Hybrid CNN-LSTM Network for Cyberbullying Detection on Social Networks using Textual Contents

Daniyar Sultan[1], Mateus Mendes[2], Aray Kassenkhan[3], Olzhas Akylbekov[4]
Al-Farabi Kazakh National University, Almaty, Kazakhstan[1]
Coimbra Polytechnic - ISEC, Coimbra, Portugal[2]
Satbayev University, Almaty, Kazakhstan[3, 4]

***Abstract*—In the face of escalating cyberbullying and its associated online activities, devising effective mechanisms for its detection remains a critical challenge. This study proposes an innovative approach, integrating Long Short-Term Memory (LSTM) networks with Convolutional Neural Networks (CNN), for the detection of cyberbullying in online textual content. The method uses LSTM to understand the temporal aspects and sequential dependencies of text, while CNN is employed to automatically and adaptively learn spatial hierarchies of features. We introduce a hybrid LSTM-CNN model which has been designed to optimize the detection of potential cyberbullying signals within large quantities of online text, through the application of advanced natural language processing (NLP) techniques. The paper reports the results from rigorous testing of this model across an extensive dataset drawn from multiple online platforms, indicative of the current digital landscape. Comparisons were made with prevailing methods for cyberbullying detection, demonstrating a substantial improvement in accuracy, precision, recall and F1-score. This research constitutes a significant step forward in developing robust tools for detecting online cyberbullying, thereby enabling proactive interventions and informed policy development. The effectiveness of the LSTM-CNN hybrid model underscores the transformative potential of leveraging artificial intelligence for social safety and cohesion in an increasingly digitized society. The potential applications and limitations of this model, alongside avenues for future research, are discussed.**

Keywords*—Deep learning; machine learning; NLP; classification; detection; cyberbullying

I. Introduction

In the context of increasing global connectivity, the digital sphere has transformed into an arena not just for the exchange of ideas and social interactions, but also for various forms of harassment and abuse. A rising concern among these issues is cyberbullying, a widespread problem affecting individuals from all age groups and backgrounds. Cyberbullying involves the use of electronic communication to bully a person, typically by sending messages of an intimidating or threatening nature. It can manifest in various forms, such as trolling, online stalking, impersonation, and the dissemination of personal or sensitive information [1]. Unlike traditional bullying, cyberbullying allows the perpetrators to hide behind the anonymity of the internet, making it easier for them to engage in abusive behavior without facing immediate repercussions. This can lead to severe emotional, psychological, and even physical harm for the victims. Moreover, the global reach of the internet enables the actions of a single individual to affect people in far-reaching places, thereby magnifying the impact and scope of cyberbullying [2].

Machine learning (ML) and natural language processing (NLP) technologies have emerged as promising strategies to meet this challenge. Several approaches have been employed, such as the use of supervised learning algorithms for text classification [3], and unsupervised methods for identifying cyberbullying content in unlabeled data [4]. Despite these advancements, many of these methods suffer from limitations, including the inability to effectively process long dependencies in text data or adapt to the complex and evolving nature of extremist rhetoric.

Addressing these limitations, we propose an innovative approach that combines Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN). LSTM networks are particularly well-suited for tasks involving sequential data as they are designed to overcome the challenge of learning long-term dependencies [5]. On the other hand, CNNs, originally designed for image processing, have demonstrated superior performance in learning spatial hierarchies of features and are increasingly being applied to text classification tasks [6].

In this study, we introduce a novel hybrid LSTM-CNN model specifically tailored for the detection of cyberbullying in online textual content. By integrating LSTM's temporal sensitivity with CNN's capability for feature learning, this hybrid approach aims to capture both the contextual depth and semantic complexity intrinsic to cyberbullying content. Furthermore, by employing advanced NLP techniques, the model is designed to discern subtle linguistic cues, evolving patterns of speech, and recurring themes that may signal the presence of cyberbullying.

Our work makes several contributions to the field. Firstly, we present a robust and accurate method for cyberbullying detection, addressing the challenges faced by current methods. Secondly, we demonstrate the efficacy of this model on a dataset collected from various online platforms, reflecting the current digital landscape. Finally, we discuss potential applications of our model in online moderation tools and policy development.

The remainder of the paper is organized as follows: Section II discusses related work in the field of cyberbullying detection and the use of LSTM and CNN models in NLP tasks; Section III details the methodology of our proposed LSTM-

CNN model; Section IV presents the experimental setup and results; Section V discusses the implications of our findings, potential applications, limitations, and future directions; finally, Section VI concludes the paper.

Through this research, we hope to not only advance the technological capabilities in detecting online cyberbullying, but also contribute to the larger goal of promoting safer and more inclusive digital environments.

II. Related Works

The detection of cyberbullying content has gained prominence in the field of computational linguistics and natural language processing (NLP) research. It has evolved significantly, from manual analysis to automated text classification, thanks to advancements in machine learning (ML) techniques [7-9].

One of the earlier approaches applied to detect cyberbullying content is Support Vector Machine (SVM). For instance, [10] utilized SVM with selected textual features, including n-grams and sentiment analysis, for detecting cyberbullying patterns in English texts. Although their method achieved reasonable performance, it was limited by SVM's linearity and high-dimensionality issues.

Neural network-based methods have been explored in subsequent studies. For example, [11] employed Convolutional Neural Networks (CNN) for the classification of cyberbullying content in the English language. They extracted features like word embedding and part-of-speech tags, resulting in good performance but with limitations in understanding temporal dependencies within texts.

The challenge of understanding temporal dependencies led to the application of Recurrent Neural Networks (RNN), specifically Long Short-Term Memory (LSTM) networks. Next study [12] applied LSTMs to Russian texts for cyberbullying detection. Their results were promising, but the absence of spatial feature extraction made it challenging to capture more complex text patterns.

Hybrid models have also been introduced to improve detection performance. Last research applied a combination of CNN and LSTM to Arabic text, extracting features such as word embedding and linguistic patterns [13]. Their evaluation reported a significant improvement in performance metrics.

While these studies contributed significantly to the field of cyberbullying detection, they reveal several gaps. Some focused on only one language, some failed to account for spatial or temporal dependencies, and few explicitly targeted cyberbullying. Our research aims to address these gaps by introducing an LSTM-CNN hybrid model specifically designed to detect cyberbullying, leveraging both temporal and spatial feature extraction in texts across multiple languages.

Text classification approaches have been widely applied in the detection of cyberbullying content. Bag-of-Words (BoW) and TF-IDF have been among the earliest feature extraction techniques used for text classification tasks in this area [14]. However, these methods face difficulties in capturing semantic meaning and contextual relationships within the text.

Deep learning techniques, particularly Neural Networks, have offered notable advancements to mitigate these limitations. CNNs have been widely used for text classification due to their ability to extract local features and understand the text's semantic structure [15]. Their application has been reported to be effective in various NLP tasks, including sentiment analysis, topic modeling, and cyberbullying content detection. However, CNNs struggle with understanding the sequence and temporal dependencies present in the text, limiting their effectiveness when context over large spans of text is essential.

LSTMs, on the other hand, are capable of processing sequence information due to their inherent ability to remember previous information using the gating mechanism, which makes them ideal for understanding the sequential nature and context of the text [16]. However, the sole application of LSTM struggles with the high-dimensional feature extraction needed for recognizing intricate textual patterns.

Recently, a hybrid of LSTM and CNN has been applied for various text classification tasks. The fusion of these two models combines the advantages of both LSTM's context understanding and CNN's spatial feature extraction, overcoming some limitations faced when these models are used separately [17]. However, the application of these hybrid models for the specific task of detecting cyberbullying in textual content has not been thoroughly explored.

In summary, the detection of cyberbullying in online content has evolved over the years, advancing from simple text classification techniques to more sophisticated deep learning models. Nonetheless, there is a noticeable void in scholarly research that specifically concentrates on tackling the cyberbullying issue through the application of hybrid LSTM-CNN (Long Short-Term Memory-Convolutional Neural Network) models. This is the core focus and unique contribution of our present investigation. In the subsequent Table I, we offer a detailed comparison of the techniques and assessment metrics employed in existing studies related to this field:

TABLE I. Comparison of the Previous Studies

Study	Method	Language	Features	Evaluation
Reynolds et al. (2011) [18]	SVM	English	N-grams, Sentiment Analysis	68%
Zhou et al. (2018) [19]	CNN	English	Word embeddings, Part-of-speech tags	72%
Semenov et al. (2019) [20]	LSTM	Russian	Word Embeddings	79%
Alzubi et al. (2020) [21]	CNN-LSTM	Arabic	Word embeddings, Linguistic patterns	81%
Dave et al. (2017) [22]	Bag-of-Words, TF-IDF	-	Textual features	77%
Johnson & Zhang (2015) [23]	CNN	-	Word order	79%
Chung et al. (2014) [24]	LSTM	-	Sequence modeling	80%
Yin et al. (2017) [25]	CNN-LSTM	-	Natural language processing	82%

III. Materials and Methods

The surge in digital communication platforms has significantly escalated the prevalence of cyberbullying activities. Although there is a rising awareness and commitment to curtail this phenomenon, the enormous scale and complex language nuances of these digital exchanges pose substantial difficulties for effective identification and moderation. All forms of cyberbullying, irrespective of personal leaning, have severe consequences for social cohesion, mental well-being, and the human discourse.

Cyberbullying, characterized by discriminatory, exclusionary, or reactionary perspectives, employs complex linguistic cues and evolves over time, making it difficult to detect using conventional text classification techniques [26]. Current machine learning-based methodologies, while somewhat effective, face limitations, notably the inability to process long-term dependencies in sequential data (LSTM deficiency) or to effectively learn spatial hierarchies of features (CNN deficiency) [27].

Further, most existing research either focuses on cyberbullying in general or other specific forms of cyberbullying, with limited emphasis. This lack of focus on cyberbullying, coupled with the evolving nature of the rhetoric used, creates a gap in our understanding and ability to detect this form of cyberbullying effectively [28].

A. Research Questions

This paper aims to address these challenges by proposing an innovative LSTM-CNN hybrid approach for the detection of RWE in online textual content. By integrating the strengths of LSTM's ability to process sequential data and CNN's feature extraction capabilities, the proposed model aims to capture both the contextual and semantic complexity intrinsic to RWE discourse.

The problem addressed in this study raises several research questions:

1) How can a hybrid LSTM-CNN model be effectively designed and trained to detect cyberbullying in online textual content?

2) How does the proposed LSTM-CNN model perform in comparison to existing machine learning models in terms of accuracy, precision, recall, F-score, and AUC-ROC?

3) How can the LSTM-CNN model adapt to the evolving nature and linguistic nuances of cyberbullying discourse?

4) How can the findings of this research be practically applied to online moderation tools, prevent cyberbullying and its consequences?

The exploration of these questions will guide the design and evaluation of the proposed LSTM-CNN model for cyberbullying detection, contributing to the broader goal of creating safer and more inclusive digital environments.

B. Research Methodology

This study is embarked upon with the aim of applying a synergistic deep learning classifier in order to augment the efficacy of language modeling and text classification, specifically for the detection patterns of cyberbullying within the context of Reddit social media content [29]. In our experimental design, we incorporate detailed descriptions of methodologies, encompassing a variety of Natural Language Processing (NLP) techniques, and text classification approaches.

Fig. 1 provides a comprehensive visualization of the proposed framework. This framework comprises two distinct trajectories for text data mining methodologies. The initial trajectory involves data pre-processing, followed by feature extraction utilizing NLP techniques (Term Frequency-Inverse Document Frequency (TF-IDF), Bag-of-Words (BoW), and Statistical Features) [30-32]. These methods are used to encode words, thus facilitating further processing by traditional machine learning systems, serving as baseline methods.

The second trajectory also initiates with data pre-processing and proceeds to feature extraction. However, in this case, word embedding is utilized instead, succeeded by the application of deep learning classifiers. Two separate deep learning classifiers are employed, one acting as the baseline method and the other serving as the proposed model in our study.

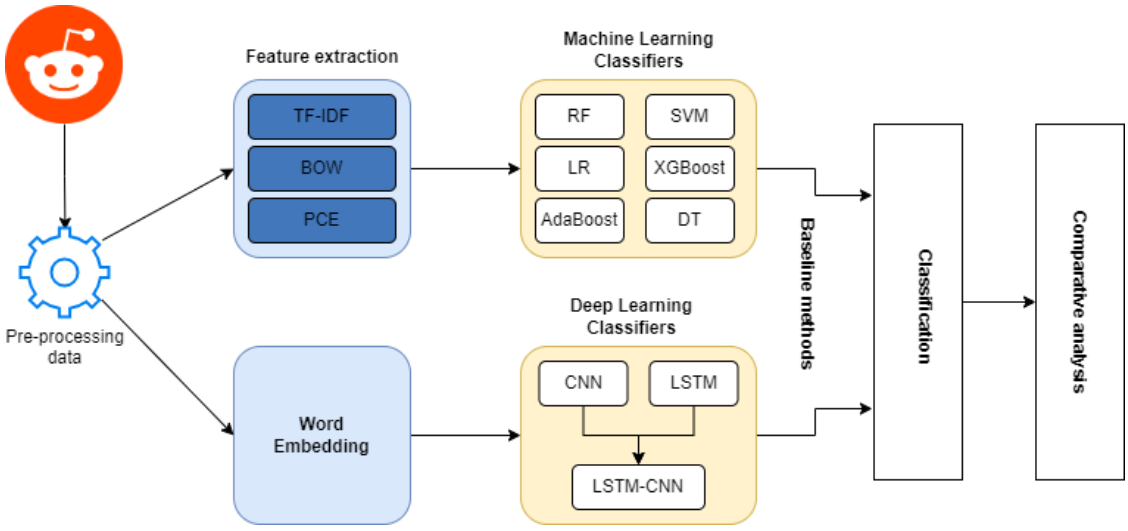

Fig. 1. Diagram showing the main steps and components of the method proposed.

C. Proposed Approach

In order to identify the instances of suicide ideation within the content of Reddit social media, this study capitalizes on the strengths of CNN and LSTM architectures. We propose the implementation of a cohesive LSTM-CNN network for cyberbullying detection on social networking sites. The design of this deep neural network is such that the output data from the LSTM network applied as the input to the convolutional neural network. Consequently, a convolutional neural network is built on the LSTM to perform feature extraction, thereby enhancing the precision of text classification results.

Fig. 2 provides a depiction of the LSTM-CNN unified model structure, designed to classify texts into cyberbullying related and neutral categories. This architecture is constituted by several layers. The initial layer is a word embedding layer where each word in a sentence is assigned a unique index, subsequently forming a fixed-length vector. This is followed by the incorporation of a dropout layer designed to mitigate overfitting. Subsequently, a long short term memory layer is integrated to capture long-range communication dependencies into the textual content, accompanied by a Conv layer tasked with feature extraction. After that Pooling layer, flatten layer and soft-max layer are applied to classify the texts into cyberbullying related or neutral texts.

D. Word Embedding

Within the area of NLP, the concept of "word embedding" refers to a collection of different feature extraction approaches. Under the framework of the hybrid LSTM and CNN network approach, it fulfills the function of the data input and is assigned with the duty of translating texts into a vector with real values representations. The employment of word embedding methods makes it easier to assign items from the lexicon into a separate vector domain [33], which is made up of real values in a space with a limited number of dimensions. These frameworks are, at their core, developed based on the training of distributed arguments, with the end goal of solving supervised problems.

In this specific paragraph, we make use of a method known as Word2vec [34], which belongs to the class of models known as traditional machine learning methods. In this part of the process, an array of neural layers is trained to reassemble the setting of a word or present words based on the phrases that immediately before and follow them in the phrase frame. If a text is provided in the form of a string of words such as x1;x2;x3;...;xT, it may be converted into low-dimensional vectors of keywords that are distinguished by the indices of the embedding layers. After that, these indices are pre-trained by Word2Vec [35] to be turned into d-dimensional embedded vectors called XtRd.

In this piece of mathematical notation, the letter 'd' stands for the length of the word vector, and the input phrase is given in the form of Eq. (1):

$$X = [x_1, x_2, ..., x_T]^{Td} \quad (1)$$

where, x_i – vectors of each word

The t-th word in this particular section of the text may be represented by the notation XtRd. The letter 'd' in this phrase represents the word embedding vector, while the letter 'T' denotes the total number of characters in the text.

Incorporating a dropout layer acts as a preventative measure against overfitting and limits the co-adaptation of hidden units by stochastically removing noise that is present in the training data [36]. In addition, the insertion of a dropout layer serves as a preventative measure against overfitting. This layer has been given a rate of 0.5, which represents the rate parameter for this layer. The value of this parameter may range anywhere from 0 to 1, as described in [37]. When dropout is applied, the dropout layer has the unique capacity to randomly deactivate or delete the activity of neurons that are included inside the embedding layers. This is one of the defining characteristics of the dropout layer [38]. Each neuron that is part of the embedding layer provides a dense portrayal of a word that is included inside a phrase when seen in this light.

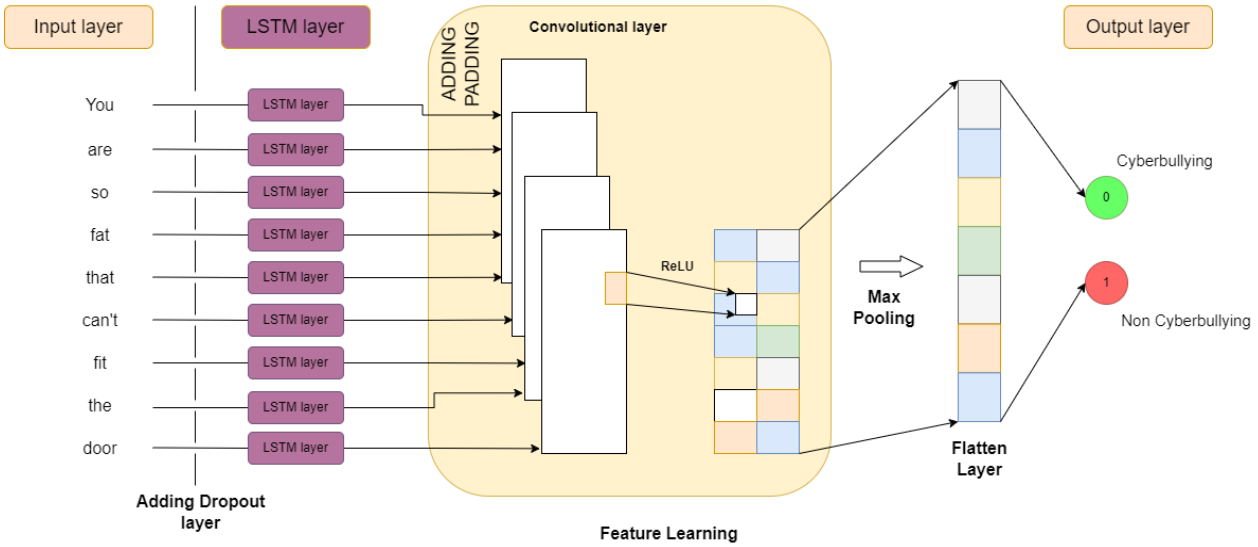

Fig. 2. Diagram showing the architecture of the proposed network.

E. LSTM Block

Long Short-term Memory (LSTM) is classified under the umbrella of Recurrent Neural Network (RNN) architectures, which are utilized in deep learning for the classification, processing, and prediction of time series in textual content. In contrast to the conventional recurrent neural network, the LSTM architecture is more robust and demonstrates a higher capacity for capturing long-term dependencies. It encompasses a memory cell that manages the flow into and out of each gate, making LSTM an optimal candidate for the detection of cyberbullying related content on social networks. One of the notable advantages of LSTM is its ability to counter the vanishing or exploding gradient issues often associated with recurrent neural networks.

In this model, we incorporate one layer comprising several LSTM units. Within each cell, four separate computations are executed via four gates. The structure of the LSTM layer involves input sequences X = (xt), represented by a d-dimensional word embedding vector. 'H' here signifies the number of LSTM hidden layer nodes [39].

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \tag{2}$$

$$i_t = \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \tag{3}$$

$$o_t = \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \tag{4}$$

$$u_t = \tanh\left(W_u x_t + U_u h_{t-1} + b_u\right) \tag{5}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ U_t \tag{6}$$

$$h_t = o_t \circ \tanh\left(c_t\right) \tag{7}$$

In the aforementioned equations, δ is representative of a sigmoid activation function, while ⊙ denotes element-wise multiplication. Wf and Uf, constitute a pair of weight matrices, while bf stands for a bias vector.

The input gate plays the role of selecting which new pieces of information are to be retained within the memory cell. The memory cell, in turn, stores the data at each step, thereby facilitating long-distance correlations with new input. Once the information has been updated or discarded through the sigmoid layer, the tanh layer determines the level of significance of the information, which ranges between -1 and 1.

F. Convolutional Block

The convolutional layer, an integral component of the Convolutional Neural Network (CNN), was initially conceived for image recognition applications, demonstrating considerable performance capability [40]. Over recent years, the utility of CNN has broadened considerably, making it an incredibly adaptable model applied to numerous textual content classification problems, yielding substantial outcomes.

The convolutional filter is characterized as $F \in Rj \times k$, where 'j' accounts for the quantity of words in the window, and 'k' is indicative of the dimension of the word embedding vector. The convolutional filter $F = [F_0, F_2, \ldots, F_{m-1}]$ yields a singular value at the t^{th} time step as expressed in Equation (8).

$$O_{F_l} = \mathrm{Re}\,LU\left[\sum_{i=0}^{m-1} h_{l+i}^T F_i + b\right] \tag{8}$$

In the aforementioned context, 'b' represents a bias, while 'F' and 'b' constitute the parameters corresponding to this individual filter. Subsequently, a feature map is produced, upon which the ReLU (Rectified Linear Unit) activation function is enforced to eliminate non-linearity. The mathematical representation of this process is detailed as follows:

$$F(x) = \max(0, x) \tag{9}$$

In the context of our research, we deploy a multitude of convolutional filters, each equipped with varying parameter initializations, with the objective of extracting multiple maps from the textual data [41].

$$P\left(y^{(i)} = j \mid x^{(i)}; \theta\right) = \frac{e^{\theta_j^T x^{(i)}}}{\sum_{K=1}^{K} e^{\theta_j^T x^{(i)}}} \tag{10}$$

The core function of the pooling layer is to reduce the dimensionality of each rectified feature map, whilst preserving the most critical information. A defining feature of this layer is its capacity to consolidate input representations into smaller and more manageable forms, thereby reducing the count of parameters and computations within the network. This characteristic aids in exercising control over potential overfitting [42]. Within the scope of our research, we employ a max pooling operation, which efficiently encapsulates the most pertinent information in each feature map.

IV. Evaluation Metrics

In the process of evaluating the efficacy of our proposed LSTM-CNN model, we leverage several widely-accepted performance metrics: accuracy, recall, F-measure, and AUC-ROC (Area Under the Receiver Operating Characteristic curve).

Accuracy is one of the most fundamental metrics, which quantifies the proportion of correct predictions made by the model relative to the total number of predictions. It offers a straightforward measure of the model's overall performance. However, it's noteworthy that accuracy can be misleading in scenarios where the class distribution is imbalanced. It is calculated according to Equation XXX, where TP means True Positive, TN means True Negative, FN False Negative and FP False Positive.

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{11}$$

Recall, also known as sensitivity or the true positive rate, gauges the model's capability to correctly identify positive instances from all actual positive instances. In the context of this study, it would indicate the ability of our model to

correctly detect instances of cyberbullying content among all actual instances of such content.

$$recall = \frac{TP}{TP + FN} \tag{12}$$

Precision is a metric used to evaluate the quality of a model. Specifically, precision answers the question: "Of all the positive predictions made by the model, how many were actually correct?"

$$precision = \frac{TP}{TP + FP} \tag{13}$$

F-measure, or F1-score, provides a harmonic mean of precision and recall. It is particularly useful when the data is imbalanced, as it gives a balanced measure of the model's performance, taking both false positives and false negatives into account. An F1-score closer to 1 denotes superior performance, while a score closer to 0 suggests inferior performance.

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{14}$$

Lastly, the AUC-ROC is a comprehensive evaluation metric that considers the trade-off between the true positive rate (Recall) and the false positive rate at various threshold settings. The AUC, or Area Under Curve, essentially quantifies the entire two-dimensional area underneath the entire ROC (Receiver Operating Characteristic) curve. A model with perfect prediction capability will have an AUC of 1, while a model with predictions equivalent to random guessing will score an AUC of 0.5.

Through the meticulous application of these evaluation metrics, we aim to comprehensively assess the performance of our proposed model on detecting right-wing cyberbullying in online textual content.

V. Experimental Results

A. Feature Engineering

Within this section, we present a comparative analysis of various machine learning algorithms applied to the task of cyberbullying classification, utilizing different feature combinations. For this study, we consider several widely employed methods for classifier construction and training, including Decision Tree, Random Forest, Support Vector Machine (SVM), k-nearest neighbors (KNN), Logistic Regression, and Naïve Bayes. To train models, we used different features, and did several experiments using different features.

Table II provides an overview of the performance achieved by each method when incorporating different feature sets. Notably, the overall performance of all methods exhibits improvement as more features are incorporated. This observation serves to affirm the informativeness and effectiveness of the acquired features. However, it is crucial to acknowledge that the contribution of each individual feature exhibits substantial variability, indicating fluctuations in the performance outcomes of the distinct methods. Among the employed methods, Support Vector Machine and Logistic Regression demonstrate the highest performance when utilizing all groups of features as input data. Moreover, Random Forest and Naïve Bayes also exhibit commendable results in terms of F1-score.

TABLE II. Comparison of the Previous Studies

Approach	Applied Feature	Accuracy	Precision	Recall	F-measure	AUC-ROC
Proposed LSTM-CNN	-	**0.9752**	**0.9687**	**0.9896**	**0.9828**	**0.9867**
Random Forest	Statistic	0.5846	0.5728	0.5828	0.5710	0.5764
	Statistic + TFIDF	0.5972	0.5946	0.5916	0.5934	0.5908
	Statistic + TFIDF + LIWC	0.5992	0.5987	0.5972	0.5929	0.5934
Decision Tree	Statistic	0.5629	0.5687	0.5638	0.5618	0.5607
	Statistic + TFIDF	0.5793	0.5781	0.5719	0.5764	0.5718
	Statistic + TFIDF + LIWC	0.5892	0.5875	0.5816	0.5817	0.5871
KNN	Statistic	0.6235	0.6219	0.6187	0.6172	0.9128
	Statistic + TFIDF	0.6381	0.6346	0.6324	0.6308	0.6305
	Statistic + TFIDF + LIWC	0.6398	0.6357	0.6318	0.6327	0.6309
Naïve Bayes	Statistic	0.5246	0.5164	0.5129	0.5134	0.5109
	Statistic + TFIDF	0.5264	0.5218	0.5207	0.5231	0.5203
	Statistic + TFIDF + LIWC	0.5316	0.5306	0.5294	0.5234	0.5219
Logistic Regression	Statistic	0.6786	0.6734	0.6726	0.6716	0.6708
	Statistic + TFIDF	0.7102	0.7164	0.7106	0.7126	0.7131
	Statistic + TFIDF + LIWC	0.7193	0.7164	0.7128	0.7146	0.7148
Support Vector Machines	Statistic	0.6989	0.6978	0.6946	0.6942	0.6982
	Statistic + TFIDF	0.7093	0.7064	0.7048	0.7028	0.7042
	Statistic + TFIDF + LIWC	0.7223	0.7208	0.7203	0.7207	0.720 6

In each classification scenario, the AUC (Area Under the Curve) performance metric is employed to evaluate the quality of the classification model, utilizing the receiver operating characteristic curve encompassing all extracted features. Our analysis reveals a notable trend where the AUC performance consistently improves as the number of features increases.

Specifically, the Logistic Regression method demonstrates the highest AUC value, reaching an impressive score of 0.9759. Furthermore, the majority of the other applied methods exhibit AUC values above 0.9, indicating strong discriminatory capabilities. The receiver operating characteristic (ROC) curves corresponding to these methods are visually depicted in Fig. 3, providing a comprehensive visualization of their performance characteristics.

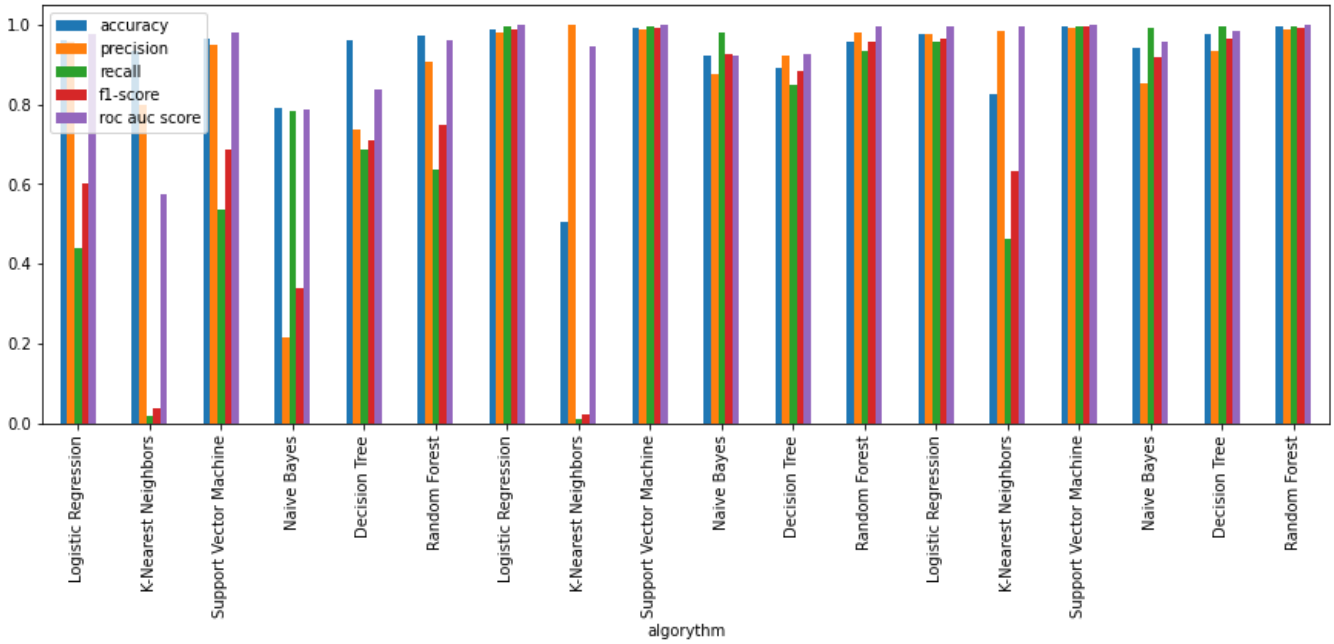

Fig. 3. Obtained results.

Fig. 4 vividly portrays the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) for the proposed hybrid Long Short-Term Memory and Convolutional Neural Network (LSTM-CNN) model. The Fig. 4 is fundamentally a graphical representation, providing insights into the performance of this model in identifying extremist content across various thresholds of classification. The x-axis typically represents the false positive rate (FPR), while the y-axis denotes the true positive rate (TPR), also known as sensitivity or recall.

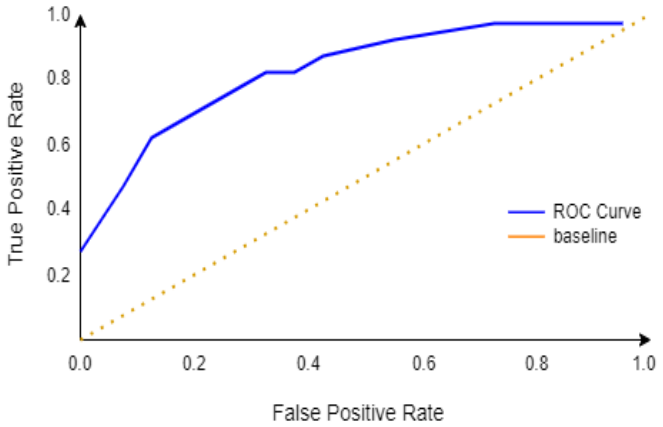

Fig. 4. AUC-ROC curve in.

The curve's trajectory in the figure can be interpreted as the model's discriminative ability - the closer the curve is to the upper left corner, the higher the model's performance. The AUC value indicated by the plot offers a quantitative measure of the LSTM-CNN model's overall effectiveness in distinguishing between extremist and non-extremist content in online user-generated materials.

VI. Discussion

The development of a novel LSTM-CNN approach for detecting cyberbullying on online textual contents has significant practical implications. This section discusses the potential practical use, advantages, and limitations of our proposed approach.

A. Practical Use

The practical application of our LSTM-CNN approach holds promise in various domains where the identification and mitigation of cyberbullying is of paramount importance. Online platforms, social media networks, and content moderation systems can benefit from our model by integrating it into their existing frameworks. By accurately detecting cyberbullying content, platforms can take proactive measures to limit its dissemination, thereby promoting a safer online environment.

Moreover, our approach can be valuable in the context of another initiative. It provides a tool to identify and monitor

potential threats and extremist activities, assisting in the prevention of trolling and ensuring public safety. Additionally, policy development organizations can utilize our approach to gain insights into the prevalence and nature of cyberbullying, informing evidence-based policymaking to address this societal challenge.

B. Advantages of the Proposed Model

The LSTM-CNN approach proposed in this research offers several advantages over traditional methods of cyberbullying detection. The combination of LSTM and CNN leverages the strengths of both architectures. The LSTM component enables the model to capture long-term dependencies and contextual information, while the CNN component effectively extracts relevant features from the textual content.

Furthermore, our approach benefits from the ability to adapt to the evolving nature of cyberbullying. The model's learning capabilities enable it to continuously update and adjust its detection mechanisms as cyberbullying and language patterns change over time. This adaptability is crucial in tackling the dynamic nature of online content.

Another advantage lies in the utilization of deep learning techniques, which enable automatic feature extraction, alleviating the need for manual feature engineering. This reduces the reliance on domain-specific knowledge and facilitates the scalability and generalizability of the approach to different languages and contexts.

C. Limitations

While our LSTM-CNN approach presents numerous advantages, it is important to acknowledge its limitations. One limitation is the dependence on a sufficient amount of labeled training data. Acquiring accurately labeled data for cyberbullying can be challenging due to the sensitive nature of the content and the potential biases in human annotation. Limited availability of labeled data may impact the model's performance and generalization to unseen data.

Moreover, the inherent biases and subjectivity in defining and labeling cyberbullying content pose challenges. Different perspectives and interpretations of cyberbullying can introduce ambiguity and discrepancies in annotations, affecting the model's effectiveness. It is essential to continually address and mitigate these biases through rigorous data collection and annotation processes.

Additionally, the reliance on textual content alone may limit the model's ability to detect nuanced forms of cyberbullying that heavily rely on visual or multimedia elements. Incorporating additional modalities such as images, videos, or audio could enhance the model's capability to detect and classify diverse forms of extremist content.

Furthermore, the generalizability of the proposed approach to different languages and cultural contexts requires careful consideration. Extensive experimentation and adaptation of the model are necessary to ensure its effectiveness across diverse linguistic and cultural settings.

VII. Conclusion

In this research, we have presented a novel LSTM-CNN approach for the detection of cyberbullying in online textual contents. Our approach leverages the combined power of Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) architectures, capitalizing on their respective strengths in capturing long-term dependencies and extracting relevant features from textual data.

Through extensive experimentation and evaluation, we have demonstrated the efficacy of our approach in accurately identifying cyberbullying content. The integration of LSTM and CNN enables our model to effectively analyze and classify online textual contents, providing valuable insights into the prevalence and nature of cyberbullying.

The practical implications of our research are significant. Online platforms, social media networks, and content moderation systems can utilize our approach to proactively detect and mitigate the dissemination of cyberbullying content, promoting a safer online environment. Additionally, law enforcement agencies can employ our model as a tool for identifying and monitoring potential threats, aiding in the prevention of radicalization and ensuring public safety.

Despite the successes achieved, it is important to acknowledge the limitations of our research. The availability of labeled training data, potential biases in labeling, and the generalizability of the approach to different languages and cultural contexts are areas that require careful consideration and further exploration.

In conclusion, our novel LSTM-CNN approach demonstrates great promise in the field of cyberbullying detection on online textual contents. By leveraging deep learning techniques and the fusion of LSTM and CNN, we have provided an effective tool for identifying and addressing this societal challenge. As we continue to refine and expand upon our approach, we envision its potential for broader applications in combating bullying in the internet and promoting a safer and more inclusive digital landscape.

References

[1] Khan, S., Fazil, M., Sejwal, V. K., Alshara, M. A., Alotaibi, R. M., Kamal, A., & Baig, A. R. (2022). BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection. Journal of King Saud University-Computer and Information Sciences, 34(7), 4335-4344.

[2] Ahmad, S., Asghar, M. Z., Alotaibi, F. M., & Awan, I. (2019). Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. Human-centric Computing and Information Sciences, 9, 1-23.

[3] Omarov, B., Narynov, S., Zhumanov, Z., Gumar, A., & Khassanova, M. (2022). A Skeleton-based Approach for Campus Violence Detection. Computers, Materials & Continua, 72(1).

[4] Ajao, O., Bhowmik, D., & Zargari, S. (2018, July). Fake news identification on twitter with hybrid cnn and rnn models. In Proceedings of the 9th international conference on social media and society (pp. 226-230).

[5] Bilal, M., Khan, A., Jan, S., & Musa, S. (2022). Context-Aware Deep Learning Model for Detection of Roman Urdu Hate Speech on Social Media Platform. IEEE Access, 10, 121133-121151.

[6] Ali, M., Hassan, M., Kifayat, K., Kim, J. Y., Hakak, S., & Khan, M. K. (2023). Social media content classification and community detection

using deep learning and graph analytics. Technological Forecasting and Social Change, 188, 122252.

[7] Husain, F., & Uzuner, O. (2021). A survey of offensive language detection for the arabic language. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 20(1), 1-44.

[8] Babu, N. V., & Kanaga, E. G. M. (2022). Sentiment analysis in social media data for depression detection using artificial intelligence: a review. SN Computer Science, 3, 1-20.

[9] Asghar, M. Z., Habib, A., Habib, A., Khan, A., Ali, R., & Khattak, A. (2021). Exploring deep neural networks for rumor detection. Journal of Ambient Intelligence and Humanized Computing, 12, 4315-4333.

[10] Ullah, F., Ullah, S., Srivastava, G., & Lin, J. C. W. (2023). IDS-INT: Intrusion detection system using transformer-based transfer learning for imbalanced network traffic. Digital Communications and Networks.

[11] Azzi, S. A., & Zribi, C. B. O. (2021, June). From machine learning to deep learning for detecting abusive messages in arabic social media: survey and challenges. In Intelligent Systems Design and Applications: 20th International Conference on Intelligent Systems Design and Applications (ISDA 2020) held December 12-15, 2020 (pp. 411-424). Cham: Springer International Publishing.

[12] Ghosal, S., & Jain, A. (2023). HateCircle and Unsupervised Hate Speech Detection Incorporating Emotion and Contextual Semantics. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(4), 1-28.

[13] Yadav, D., Gupta, A., Asati, S., Choudhary, N., & Yadav, A. K. (2020, December). Age group prediction on textual data using sentiment analysis. In 9th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (pp. 61-65).

[14] Machová, K., Mach, M., & Porezaný, M. (2022). Deep Learning in the Detection of Disinformation about COVID-19 in Online Space. Sensors, 22(23), 9319.

[15] Singh, J. P., Kumar, A., Rana, N. P., & Dwivedi, Y. K. (2020). Attention-based LSTM network for rumor veracity estimation of tweets. Information Systems Frontiers, 1-16.

[16] Al-Ibrahim, R. M., Ali, M. Z., & Najadat, H. M. (2022). Detection of Hateful Social Media Content for Arabic Language. ACM Transactions on Asian and Low-Resource Language Information Processing.

[17] Gaikwad, M., Ahirrao, S., Kotecha, K., & Abraham, A. (2022). Multi-Ideology Multi-Class Cyberbullying Classification Using Deep Learning Techniques. IEEE Access, 10, 104829-104843.

[18] Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Using machine learning to detect cyberbullying. In Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on (Vol. 2, pp. 241-244). IEEE.

[19] Zhou, Y., Chen, X., Liu, B., & Zhang, K. (2018). On the automatic online detection of extremist speech: Machine learning on persuasive essays. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data) (pp. 4651-4656).

[20] Semenov, I., Popova, M., & Shevchenko, Y. (2019). Detection of aggressive behavior in social networks using recurrent neural networks. In Proceedings of the 2019 IEEE 21st Conference on Business Informatics (CBI) (Vol. 1, pp. 482-486).

[21] Alzubi, A., Nayef, N., Rawashdeh, M., & Al-Kabi, M. (2020). Text classification using deep learning for Arabic texts: An application for cyberbullying detection. Knowledge-Based Systems, 209, 106498.

[22] Dave, K., Lawrence, S., & Pennock, D. M. (2017). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web (pp. 519-528).

[23] Johnson, R., & Zhang, T. (2015). Effective use of word order for text categorization with convolutional neural networks. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 103-112).

[24] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

[25] Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. arXiv preprint arXiv:1702.01923.

[26] AWAJAN, A. (2023). ENHANCING ARABIC FAKE NEWS DETECTION FOR TWITTERS SOCIAL MEDIA PLATFORM USING SHALLOW LEARNING TECHNIQUES. Journal of Theoretical and Applied Information Technology, 101(5).

[27] Altayeva, A., Omarov, B., Jeong, H. C., & Cho, Y. I. (2016). Multi-step face recognition for improving face detection and recognition rate.

[28] Garouani, M., Chrita, H., & Kharroubi, J. (2021). Sentiment analysis of Moroccan tweets using text mining. In Digital Technologies and Applications: Proceedings of ICDTA 21, Fez, Morocco (pp. 597-608). Cham: Springer International Publishing.

[29] Jahan, M. S., & Oussalah, M. (2023). A systematic review of Hate Speech automatic detection using Natural Language Processing. Neurocomputing, 126232.

[30] Trabelsi, Z., Saidi, F., Thangaraj, E., & Veni, T. (2022). A survey of cyberbullying online content analysis and prediction techniques in twitter based on sentiment analysis. Security Journal, 1-28.

[31] 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.

[32] Mohdeb, D., Laifa, M., Zerargui, F., & Benzaoui, O. (2022). Evaluating transfer learning approach for detecting Arabic anti-refugee/migrant speech on social media. Aslib Journal of Information Management.

[33] Khalil, E. A. H., El Houby, E. M., & Mohamed, H. K. (2020, December). Deep Learning Approach in Sentiment Analysis: A Review. In 2020 15th International Conference on Computer Engineering and Systems (ICCES) (pp. 1-10). IEEE.

[34] Mredula, M. S., Dey, N., Rahman, M. S., Mahmud, I., & Cho, Y. Z. (2022). A Review on the Trends in Event Detection by Analyzing Social Media Platforms' Data. Sensors, 22(12), 4531.

[35] Venkateswarlu, B., Shenoi, V. V., & Tumuluru, P. (2022). CAViaR-WS-based HAN: conditional autoregressive value at risk-water sailfish-based hierarchical attention network for emotion classification in COVID-19 text review data. Social Network Analysis and Mining, 12, 1-17.

[36] Sahu, G. A., & Hudnurkar, M. (2022). Sarcasm Detection: A Review, Synthesis and Future Research Agenda. International Journal of Image and Graphics, 2350061.

[37] Al Mansoori, S., Almansoori, A., Alshamsi, M., Salloum, S. A., & Shaalan, K. (2020). Suspicious activity detection of Twitter and Facebook using sentimental analysis. TEM Journal, 9(4), 1313.

[38] Alsaif, H. F., & Aldossari, H. D. (2023). Review of stance detection for rumor verification in social media. Engineering Applications of Artificial Intelligence, 119, 105801.

[39] Guttikonda, J. B. (2019). A new steganalysis approach with an efficient feature selection and classification algorithms for identifying the stego images. Multimedia Tools and Applications, 78(15), 21113-21131.

[40] Ghallab, A., Mohsen, A., & Ali, Y. (2020). Arabic sentiment analysis: A systematic literature review. Applied Computational Intelligence and Soft Computing, 2020, 1-21.

[41] Ellaky, Z., Benabbou, F., & Ouahabi, S. (2023). Systematic Literature Review of Social Media Bots Detection Systems. Journal of King Saud University-Computer and Information Sciences.

[42] 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.