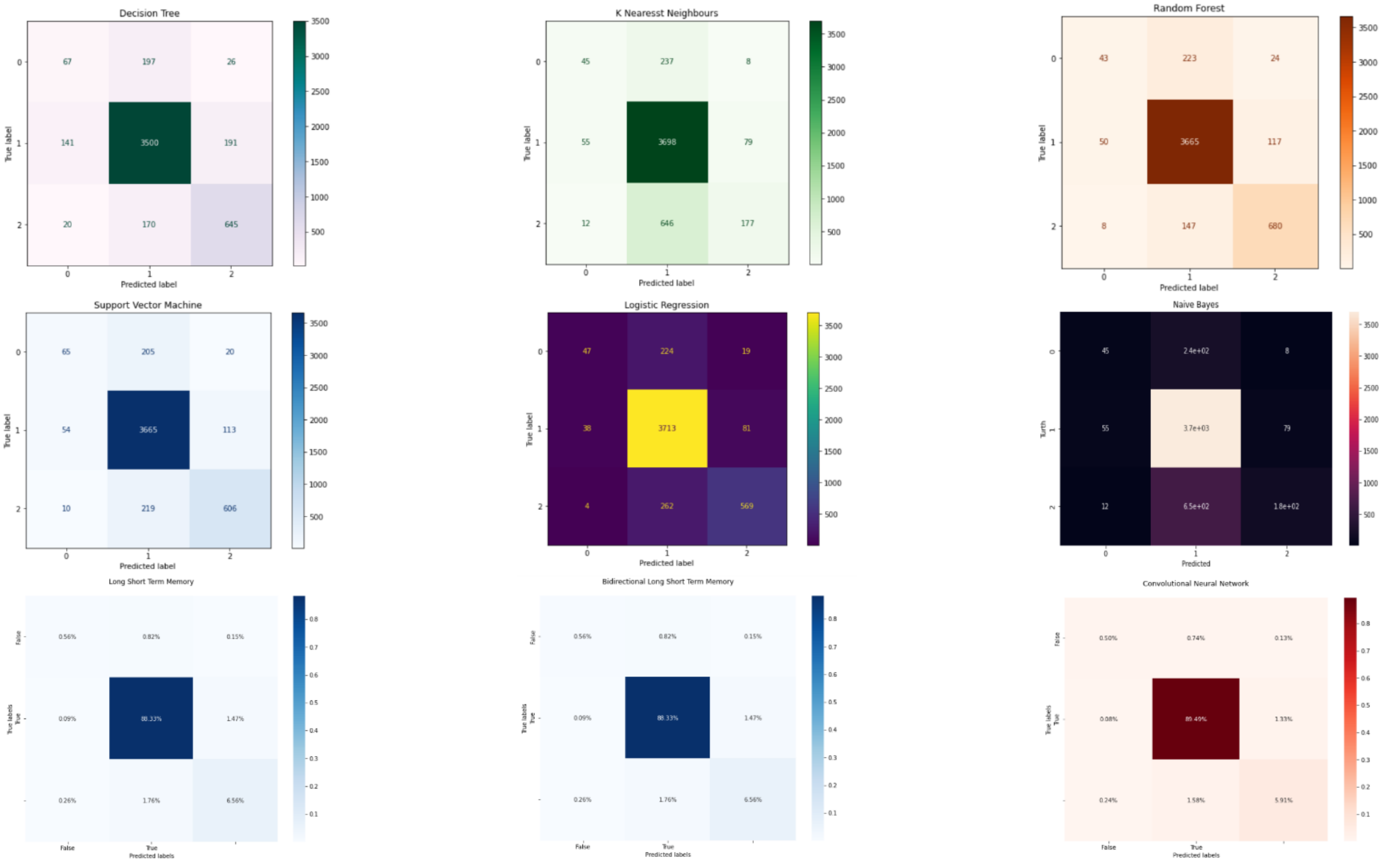

Fig. 6. Confusion matrices for hate speech detection using different machine learning methods.

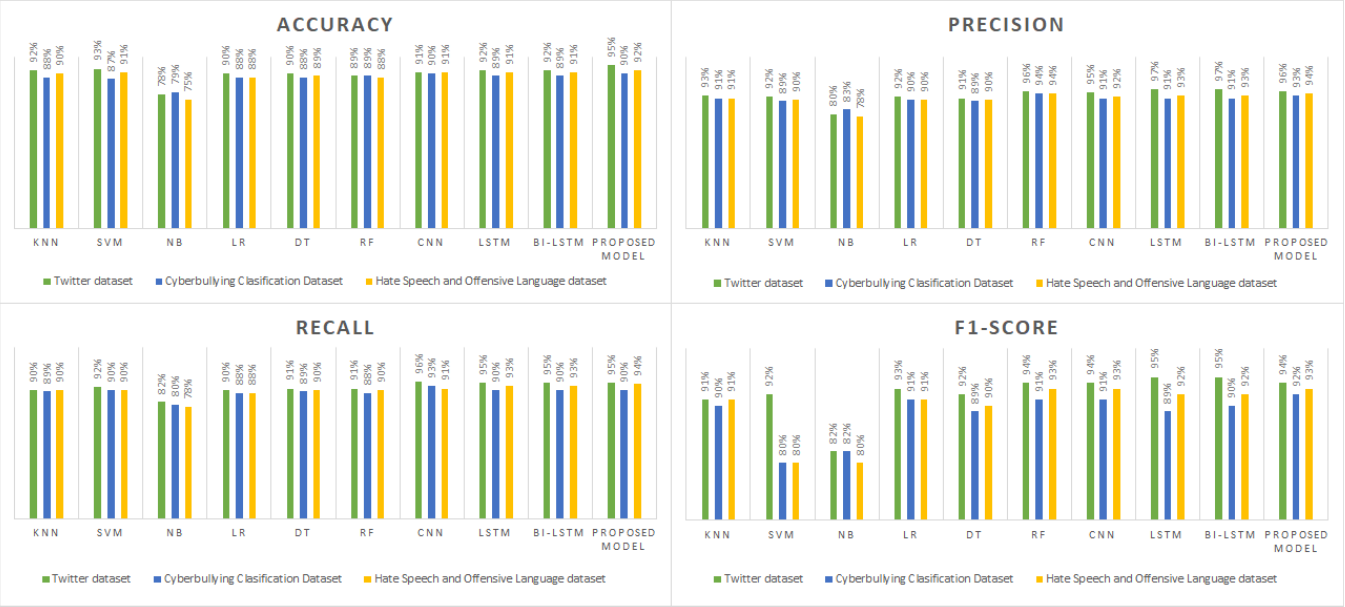

Fig. 7. Evaluation parameters for different datasets using machine learning and deep learning models.

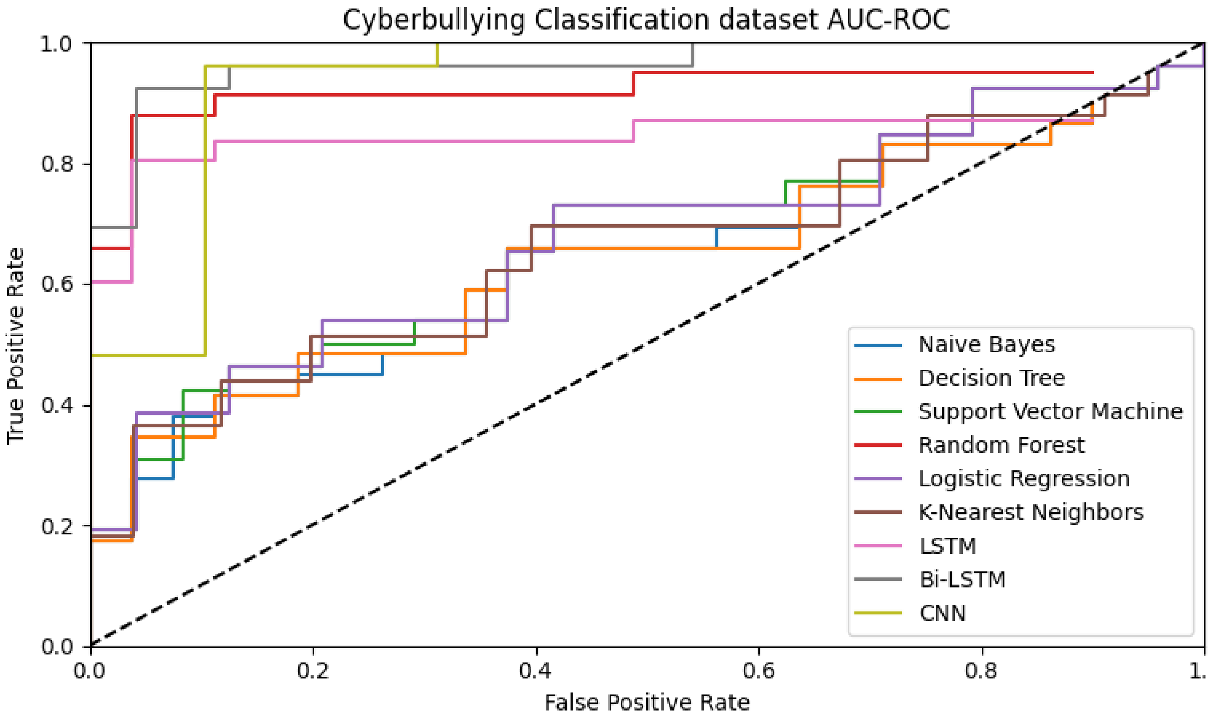

Fig. 8. ROC curve of applied machine learning and deep leaerning techniques for hate speech detection

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

In light of the compelling performance metrics attained, the proposed methodology emerges as a viable approach for the identification of cyberbullying activities within social networking platforms. Furthermore, when evaluated against all performance benchmarks, the introduced deep neural network stands paramount, particularly in discerning instances of cyberbullying.

The efficacy of the proposed deep neural network can be attributed not only to the refinement of weights and biases but also to its optimization leading to reduced training duration. This streamlined process, indicative of the method's robustness, fosters favorable outcomes, reinforcing the technique's applicability and effectiveness.

Crucially, the findings suggest that the innovative application of deep neural networks, as advocated in this study, exhibits a high degree of adaptability, capable of accommodating texts of varying lengths. This flexibility, intrinsic to the proposed model, signifies its potential for broader applicability and scalability within current digital communication contexts, thereby bolstering its practicality in real-world scenarios.

VI. Discussion

This research ventured into the critical realm of cyberbullying detection within social networking sites, recognizing the profound impact that online harassment can have on individuals and communities. The study's underpinning was the development and assessment of a novel deep learning strategy designed to efficiently and accurately identify instances of cyberbullying, a task that traditional machine learning models have approached with varying degrees of success.

One of the salient aspects of this research was its emphasis on deep learning models, particularly BiLSTM, and their capacity to outstrip the performance of conventional machine learning approaches in this domain. These models, known for their proficiency in handling sequential data, proved adept at capturing the nuanced context embedded in human language, a critical factor in accurately detecting cyberbullying.

The superiority of the BiLSTM model, as evidenced through various performance metrics, underscores a pivotal shift in computational linguistics, highlighting the increasing relevance of models that understand the intricacies of language and context. This contextual understanding is paramount in the realm of cyberbullying, where the intent behind words can be just as harmful as the content itself. The model's ability to discern subtle nuances comes from its architectural advantage, allowing it to retain information over prolonged sequences and thereby understanding context better than its machine learning counterparts.

However, the journey to this point was not without its challenges. One of the primary obstacles was the variability of language used in cyberbullying. Slang, misspellings, regional dialects, and code-switching are rampant in online communications, presenting hurdles in training models that traditionally rely on standard language rules. The research navigated this by enriching the training data and iteratively refining the neural network parameters, which was pivotal in enhancing the model's ability to understand and interpret the eclectic nature of online discourse.

Moreover, the ethical implications of automated cyberbullying detection were considered, acknowledging the delicate balance between flagging harmful content and preserving user privacy and freedom of speech. The model's design required careful consideration to respect users' digital rights while maintaining its commitment to creating safer online environments. This dual commitment is reflected in the model's methodology, emphasizing user protection while striving for comprehensive detection and minimal false positives.

Comparatively, the study's findings align with the current trajectory in cyberbullying research and technological advancements in artificial intelligence. They underscore the potential deep learning holds in transforming safety measures on social networking platforms. However, they also bring to light certain limitations that future studies will need to address.

Firstly, while the model showcased high efficiency, the question of scalability remains. As social media content continues to grow exponentially, the ability of this model to process vast quantities of data without compromising performance is something future research needs to explore. Additionally, the adaptability of the model in real-time detection scenarios are aspects that necessitate further investigation, considering the dynamic nature of online interactions.

Secondly, the diversity in data sets poses both a challenge and an opportunity. The model's performance across various demographics, cultures, and languages is a testament to its robustness. However, there's a recognized need for more diverse and inclusive data sets to ensure the model's efficacy across broader spectrums of society. This inclusivity extends beyond linguistic diversity to encompass different forms of cyberbullying, acknowledging that online harassment transcends overt language to include subtler, equally damaging forms.

Furthermore, the research's focus on deep learning, while justified, also highlights the need for interdisciplinary approaches in future studies. The psychological, sociological, and cultural dimensions of cyberbullying demand a holistic approach to technology-based solutions. Collaborations across various fields could enhance the technological frameworks proposed by this study, ensuring they are grounded in the multifaceted reality of cyberbullying.

In conclusion, this study marks a significant step forward in employing advanced AI technologies in the battle against online harassment. However, it also serves as a reminder of the work still required to perfect these systems. As we move forward, the goal remains clear: to harness the power of technology in creating online spaces where safety, respect, and freedom of expression coexist. The journey, though complex, holds the promise of achieving a more harmonious digital society, and this research serves as both a catalyst and a beacon in that quest.

VII. Conclusion

The journey through this research has underscored the intricate challenges and profound necessities within the realm of detecting and mitigating cyberbullying across social media platforms. In an era where digital interactions are an extension of our social fabric, ensuring the virtual environment's safety becomes paramount. This study ventured beyond traditional machine learning methodologies, embracing the nuanced capabilities of deep learning mechanisms, particularly through the adoption of the BiLSTM model. The findings reaffirm the assertion that understanding the sequential and contextual aspects of language is crucial in the accurate detection of cyberbullying. By leveraging the enhanced memory and processing capabilities of BiLSTM, the research demonstrated notable success in identifying offensive content, thereby holding significant implications for safeguarding online communities. However, it was observed that the battle against online harassment is an ongoing process, necessitating continuous advancements and iterations within technological applications.

As we envisage the future of cyber safety, the conclusions drawn here are not terminal but rather serve as a springboard for further exploration and innovation. The success of the proposed model underscores the potential within deep learning methodologies, offering a beacon of hope for substantial progress in this domain. Nonetheless, the complexities of human interaction, the ever-evolving nature of language, and the ethical considerations in digital monitoring present ongoing challenges that must steer future research directions. Collaborative, interdisciplinary approaches may also be essential in addressing these multifaceted issues, uniting technological prowess with psychological, cultural, and linguistic expertise. As we forge ahead, the objective remains steadfast: to refine and enhance these technological guardians to preserve the dignity, safety, and well-being of individuals in the digital sphere, ensuring that our virtual environments are reflective of the respect and security we strive for in our broader societies.

References

[1] Anand, M., Sahay, K. B., Ahmed, M. A., Sultan, D., Chandan, R. R., & Singh, B. (2023). Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques. Theoretical Computer Science, 943, 203-218.

[2] Fale, P. N., Goyal, K. K., & Shivani, S. (2023, April). A hybrid deep learning approach for abusive text detection. In AIP Conference Proceedings (Vol. 2753, No. 1). AIP Publishing.

[3] Al-Sarem, M., Alsaeedi, A., Saeed, F., Boulila, W., & AmeerBakhsh, O. (2021). A novel hybrid deep learning model for detecting COVID-19-related rumors on social media based on LSTM and concatenated parallel CNNs. Applied Sciences, 11(17), 7940.

[4] Ayo, F. E., Folorunso, O., Ibharalu, F. T., & Osinuga, I. A. (2020). Hate speech detection in Twitter using hybrid embeddings and improved cuckoo search-based neural networks. International Journal of Intelligent Computing and Cybernetics, 13(4), 485-525.

[5] Kumar, A., Saumya, S., & Singh, A. (2023). Detecting Dravidian Offensive Posts in MIoT: A Hybrid Deep Learning Framework. ACM Transactions on Asian and Low-Resource Language Information Processing.

[6] Khan, A. A., Iqbal, M. H., Nisar, S., Ahmad, A., & Iqbal, W. (2023). Offensive Language Detection for Low Resource Language Using Deep Sequence Model. IEEE Transactions on Computational Social Systems.

[7] Ahmad, G. I., Singla, J., Anis, A., Reshi, A. A., & Salameh, A. A. (2022). Machine Learning Techniques for Sentiment Analysis of Code-Mixed and Switched Indian Social Media Text Corpus-A Comprehensive Review. International Journal of Advanced Computer Science and Applications, 13(2).

[8] Elzayady, H., Mohamed, M. S., Badran, K. M., & Salama, G. I. (2023). A hybrid approach based on personality traits for hate speech detection in Arabic social media. International Journal of Electrical and Computer Engineering, 13(2), 1979.

[9] Toktarova, A., Syrlybay, D., Myrzakhmetova, B., Anuarbekova, G., Rakhimbayeva, G., Zhylanbaeva, B., ... & Kerimbekov, M. (2023). Hate speech detection in social networks using machine learning and deep learning methods. International Journal of Advanced Computer Science and Applications, 14(5).

[10] Omarov, B., Altayeva, A., & Cho, Y. I. (2017). Smart building climate control considering indoor and outdoor parameters. In Computer Information Systems and Industrial Management: 16th IFIP TC8 International Conference, CISIM 2017, Bialystok, Poland, June 16-18, 2017, Proceedings 16 (pp. 412-422). Springer International Publishing.

[11] Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia tools and applications, 80(8), 11765-11788.

[12] Islam, M. R., Liu, S., Wang, X., & Xu, G. (2020). Deep learning for misinformation detection on online social networks: a survey and new perspectives. Social Network Analysis and Mining, 10, 1-20.

[13] Haq, I., Qiu, W., Guo, J., & Tang, P. (2023). Pashto offensive language detection: a benchmark dataset and monolingual Pashto BERT. PeerJ Computer Science, 9, e1617.

[14] B. Omarov, A. Suliman and K. Kushibar, "Face recognition using artificial neural networks in parallel architecture", Journal of Theoretical and Applied Information Technology, vol. 91, no. 2, pp. 238-248.

[15] Sharma, D. K., Singh, B., Agarwal, S., Pachauri, N., Alhussan, A. A., & Abdallah, H. A. (2023). Sarcasm Detection over Social Media Platforms Using Hybrid Ensemble Model with Fuzzy Logic. Electronics, 12(4), 937.

[16] Weitzel, L., Daroz, T. H., Cunha, L. P., Von Helde, R., & de Morais, L. M. (2023, June). Investigating Deep Learning Approaches for Hate Speech Detection in Social Media: Portuguese-BR tweets. In 2023 18th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-5). IEEE.

[17] Al Banna, M. H., Ghosh, T., Nahian, M. J. A., Kaiser, M. S., Mahmud, M., Taher, K. A., ... & Andersson, K. (2023). A Hybrid Deep Learning Model to Predict the Impact of COVID-19 on Mental Health from Social Media Big Data. IEEE Access.

[18] Abbes, M., Kechaou, Z., & Alimi, A. M. (2023, July). Deep learning approach for Tunisian hate Speech detection on Facebook. In 2023 IEEE Symposium on Computers and Communications (ISCC) (pp. 739-744). IEEE.

[19] Rathakrishnan, A., & Sathiyanarayanan, R. (2023). Rumor detection on social media using deep learning algorithms with fuzzy inference system for healthcare analytics system using COVID-19 dataset. International Journal of Computational Intelligence and Applications, 22(01), 2341008.

[20] Murshed, B. A. H., Abawajy, J., Mallappa, S., Saif, M. A. N., & Al-Ariki, H. D. E. (2022). DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform. IEEE Access, 10, 25857-25871.

[21] Banna, M. H. A., Ghosh, T., Nahian, M. J. A., Kaiser, M. S., Mahmud, M., Taher, K. A., ... & Andersson, K. (2023). A Hybrid Deep Learning Model to Predict the Impact of COVID-19 on Mental Health from Social Media Big Data. IEEE Access, 11, 77009-77022.

[22] Omarov, B., Omarov, B., Shekerbekova, S., Gusmanova, F., Oshanova, N., Sarbasova, A., ... & Sultan, D. (2019). Applying face recognition in video surveillance security systems. In Software Technology: Methods and Tools: 51st International Conference, TOOLS 2019, Innopolis, Russia, October 15–17, 2019, Proceedings 51 (pp. 271-280). Springer International Publishing.

[23] Rehman, A. U., Malik, A. K., Raza, B., & Ali, W. (2019). A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis. Multimedia Tools and Applications, 78, 26597-26613.

[24] Nagar, S., Barbhuiya, F. A., & Dey, K. (2023). Towards more robust hate speech detection: using social context and user data. Social Network Analysis and Mining, 13(1), 47.

[25] Mazari, A. C., & Kheddar, H. (2023). Deep Learning-based Analysis of Algerian Dialect Dataset Targeted Hate Speech, Offensive Language and Cyberbullying. International Journal of Computing and Digital Systems.

[26] Singh, N. K., Singh, P., Das, P., & Chand, S. XRBi-GAC: A hybrid deep learning framework for multilingual toxicity detection. Journal of Intelligent & Fuzzy Systems, (Preprint), 1-13.

[27] Bhuvaneswari, M., & Prabha, V. L. A deep learning approach for the depression detection of social media data with hybrid feature selection and attention mechanism. Expert Systems, e13371.

[28] Yafooz, W. M., Al-Dhaqm, A., & Alsaeedi, A. (2023). Detecting Kids Cyberbullying Using Transfer Learning Approach: Transformer Fine-Tuning Models. In Kids Cybersecurity Using Computational Intelligence Techniques (pp. 255-267). Cham: Springer International Publishing.

[29] Sari, T. I., Ardilla, Z. N., Hayatin, N., & Maskat, R. (2022). Abusive comment identification on Indonesian social media data using hybrid deep learning. IAES International Journal of Artificial Intelligence, 11(3), 895.

[30] Fati, S. M., Muneer, A., Alwadain, A., & Balogun, A. O. (2023). Cyberbullying Detection on Twitter Using Deep Learning-Based Attention Mechanisms and Continuous Bag of Words Feature Extraction. Mathematics, 11(16), 3567.

[31] Sharma, G., Brar, G. S., Singh, P., Gupta, N., Kalra, N., & Parashar, A. (2022, November). An Exploration of Machine Learning and Deep Learning Techniques for Offensive Text Detection in Social Media—A Systematic Review. In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2022, Volume 3 (pp. 541-559). Singapore: Springer Nature Singapore.

[32] Mundra, S., & Mittal, N. (2023). CMHE-AN: Code mixed hybrid embedding based attention network for aggression identification in hindi english code-mixed text. Multimedia Tools and Applications, 82(8), 11337-11364.

[33] Kemal, B. S., Abebe, T. U., Pendem, G. K., Krishna, T. G., & Gemeda, K. A. (2023). Bilingual Social Media Text Hate Speech Detection For Afaan Oromo And Amharic Languages Using Deep Learning. Journal of Namibian Studies: History Politics Culture, 34, 250-281.

[34] Paul, S., Saha, S., & Singh, J. P. (2023). COVID-19 and cyberbullying: deep ensemble model to identify cyberbullying from code-switched languages during the pandemic. Multimedia tools and applications, 82(6), 8773-8789.

[35] Fha, S., Sharma, U., & Naleer, H. M. M. (2023). Development of an Efficient Method to Detect Mixed Social Media Data with Tamil-English Code Using Machine Learning Techniques. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(2), 1-19.

[36] Nagulapati, V. S., Rapelli, S. R., Fadlullah, Z. M., Fouda, M. M., Alasmary, W., & Guizani, M. (2022, May). On Improving Automated Detection of Cyber-Bully in Social Networks with Constrained Datasets: A Hierarchical Deep Learning Approach. In ICC 2022-IEEE International Conference on Communications (pp. 1746-1751). IEEE.

[37] Elzayady, H., Mohamed, M. S., Badran, K., Salama, G., & Abdel-Rahim, A. (2023). Arabic Hate Speech Identification by Enriching MARBERT Model with Hybrid Features. In Intelligent Sustainable Systems: Selected Papers of WorldS4 2022, Volume 2 (pp. 559-566). Singapore: Springer Nature Singapore.

[38] Vatambeti, R., Mantena, S. V., Kiran, K. V. D., Manohar, M., & Manjunath, C. (2023). Twitter sentiment analysis on online food services based on elephant herd optimization with hybrid deep learning technique. Cluster Computing, 1-17.

[39] Sharif, O., & Hoque, M. M. (2022). Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers. Neurocomputing, 490, 462-481.\

[40] Murshed, B. A. H., Suresha, Abawajy, J., Saif, M. A. N., Abdulwahab, H. M., & Ghanem, F. A. (2023). FAEO-ECNN: cyberbullying detection in social media platforms using topic modelling and deep learning. Multimedia Tools and Applications, 1-40.

[41] Yadav, D., & Sain, M. K. (2023). Comparative Analysis and Assesment on Different Hate Speech Detection Learning Techniques. JOURNAL OF ALGEBRAIC STATISTICS, 14(1), 29-48.

[42] Quoc Tran, K., Trong Nguyen, A., Hoang, P. G., Luu, C. D., Do, T. H., & Van Nguyen, K. (2023). Vietnamese hate and offensive detection using PhoBERT-CNN and social media streaming data. Neural Computing and Applications, 35(1), 573-594.

[43] Hasan, M., Islam, L., Jahan, I., Meem, S. M., & Rahman, R. M. (2023). Natural Language Processing and Sentiment Analysis on Bangla Social Media Comments on Russia–Ukraine War Using Transformers. Vietnam Journal of Computer Science, 1-28.

[44] D. Sultan, B. Omarov, Z. Kozhamkulova, G. Kazbekova, L. Alimzhanova et al., "A review of machine learning techniques in cyberbullying detection," Computers, Materials & Continua, vol. 74, no.3, pp. 5625–5640, 2023.

[45] SIMON, Y., Baha, B. Y., & Garba, E. J. (2022). A MULTI-PLATFORM APPROACH USING HYBRID DEEP LEARNING MODELS FOR AUTOMATIC DETECTION OF HATE SPEECH ON SOCIAL MEDIAHate speech on online social networks is a general problem across social media platforms that has the potential of causing physical harm to t. BIMA JOURNAL OF SCIENCE AND TECHNOLOGY (2536-6041), 6(02), 77-90.

[46] Hamza, M. A., Alshahrani, H. J., Tarmissi, K., Yafoz, A., Aziz, A. S. A., Mahzari, M., ... & Yaseen, I. (2023). Improved Attentive Recurrent Network for Applied Linguistics-Based Offensive Speech Detection. Computer Systems Science & Engineering, 47(2).

[47] Gongane, V. U., Munot, M. V., & Anuse, A. D. (2022). Detection and moderation of detrimental content on social media platforms: current status and future directions. Social Network Analysis and Mining, 12(1), 129.

[48] Libina, M., Sasipriya, G., & Rajasekar, V. (2023, April). An Automatic Method to Prevent and Classify Cyber Bullying Incidents Using Machine Learning Approach. In 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM) (pp. 1-7). IEEE.

[49] Tursynova, A., & Omarov, B. (2021, November). 3D U-Net for brain stroke lesion segmentation on ISLES 2018 dataset. In 2021 16th International Conference on Electronics Computer and Computation (ICECCO) (pp. 1-4). IEEE.

[50] Wadud, M. A. H., Kabir, M. M., Mridha, M. F., Ali, M. A., Hamid, M. A., & Monowar, M. M. (2022). How can we manage offensive text in social media-a text classification approach using LSTM-BOOST. International Journal of Information Management Data Insights, 2(2), 100095.

[51] Fazil, M., Khan, S., Albahlal, B. M., Alotaibi, R. M., Siddiqui, T., & Shah, M. A. (2023). Attentional multi-channel convolution with bidirectional LSTM cell toward hate speech prediction. IEEE Access, 11, 16801-16811.

[52] Kaya, S., & Alatas, B. (2022). A New Hybrid LSTM-RNN Deep Learning Based Racism, Xenomy, and Genderism Detection Model in Online Social Network. International Journal of Advanced Networking and Applications, 14(2), 5318-5328.

[53] Akhter, M. P., Jiangbin, Z., Naqvi, I. R., AbdelMajeed, M., & Zia, T. (2021). Abusive language detection from social media comments using conventional machine learning and deep learning approaches. Multimedia Systems, 1-16.

[54] Shannaq, F., Hammo, B., Faris, H., & Castillo-Valdivieso, P. A. (2022). Offensive language detection in Arabic social networks using evolutionary-based classifiers learned from fine-tuned embeddings. IEEE Access, 10, 75018-75039.

[55] Iqbal, A., Shahzad, K., Khan, S. A., & Chaudhry, M. S. (2023). The relationship of artificial intelligence (AI) with fake news detection (FND): a systematic literature review. Global Knowledge, Memory and Communication.

[56] Aurpa, T. T., Sadik, R., & Ahmed, M. S. (2022). Abusive Bangla comments detection on Facebook using transformer-based deep learning models. Social Network Analysis and Mining, 12(1), 24.

[57] Nath, N., George, J. P., Kesan, A., & Rodrigues, A. (2022). An Efficient Deep Learning-Based Hybrid Architecture for Hate Speech Detection in Social Media. In Data Science and Security: Proceedings of IDSCS 2022 (pp. 347-355). Singapore: Springer Nature Singapore.

[58] Abarna, S., Sheeba, J. I., & Devaneyan, S. P. A novel ensemble model for identification and classification of cyber harassment on social media platform. Journal of Intelligent & Fuzzy Systems, (Preprint), 1-24.

[59] Elzayady, H., Mohamed, M. S., Badran, K., & Salama, G. (2022, July). Improving Arabic hate speech identification using online machine learning and deep learning models. In Proceedings of Seventh International Congress on Information and Communication Technology: ICICT 2022, London, Volume 2 (pp. 533-541). Singapore: Springer Nature Singapore.

[60] Ghosh, T., Al Banna, M. H., Al Nahian, M. J., Uddin, M. N., Kaiser, M. S., & Mahmud, M. (2023). An attention-based hybrid architecture with explainability for depressive social media text detection in Bangla. Expert Systems with Applications, 213, 119007.

[61] Kumar, R., & Bhat, A. (2022). A study of machine learning-based models for detection, control, and mitigation of cyberbullying in online social media. International Journal of Information Security, 21(6), 1409-1431.

[62] Lin, H., Siarry, P., Gururaj, H. L., Rodrigues, J., & Jain, D. K. (2022). Special issue on deep learning methods for cyberbullying detection in multimodal social data. Multimedia Systems, 28(6), 1873-1875.

[63] Karwa, R. R., & Gupta, S. R. (2022). Automated hybrid Deep Neural Network model for fake news identification and classification in social networks. Journal of Integrated Science and Technology, 10(2), 110-119.

[64] Kumar, A., & Sachdeva, N. (2022). Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data. Multimedia systems, 28(6), 2027-2041.

[65] Abdelhakim, M., Liu, B., & Sun, C. (2023). Ar-PuFi: A Short-Text Dataset to Identify the Offensive Messages Towards Public Figures in the Arabian Community. Expert Systems with Applications, 120888.

[66] de Pablo, Á., Araque, O., & Iglesias, C. A. (2022). Transfer Learning with Social Media Content in the Ride-Hailing Domain by Using a Hybrid Machine Learning Architecture. Electronics, 11(2), 189.

[67] Shanmugavadivel, K., Sathishkumar, V. E., Raja, S., Lingaiah, T. B., Neelakandan, S., & Subramanian, M. (2022). Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed data. Scientific Reports, 12(1), 21557.

[68] Althobaiti, M. J. (2022). Bert-based approach to arabic hate speech and offensive language detection in twitter: Exploiting emojis and sentiment analysis. International Journal of Advanced Computer Science and Applications, 13(5).

[69] Nascimento, F. R., Cavalcanti, G. D., & Da Costa-Abreu, M. (2023). Exploring automatic hate speech detection on social media: a focus on content-based analysis. SAGE Open, 13(2), 21582440231181311.

[70] Del Valle-Cano, G., Quijano-Sánchez, L., Liberatore, F., & Gómez, J. (2023). SocialHaterBERT: A dichotomous approach for automatically detecting hate speech on Twitter through textual analysis and user profiles. Expert Systems with Applications, 216, 119446.