Computers, Materials & Continua
DOI: 10.32604/cmc.2023.032993
Article

Tech Science Press

Cyberbullying-related Hate Speech Detection Using Shallow-to-deep Learning

Daniyar Sultan[1,2], Aigerim Toktarova[3,*], Ainur Zhumadillayeva[4], Sapargali Aldeshov[5,6], Shynar Mussiraliyeva[1], Gulbakhram Beissenova[6,7], Abay Tursynbayev[8], Gulmira Baenova[4] and Aigul Imanbayeva[6]

[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan
[2]International Information Technology University, Almaty, Kazakhstan
[3]Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan
[4]L.N.Gumilyov Eurasian National University, Astana, Kazakhstan
[5]South Kazakhstan State Pedagogical University, Shymkent, Kazakhstan
[6]M. Auezov South Kazakhstan University, Shymkent, Kazakhstan
[7]University of Friendship of People's Academician A. Kuatbekov, Shymkent, Kazakhstan
[8]National Academy of Education named after Y. Altynsarin, Astana, Kazakhstan
*Corresponding Author: Aigerim Toktarova. Email: aikerimtoktarova@gmail.com
Received: 03 June 2022; Accepted: 12 July 2022

Abstract: Communication in society had developed within cultural and geographical boundaries prior to the invention of digital technology. The latest advancements in communication technology have significantly surpassed the conventional constraints for communication with regards to time and location. These new platforms have ushered in a new age of user-generated content, online chats, social network and comprehensive data on individual behavior. However, the abuse of communication software such as social media websites, online communities, and chats has resulted in a new kind of online hostility and aggressive actions. Due to widespread use of the social networking platforms and technological gadgets, conventional bullying has migrated from physical form to online, where it is termed as Cyberbullying. However, recently the digital technologies as machine learning and deep learning have been showing their efficiency in identifying linguistic patterns used by cyberbullies and cyberbullying detection problem. In this research paper, we aimed to evaluate shallow machine learning and deep learning methods in cyberbullying detection problem. We deployed three deep and six shallow learning algorithms for cyberbullying detection problems. The results show that bidirectional long-short-term memory is the most efficient method for cyberbullying detection, in terms of accuracy and recall.

Keywords: Cyberbullying; machine learning; deep learning; classification; NLP

1 Introduction

Threats, online harassment, disgrace, fear, and other forms of cyberbullying are characterized as new forms of violence or bullying that are perpetrated through technical gadgets and the World Wide Web [1,2]. With the development of the digital technologies and social media, the internet has become a global means of communication, allowing individuals from all walks of life to express themselves. As a result, the internet has amassed a vast quantity of data and has evolved into a strong source of information.

Cyberbullying has grown in popularity as a result of technical advancement and increased accessibility. In addition to the emotional, social, and intellectual consequences of cyberbullying and online harassment, many victims of cyberbullying have attempted suicide as a result of their psychological trauma [3]. Cyberbullying has been recognised as a growing global issue, with preventative measures being considered for prospective victims. Consequently, research studies should look at detecting cyberbullying and devising preventive methods [4,5].

Traditional approaches are, however, difficult to scale and assess in this situation. These strategies are often based on human patterns and social networking sites with comparatively small data sampling [6]. When these technologies are applied to massive online social networks with vast size and scope, various challenges arise. Besides, the rapid rise of social networks encourages and disseminates violent behavior by offering venues and media platforms for perpetrating and propagating it. On the other side, social networks provide valuable information for studying human behavior and interactions on a wide scale, and researchers may utilize this information to create effective ways for identifying and limiting misbehaving and/or violent behavior. Besides, social networks offer a platform for criminals to conduct illicit acts. To identify and limit aggression and violence in complex systems, the approaches that target both components as content and network should be improved.

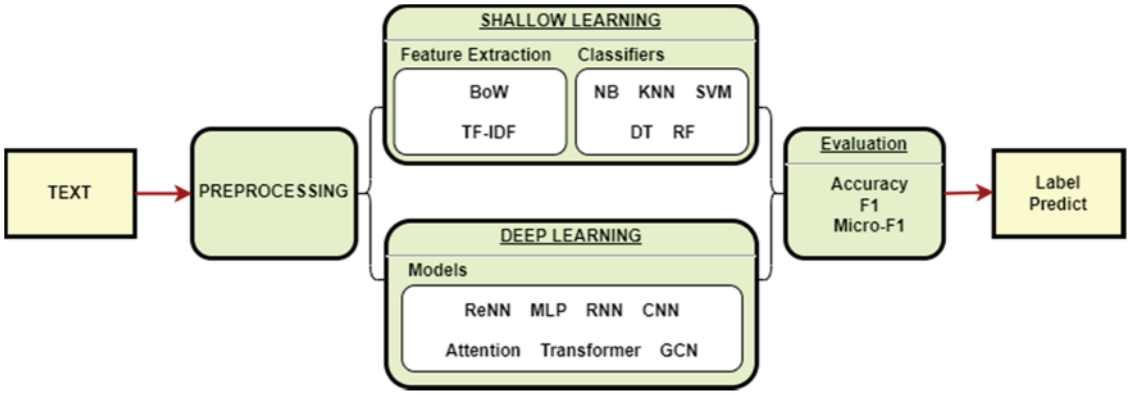


Figure 1: Flowchart of Shallow-to-Deep cyberbullying detection

In light of shallow and deep models, Fig. 1 depicts a flowchart of the methods involved in text categorization. Numerical, picture, and signal data are not the same as text data. It necessitates the cautious use of natural language processing (NLP) methods. Preprocessing text data for the model is the first and most crucial step. Shallow learning algorithms often need artificial ways to generate acceptable sample features, which are subsequently classified using traditional machine learning techniques. As a result, feature extraction severely limits the method's practicality.

In this paper, we applied six shallow learning and three deep learning methods on cyberbullying detection problem in social media for assessing the effectiveness of each method and to determine the

most advantageous features for the given problem. In order to achieve the assigned goal, firstly we review the state-of-the art researches. The result of literature review is described in Section 2. Section 3 explains problem statement. Section 4 explains materials and methods including description of each algorithm and architecture of deep learning techniques for cyberbullying detection problem. Section 5 demonstrates obtained experiment results. Section 6 discuss the results by indicating advantages and disadvantages of each method. At the end of the paper, we conclude the study referencing the main results and future work.

2 Related Works

Cyberbullying detection has been widely researched, beginning with user studies in the social sciences and psychology sectors, and more recently shifting to computer science with the goal of building models for the automated identification. It is worth noting that the first research on the automated identification of cyberbullying was published in 2009 [7]. To identify online harassment, the researchers used feature learning approaches and a linear kernel classifier on three separate data sets. The first dataset came from a chat-style platform, while the others came from discussion-style groups. Although the experimental findings were insufficient, the study inspired more investigation with regard to the issue.

Detecting abusive language and cyberbullying via social media has become more critical in order to avoid future damage. As a result, various researches have been conducted for over a decade to call for action on the issue of cyberbullying. The next research [8] utilized a machine learning technique to identify vandalism in Wikipedia. To identify cyber-violence from sexual abusers via the internet, Sadiq et al. [9] applied the methods of text analysis and communication theory. Sarac Essiz et al. [10] used an extension of the least-square support vector machine (SVM) to identify online spam assaults in social media. To assist online harassment, Gomez et al. [11] applied traditional machine learning methods for text classification. Many authors have tried to identify cyberbullying using various machine learning algorithms all throughout the internet, particularly in social media [12]. The use of expanded machine learning approaches to identify cyberbullying has achieved significant progress in recent years [13].

To train embeddings, a recent research in cyberbullying detection has mostly focused on employing recurrent neural networks (RNN) and pre-trained models using plain tokens [14]. Other studies offer alternative solutions [15], by using psychological variables such as personalities, attitudes, and emotions in order to improve automated cyberbullying-related texts classification. However, these results were obtained by means of the rudimentary models. Despite its promising prospects, the use of language preprocessing and linguistic word embedding to enhance classification performance has not been researched further.

The researchers have been working on cyberbully identification for many years, hoping to discover a mechanism to manage or prevent cyberbullying on social networking sites. Victims of cyberbullying are unable to deal with the emotional toll of receiving aggressive, frightening, demeaning and angry texts. The problem of cyberbullying must be explored in terms of identification, prevention and mitigation in order to limit its destructive impacts.

Reference [16] investigated the prediction performance of n-grams, part-of-speech, and sentiment analysis data based on lexicons for cyberbullying identification. Similar characteristics were used not only to recognize poorly graded cyberbullying, but also to detect finer-grained cyberbullying classes [17]. Content-based characteristics are often utilized in recent methods to cyberbullying detection [18],

due to their perceived simplicity. In fact, according to [19], more than 41 articles have used content-based characteristics to identify cyberbullying, indicating that this sort of information is critical for the job.

However, semantic features generated from topic design model [20], word embeddings, and knowledge discovery [21] are increasingly being integrated with content-based attributes. In recent years, several experiments have been undertaken on the contribution of machine learning algorithms to social network data analysis. Numerous models, techniques, and approaches were applied for processing big and unstructured data as a result of machine learning research [22]. Social media content has been widely analyzed for spam, phishing, and cyberbullying prediction [23–25]. Spam spread, phishing, malware dissemination, and cyberbullying are all examples of aggressive conduct. Due to the fact that social networking sites provide users the freedom to publish on their services, textual cyberbullying has been the most common hostile conduct [26–28].

In conclusion, it is important to point out the significance of the study as well as its uniqueness concerning the investigation of the cyberbullying phenomena caused by students' hatred against other race, ethnic, or religious groups. This feature, which has already been emphasized, is of utmost significance in current online learning settings, which mirrors the cultural variety of modern world [29]. In order to achieve the desirable outcome, the research makes use of a tool that has been modified and shown to be reliable. This instrument is made up of questions that demonstrate anti-values such as intolerance, xenophobia, or a lack of empathy towards cultural variety [30]. In contrast to the findings of other studies [31,32], in which cyberbullying was analyzed as a broad construct of the connections that students display while they are online, our results show that cyberbullying is a specific kind of behavior. Even if cases of cyberbullying take place in multicultural settings, this kind of motive for cyberbullying is still quite clear to see. In addition, the study takes into consideration the cultures, races, and ethnicities of the multicultural sample that was used in the first place. Due to the fact that the findings allow us to clearly outline which of the evaluated groups are more likely to be victims due to the non-acceptance of the cultural context, lifestyles, or physical features of the other, we can assume that the results have given us the ability to clearly highlight which of the evaluated groups.

3 Problem Statement

In comparison with the challenge of cyberbullying classification, the issue of early detection of cyberbullying on social networking sites might be regarded to be distinct. In this instance, there is a collection of social media sessions, which we will refer to as "S". Thus, there is a possibility that some are acts of cyberbullying. Eq. (1) shows how a series of social network sessions is characterized:

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

where $|S|$ means the number of sessions, s_i refers to session i.

Each session, $s \in S$, is composed of many posts placed one after another, denoted as P_s, and a indication of binary value b_s, determines whether or not the exact interaction in question constitutes cyberbullying ($b_s = true$) or not ($b_s = false$). The order in which postings are made during a particular session will shift throughout the course of time and be determined by:

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

where the tuple $\left\langle P_k^S, t_k^S\right\rangle$, $k \in [1, n]$ represents the k-th post for social network session and s is the timestamp when post P_k^S is published.

At the same time, a vector of characteristics P_k^S is used to uniquely identify each post:

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

Presented with a social media session s, the goal is to determine if the session is related to cyberbullying while analyzing as few postings P_s as feasible. Therefore, the goal is to learn a function $f\left(b_s \mid s_i, \left\langle P_1^{s_i}, t_1^{s_i}\right\rangle, \ldots, \left\langle P_k^{s_k}, t_1^{s_k}\right\rangle\right)$ to classify if a text is cyberbullying-related or non-related.

4 Materials and Methods

This section discusses the evaluation metrics of each classifier that was utilized, in addition to describing the dataset that was utilized for the purpose of detecting instances of cyberbullying on social network.

4.1 Dataset

Theoretical and practical difficulties arise when we attempt to identify cases of cyberbullying on social networking sites by searching for cyberbullying-related keywords or by using machine learning for classification purpose. From a more pragmatic point of view, the researchers are still working on identifying and categorizing inappropriate information based on the learning model. To create a cyberbullying detection model that is both effective and economical, the classification performance and the implementation of the appropriate model continue to be significant challenges. In this work, we evaluated six classification models that are often used in the process of detecting material related to cyberbullying using three datasets as Twitter dataset [33], Cyberbullying Classification Dataset [34], Hate Speech and Offensive Language Dataset [35]. Hence, our dataset is comprised of two sections, each of which is derived from one of two different sources. The development of each classifier will proceed in accordance with the performance measures that were covered in next sections.

4.2 Model Overview

Fig. 2 is an illustration of the model that has been developed for the detection of cyberbullying. This model consists of four stages: preprocessing stage, feature extraction, classification stage, and assessment stage. This section devotes considerable attention to analyzing each step more thoroughly.

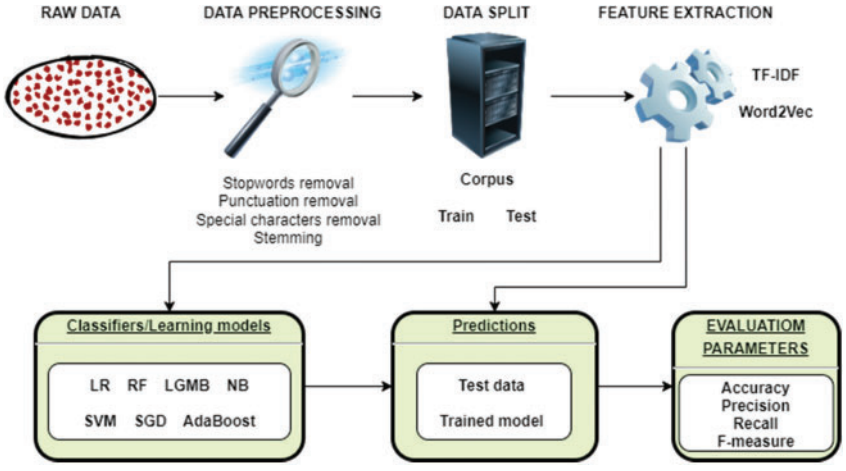


Figure 2: Flowchart of the shallow-to-deep cyberbullying detection

4.3 Feature Extraction

Term frequency-inverse document frequency (TF-IDF). In the process of classifying texts related to cyberbullying, feature extraction is an essential stage. In order to extract features for the model, we employed the TF-IDF and Word2Vec approaches. This technique for text feature extraction uses word statistics as its foundation, and it is a hybrid of two other algorithms known as term frequency (TF) and inverse document frequency (IDF), which stands for term frequency-inverse document frequency. This approach simply takes into account those word expressions that are consistent across all of the texts [36]. As a result, the TF-IDF algorithm is one of the feature extraction methods that is frequently applied in text detection [37]. In order to process text, Word2Vec is a neural network with two layers that "vectorizes" the words. It takes a corpora of text as its input and produces a collection of vectors as its output, which are attribute vectors that represent words inside that structure [38]. The Word2Vec technique builds a high-dimensional vector for each word by using the Skip-gram model, continuous bag-of-words (CBoW), and two hidden layers of shallow models [39]. The purpose is to raise the probability that:

$$\arg\max_{\theta} \prod_{w \in T} \left[\prod_{c \in C} p\left(c \mid w; \theta\right) \right] \tag{4}$$

where T is text, and θ is a parameter of p (c |w; θ).

Word2Vec Embeddings. Word2vec is a word-embedding tool that was developed by Google [40]. It is both effective and efficient. The word2vec program use a two-layer neural network language model to educate itself on the vector representations of each individual word. Actually, the application incorporates distinct models as CBoW and Skip-gram. These two models take opposite approaches to achieving their respective training objectives. CBoW makes an attempt to forecast a word based on the words that surround it, while Skip-gram endeavors to anticipate a window of words based on only single word. Word2vec is able to be trained on a large-scale unannotated corpus despite having limited computing resources due to the remarkably efficient design and unsupervised training technique that it utilizes. Word2vec embeddings may be learnt, and via this process, significant linguistic connections between words can be represented.

Bag-of-Words. To begin the process of extracting Bag-of-Words characteristics, initially a vocabulary that contains unigrams and bigrams must be organized. Terms with document frequencies that are lower than two are disregarded completely. Several distinct term weighting systems, such as tf-idf and binary ones, are also viable options in this context [41]. The tf-idf weighting system is the one that we use in this particular research. The following formula is used to get the tf-idf weight that corresponds to the i-th word in the j-th document:

$$w_{i,j} = TF_{i,j} \times \log\left(\frac{N}{DF_i}\right) \tag{5}$$

4.4 Machine Learning in Cyberbullying Detection

The vast majority of machine learning-based text categorization algorithms depend heavily on data as an essential component. On the other hand, data is devoid of any significance unless it is processed in such a way as to provide either information or implications. The data obtained from social networks are used in the selection process for both the training and testing datasets.

Decision Trees. In the field of machine learning, a decision tree (DT) is a well-known classification technique that is also one of the inductive learning methods that is used the most often. It can handle

training data that contains missing values as well as discrete variables. The idea of information entropy is used in the construction of decision trees, which are constructed afterwards by labelling training data [42]. Because of their resistance to noisy data and their capacity to learn disjunctive phrases, it would seem that they are appropriate for text categorization [43].

Random Forest. Random forest (RF) is an ensemble learning method that builds multitudes of decision tress. A visual approach known as a decision tree is one in which each branch node symbolizes a choice that may be made between several options. In decision trees, a graphical technique is used to evaluate and contrast the many available options.

Naïve Bayes. In the realm of machine learning, the Naïve Bayes (NB) algorithm is regarded as one of the most effective and efficient inductive learning algorithms [44], and it has been implemented as an efficient classifier in a number of different research projects concerning social media. Bayes is known as the "theorem of induction." Given a class variable y and a dependent feature vector x_1 through x_n, According to Bayes' theorem, the following connection exists:

$$p\left(y \mid x_1, x_2, \ldots, x_n\right) = \frac{p\left(y\right) p\left(x_1, x_2, \ldots, x_n \mid y\right)}{p\left(x_1, x_2, \ldots, x_n\right)} \tag{6}$$

By making the simplistic assumption of independence, we are able to draw the following equation:

$$p\left(y \mid x_1, x_2, \ldots x_{i-1,x_{i+1}}, x_n\right) = p\left(x_i \mid y\right) \tag{7}$$

Logistic Regression. This particular machine learning classifier, known as the logistic regression (LR) classifier, is one of the best-known and widely used methods. The majority of applications for them involve binary classifications, which result in outcomes that can only be either 0 or 1.

The hypothesis function for logistic regression is generally given as follows:

$$h\left(x\right) = \frac{1}{1 + e^{Tx}} \tag{8}$$

An optimized cost function can be given as:

$$Cost(h(x), y) = -y \log(\log\left(h\left(x\right)\right) - \left(1 - y\right) \log\left(1 - h\left(x\right)\right) \tag{9}$$

K nearest neighbors. K nearest neighbors (KNN) technique is a non-parametric approach that takes a majority vote to decide the class label of x_0 and seeks to identify the K neighbors that are located closest to x_0. In the absence of any previous information, the KNN classifier will often resort to using Euclidean distances as the distance measure.

Support Vector Machine. The Support Vector Machine (SVM) is a supervised approach that was developed on the basis of statistical learning theories [45]. It is often used for the purpose of identifying anomalies and network intrusions. Let $x_1, x_2, \ldots, x_n$ be training set, which are expressed as vectors in a certain space $X \subseteq Rd$. These data are labeled $y_1, y_2, \ldots, y_m$, where $y_1 \in (-1, 1)$ SVMs may be reduced to their most basic form, which is a hyperplane that divides the training data by the greatest possible margin. Class (-1), which is located on one side of the hyperplane, contrasts with class (1), which is located on the opposite side of the plane. In this scenario, non-cyberbullying behaviors are located on one side of the hyperplane, while cyberbullying behaviors are located on the other side.

4.5 Deep Learning in Cyberbullying Detection

Long Short Term Memory. The Long Short Term Memory (LSTM) is a type of Recurrent neural network that is able to successfully retain information for an extended length of time by including a

"memory cell". "The input gate," "the forgetting gate," and "the output gate" are primarily responsible for controlling the behavior of the memory cell. The input gate is responsible for activating the process of information being input into the memory cell, while the forgetting gate is responsible for selectively erasing some information that is stored in the memory cell and activating the storage in preparation for the subsequent input. Last but not least, the information that will be sent out of the memory cell is determined by the output gate.

Fig. 3 is an illustration of the LSTM network's organizational structure. Each box has a unique set of data, and the lines with arrows signify the flow of data between various sets of data. It is possible to comprehend how LSTM retains memory for an extended length of time by referring to Fig. 3.

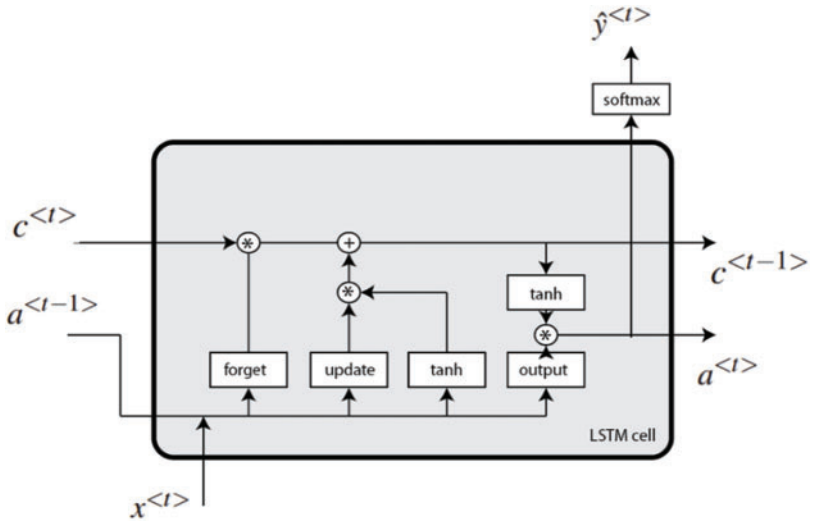


Figure 3: Flowchart of the LSTM

LSTM takes a set of input sequences as a *xi* vector, $x = (x_1, x_2, \ldots, x_t)$ and outputs a *yi* vector, $y = (y_1, y_2, \ldots, y_t)$, which is computed using to the next equations:

$$i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1} + b_i) \tag{10}$$

$$o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1} + b_o) \tag{11}$$

$$f_t = 1 - i_t) \tag{12}$$

$$tc_t = g(W_{cx}x_t + W_{cm}m_{t-1} + b_c) \tag{13}$$

$$tc_t = f_t \odot c_{t-1} + i_t \odot tb_t) \tag{14}$$

$$m_t = o_t \odot h_t \tag{15}$$

$$y_t = \varnothing(W_{ym}m_t + b_y) \tag{16}$$

Within those calculations, *i* indicates the input gate, while *o* is output gate, h_t is the forget gate. *tc* is the data supplied to the memory cell, and *c* comprises cells activation vectors, and *m* is the data the memory cell produces. *W* denotes weight matrices (e.g., *Wix* denotes the weight matrix from input *x* to the input gate *i*). *b* is the bias, while is cell input activation function and *h* is the cell output activation

function, considered as *tanh* or *inear* function. ⊙ is the point product in a matrix. ϕ is the activation function of the neural network output.

Within those calculations, i indicates the input gate, while o is output gate, f is the forget gate. tc is the data supplied to the memory cell, and c comprises cells activation vectors, and m is the data the memory cell produces. W denotes weight matrices. b is the bias, while g is cell input activation function and h is the cell output activation function, considered as *tanh* or *linear* function. ⊙ is the point product in a matrix. ϕ is the activation function of the neural network output.

Following the completion of a series of tests, we have determined that, in comparison to the ft standard equation, (14) is less complicated and requires less effort to converge. Not only the total amount of time spent in training decreases, but also the total number of iterations deteriorates as well. As a result, instead of using the ft standard equation to compute ft in the neural networks, we utilize the formula (14).

Bidirectional Long Short Term Memory. Bidirectional Long Short Term Memory (BiLSTM) is a kind of recurrent neural network that is most often used in NLP problems. In contrast to a conventional LSTM, the input travels in both ways, and the system is able to make use of knowledge through both sides. In addition to this, it is an effective method for representing the sequencing interdependence that exist between words in both the forward and backward integration of the sequence. Thus, BiLSTM consists of the addition of one extra layer of LSTM, which inverts the normal flow of data. To put it simply, it indicates that the extra LSTM layer processes the input sequence in reverse order. After that, we aggregate the outputs of the two LSTM layers in a number of different ways, including averaging, summing, multiplying, or concatenating them. To provide an example, the unrolled version of the BiLSTM is shown in Fig. 4:

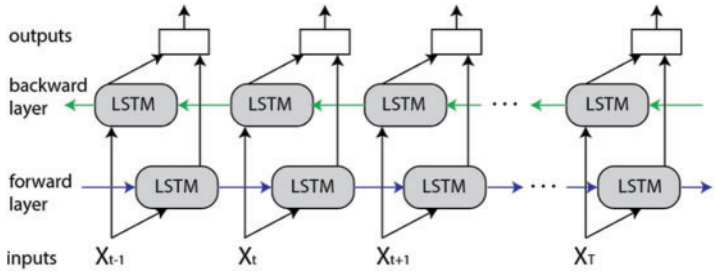


Figure 4: Flowchart of the BiLSTM

Convolutional Neural Network. Convolutional neural networks (CNN) uses convolving filters layers, which are then applied to the local characteristics of interest [46]. CNN models were initially developed for the purpose of pattern recognition; however, it has since been demonstrated that these models are also useful for natural language processing. CNNs have achieved great success in semantic text modeling, querying retrieval, text processing, and other classical NLP problems. The CNN design that Collobert et al. developed has some similarities to the network model that is shown in Fig. 5.

Let xi ∈ Rk be the k-dimensional word vector that corresponds to the i-th word in a phrase be denoted by the notation xi Rk. A phrase of length n, with extra words inserted where they're needed, is expressed as

$$x_{1:n} = x_1 \oplus x_2 \oplus \ldots \oplus x_n \tag{17}$$

here ⊕ is the operator for concatenating strings. Generally, let $x_{i:i+j}$ refers to the concatenation of next values $x_i, x_{i+1}, \ldots, x_{i+j}$. The use of a filter $w \in R^{hk}$ is required for a convolution process.

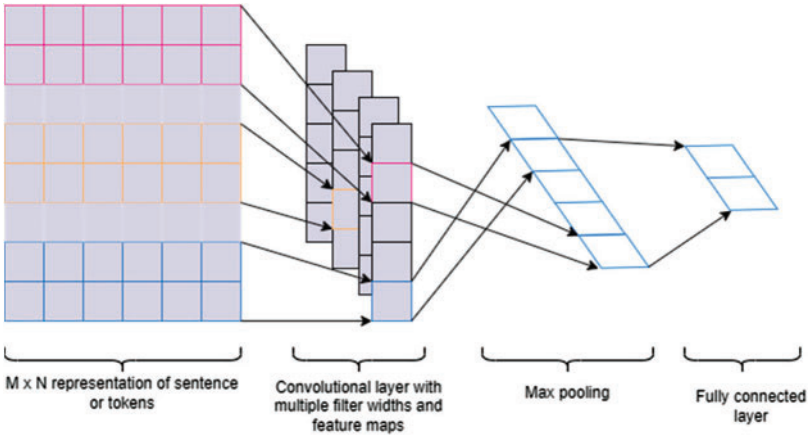


Figure 5: CNN model architecture

A convolution operation involves a filter $w \in R^{hk}$, that applied to a certain range of words, results in the production of a new feature. For instance, the c_i feature is produced from a list of words in a window $x_{i:i+h-1}$ by next equation:

$$x_{1:n} = x_1 \oplus x_2 \oplus \ldots \oplus x_n \tag{18}$$

Here $b \in R$ is a bias term and f is a non-linear function. This filter is used on every conceivable permutation of the sentence's words $\{x_{1:h}, x_{2:h+1}, \ldots, x_{n-h+1:n}\}$ to generate a feature map

$$x_{1:n} = x_1 \oplus x_2 \oplus \ldots \oplus x_n \tag{19}$$

with $c \in R^{n-h+1}$. After that, we put the feature map through a max-overtime pooling operation and use the maximum value $c^{\wedge} = \max\{c\}$ as the feature that corresponds to this specific filter [47]. The goal of this endeavor is to ensure that each feature map accurately depicts the aspect of greatest value and significance. The various sentence lengths are handled in a natural manner by this pooling approach.

We have discussed the procedure that must be followed in order to extract one feature from one filter. To get a wide variety of characteristics, the model applies many filters, each of which has a unique window size. The penultimate layer is formed by these characteristics, which are then sent to a fully linked softmax layer. The output of this layer is the probability distribution over categories.

5 Experiment Results

Threats arising in the network environment naturally stimulate the development of a research apparatus for studying their factors, mechanisms, and consequences.

5.1 Evaluation Metrics

In cyberbullying detection research uses Accuracy, Precision, Recall, F-measure, and AUC-ROC Curve as evaluation parameters.

$$accuracy = \frac{TP + TN}{P + N} \tag{20}$$

$$precision = \frac{TP}{TP + FP} \tag{21}$$

$$recall = \frac{TP}{TP + FN} \tag{22}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{23}$$

5.2 Experimental Results

In cyberbullying detection research uses Accuracy, Precision, Recall, F-measure, and area under a receiver operating characteristic (AUC-ROC). Fig. 6 demonstrates confusion matrixes for each algorithms that applied in this study that tested in cyberbullying classification dataset. Using confusion matrices, we can clear visualize exactly number of classification results in relation to other classes. There are three types of classes as cyberbullying that is noted as 1, non-cyberbullying that is noted as 0, and neutral class that is noted as 2 in our study.

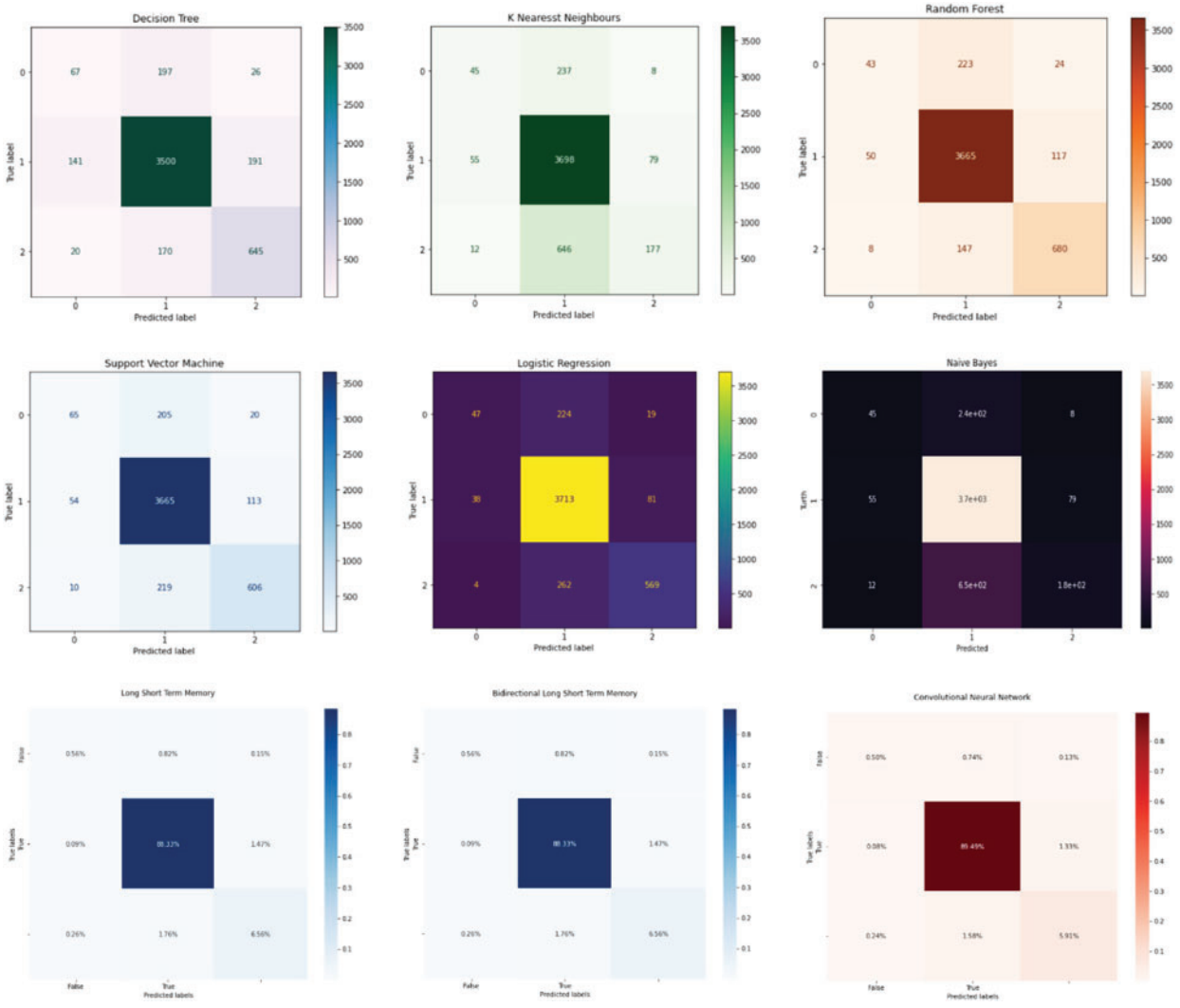


Figure 6: Comparison of confusion matrices

Fig. 7 compares the proposed model and all the applied machine learning and deep learning models in terms of accuracy, precision, recall, and F1-score. As the figure demonstrates, the proposed deep neural network outperforms all the applied machine learning and deep models in terms of accuracy, precision, F1-score in all the three datasets.

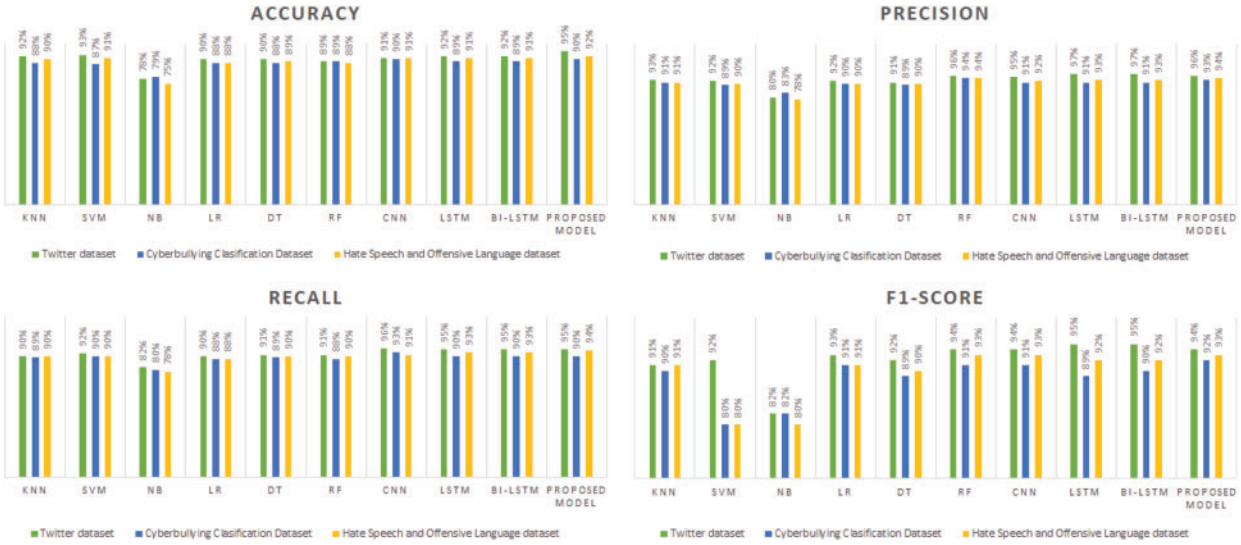
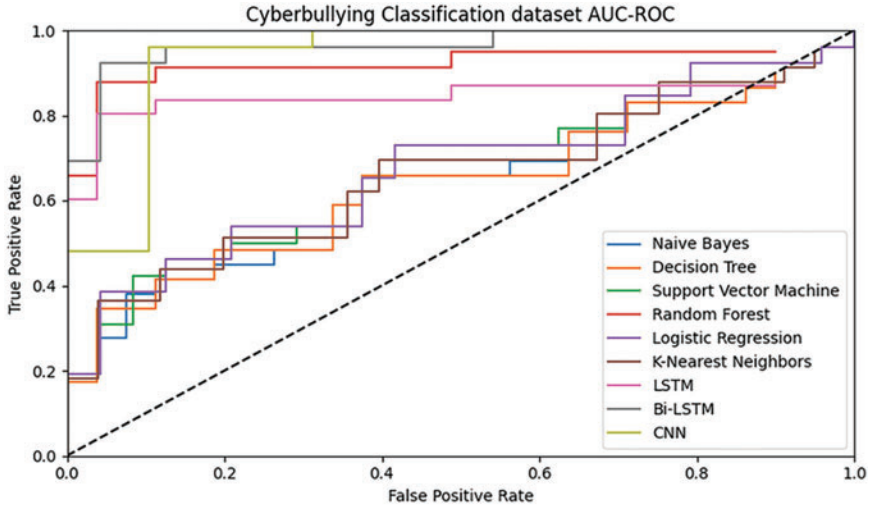


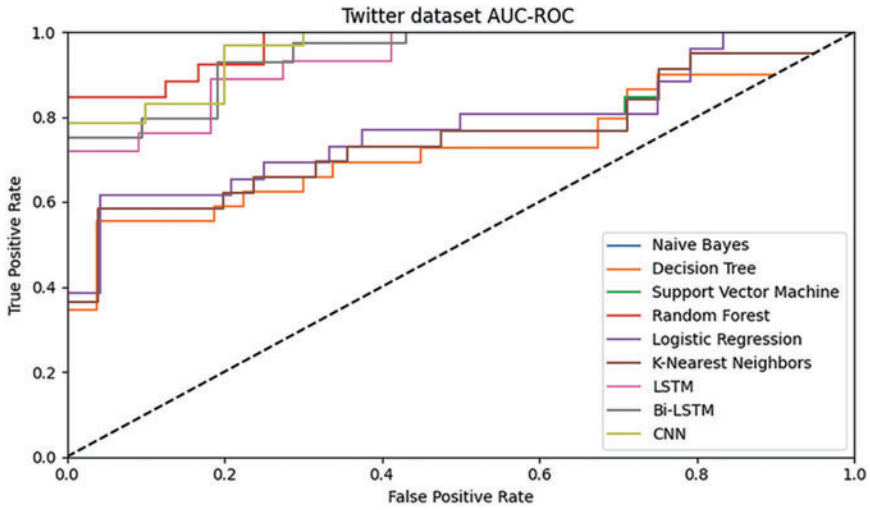
Figure 7: Comparison of the results for different datasets

The area under the receiver operating characteristic curve with all extracted characteristics is the AUC performance assessment in each classification. Fig. 8 compares AUC-ROC curves of all the applied methods and the proposed approach. As, it is noted, deep models demonstrated higher value than machine learning techniques. The figure shows that the proposed model, LSTM and BiLSTM, shows the highest AUC-ROC value from the first iteration and along the all graph.

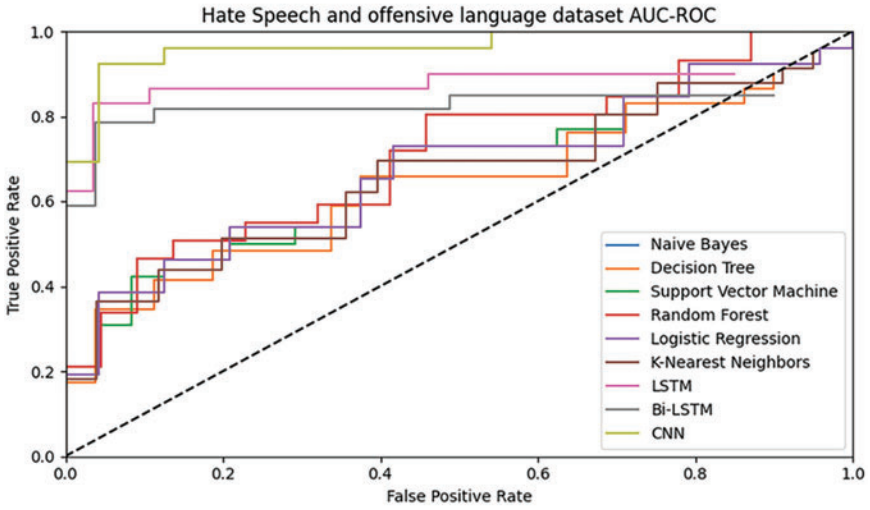


(a) The ROC curve of the applied models for Cyberbullying Detection dataset

Figure 8: (Continued)



(b) The ROC curve of the applied models for Twitter dataset



(c) The ROC curve of the applied models for Hate Speech dataset

Figure 8: The ROC curve of the applied models for different datasets

Tab. 1 demonstrates the cyberbullying classification results by applying machine learning and deep learning methods for three datasets. To evaluate the machine learning and deep learning methods we used evaluation criteria such as accuracy, precision, recall, and F1-score [48–51].

Consequently, based on its performance rates, the proposed approach may be approved as a potential approach for detecting cyberbullying in social networking sites. Furthermore, in respect of all evaluation parameters, the presented deep neural network has the best performance for detecting cyberbullying in all instances. The proposed approach yielded good results, which may be attributed to the use of the proposed deep neural network for weight and bias tweaking, as well as a decrease in training time. The findings show that the proposed deep neural network approach may be simply adapted to accommodate current short and long texts.

Table 1: Cyberbullying detection research summary

Dataset	Method	Method	Accuracy	Precision	Recall	F1-score	AUC-ROC
Twitter Dataset	Deep Models	LSTM	89.2%	89.5%	89.8%	89.6%	93%
		BiLSTM	90.1%	89.6%	90%	89.8%	93%
		CNN	90.2%	91.6%	90.4%	89.9%	94%
	Shallow Models	LR	87.3%	85.2%	86.2%	85.1%	78%
		RF	85.6%	83.9%	83.1%	83.7%	92%
		DT	87.4%	83.2%	86.3%	85.1%	80%
		NB	60.2%	52.4%	58.5%	64.2%	65%
		KNN	85.1%	85.4%	82.2%	85.6%	77%
		SVM	86.2%	85.3%	83.7%	85.8%	78%
Cyberbullying Classification Dataset	Deep Models	LSTM	89.2%	89.5%	89.8%	89.6%	92%
		BiLSTM	90.1%	89.6%	90%	89.8%	92%
		CNN	90.2%	91.6%	90.4%	89.9%	93%
	Shallow Models	LR	87.3%	85.2%	86.2%	85.1%	75%
		RF	85.6%	83.9%	83.1%	83.7%	90%
		DT	87.4%	83.2%	86.3%	85.1%	76%
		NB	60.2%	52.4%	58.5%	64.2%	68%
		KNN	85.1%	85.4%	82.2%	85.6%	77%
		SVM	86.2%	85.3%	83.7%	85.8%	78%
Hate Speech and Offensive Language Dataset	Deep Models	LSTM	89.2%	89.5%	89.8%	89.6%	91%
		BiLSTM	90.1%	89.6%	90%	89.8%	91%
		CNN	90.2%	91.6%	90.4%	89.9%	93%
	Shallow Models	LR	87.3%	85.2%	86.2%	85.1%	75%
		RF	85.6%	83.9%	83.1%	83.7%	80%
		DT	87.4%	83.2%	86.3%	85.1%	79%
		NB	60.2%	52.4%	58.5%	64.2%	67%
		KNN	85.1%	85.4%	82.2%	85.6%	78%
		SVM	86.2%	85.3%	83.7%	85.8%	78%

6 Conclusion

Social media is a relatively new human communication medium that has grown in popularity in recent years. Machine learning is utilized in a variety of applications, including social network analysis. This review provides a thorough overview of different applications that use machine learning techniques to analyze social media to detect cyberbullying and online harassment. Our paper explores each step to cyberbullying detection on social media, such as data collection, data preprocessing,

data preparation, feature selection and extraction, feature engineering, applying machine learning techniques, and text classification. Various academics have proposed different methods to address the problems of generic metadata architecture, threshold settings, and fragmentation in cyberbullying detection on social networks data streams. To address problems with cyberbullying categorization in social network data, the review also proposed a general metadata architecture for cyberbullying classification on social media. In comparison with proportionate techniques, the proposed architecture performed better across all evaluation criteria for cyberbullying and online harassment detection.

In further, a more durable automated cyberbullying detection system can be developed by considering the problems as class imbalance data, binary and multi-classification, scalability, multilingualism, threshold settings, and fragmentation.

Funding Statement: The authors received no funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

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

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

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

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

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

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

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

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

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

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

[11] C. E. Gomez, M. O. Sztainberg and R. E. Trana, "Curating cyberbullying datasets: A human-AI collaborative approach," *International Journal of Bullying Prevention*, vol. 4, no. 1, pp. 35–46, 2022.

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

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

[14] S. R. Sangwan and M. P. S. Bhatia, "Denigrate comment detection in low-resource Hindi language using attention-based residual networks," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 1, pp. 1–14, 2021.
[15] T. T. Aurpa, R. Sadik and M. S. Ahmed, "Abusive Bangla comments detection on Facebook using transformer-based deep learning models," *Social Network Analysis and Mining*, vol. 12, no. 1, pp. 1–14, 2022.
[16] R. Yan, Y. Li, D. Li, Y. Wang, Y. Zhu *et al.,* "A stochastic algorithm based on reverse sampling technique to fight against the cyberbullying," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 4, pp. 1–22, 2021.
[17] C. J. Yin, Z. Ayop, S. Anawar, N. F. Othman and N. M. Zainudin, "Slangs and short forms of malay twitter sentiment analysis using supervised machine learning," *International Journal of Computer Science & Network Security*, vol. 21, no. 11, pp. 294–300, 2021.
[18] G. Jacobs, C. Van Hee and V. Hoste, "Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text?," *Natural Language Engineering*, vol. 28, no. 2, pp. 141–166, 2022.
[19] A. Jevremovic, M. Veinovic, M. Cabarkapa, M. Krstic, I. Chorbev *et al.,* "Keeping children safe online with limited resources: Analyzing what is seen and heard," *IEEE Access*, vol. 9, no. 1, pp. 132723–132732, 2021.
[20] K. Kumari, J. P. Singh, Y. K. Dwivedi and N. P. Rana, "Multi-modal aggression identification using convolutional neural network and binary particle swarm optimization," *Future Generation Computer Systems*, vol. 118, no. 1, pp. 187–197, 2021.
[21] A. M. Abbas, "Social network analysis using deep learning: Applications and schemes," *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 1–21, 2021.
[22] S. Gupta, N. Mohan, P. Nayak, K. C. Nagaraju and M. Karanam, "Deep vision-based surveillance system to prevent train-elephant collisions," *Soft Computing*, vol. 26, no. 8, pp. 4005–4018, 2022.
[23] S. Mohammed, W. C. Fang, A. E. Hassanien and T. H. Kim, "Advanced data mining tools and methods for social computing," *The Computer Journal*, vol. 64, no. 3, pp. 281–285, 2021.
[24] B. Thuraisingham, "Trustworthy machine learning," *IEEE Intelligent Systems*, vol. 37, no. 1, pp. 21–24, 2022.
[25] V. Rupapara, F. Rustam, H. Shahzad, A. Mehmood, I. Ashraf *et al.,* "Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model," *IEEE Access*, vol. 9, no. 1, pp. 78621–78634, 2021.
[26] O. Sharif and M. M. Hoque, "Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers," *Neurocomputing*, vol. 490, no. 1, pp. 462–481, 2022.
[27] K. Kumari, J. P. Singh, Y. K. Dwivedi and N. P. Rana, "Bilingual cyber-aggression detection on social media using LSTM autoencoder," *Soft Computing*, vol. 25, no. 14, pp. 8999–9012, 2021.
[28] A. Mohamed, E. Amer, N. Eldin, M. Hossam, N. Elmasry *et al.,* "The impact of data processing and ensemble on breast cancer detection using deep learning," *Journal of Computing and Communication*, vol. 1, no. 1, pp. 27–37, 2022.
[29] A. Sheth, V. L. Shalin and U. Kursuncu, "Defining and detecting toxicity on social media: Context and knowledge are key," *Neurocomputing*, vol. 490, no. 1, pp. 312–318, 2022.
[30] U. Kursuncu, H. Purohit, N. Agarwal and A. Sheth, "When the bad is good and the good is bad: Understanding cyber social health through online behavioral change," *IEEE Internet Computing*, vol. 25, no. 1, pp. 6–11, 2021.
[31] A. M. Veiga Simão, P. Costa Ferreira, N. Pereira, S. Oliveira, P. Paulino *et al.,* "Prosociality in cyberspace: Developing emotion and behavioral regulation to decrease aggressive communication," *Cognitive Computation*, vol. 13, no. 3, pp. 736–750, 2021.
[32] G. Isaza, F. Muñoz, L. Castillo and F. Buitrago, "Classifying cybergrooming for child online protection using hybrid machine learning model," *Neurocomputing*, vol. 484, no. 1, pp. 250–259, 2022.

[33] L. Cuoghi and L. Konopelko, "Cyberbullying classification," (accessed on 25 June 2022), 2022. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification.
[34] D. Bruwaene, Q. Huang and D. Inkpen, "A multi-platform dataset for detecting cyberbullying in social media," (accessed on 25 June 2022), 2022. [Online]. Available: https://dl.acm.org/doi/abs/10.1007/s10579-020-09488-3.
[35] A. Samoshyn, "Hate speech and offensive language dataset," (accessed on 25 June 2022), 2020. [Online]. Available: https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset.
[36] G. Perasso, N. Carone and L. Barone, "Written and visual cyberbullying victimization in adolescence: Shared and unique associated factors," *European Journal of Developmental Psychology*, vol. 18, no. 5, pp. 658–677, 2021.
[37] M. Amjad, N. Ashraf, A. Zhila, G. Sidorov, A. Zubiaga *et al.,* "Threatening language detection and target identification in urdu tweets," *IEEE Access*, vol. 9, no. 1, pp. 128302–128313, 2021.
[38] Ö. Çoban, S. A. Özel and A. İnan, "Deep learning-based sentiment analysis of facebook data: The case of turkish users," *The Computer Journal*, vol. 64, no. 3, pp. 473–499, 2021.
[39] B. Omarov, N. Saparkhojayev, S. Shekerbekova, O. Akhmetova, M. Sakypbekova *et al.,* "Artificial intelligence in medicine: Real time electronic stethoscope for heart diseases detection," *CMC-Computers, Materials & Continua*, vol. 70, no. 2, pp. 2815–2833, 2022.
[40] P. Parikh, H. Abburi, N. Chhaya, M. Gupta and V. Varma, "Categorizing sexism and misogyny through neural approaches," *ACM Transactions on the Web (TWEB)*, vol. 15, no. 4, pp. 1–31, 2021.
[41] S. Kiritchenko, I. Nejadgholi and K. C. Fraser, "Confronting abusive language online: A survey from the ethical and human rights perspective," *Journal of Artificial Intelligence Research*, vol. 71, no. 1, pp. 431–478, 2021.
[42] J. A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios and R. Valencia-García, "Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings," *Future Generation Computer Systems*, vol. 114, no. 1, pp. 506–518, 2021.
[43] A. Tontodimamma, E. Nissi, A. Sarra and L. Fontanella, "Thirty years of research into hate speech: topics of interest and their evolution," *Scientometrics*, vol. 126, no. 1, pp. 157–179, 2021.
[44] X. Chen, H. Xie, G. Cheng and Z. Li, "A decade of sentic computing: topic modeling and bibliometric analysis," *Cognitive Computation*, vol. 14, no. 1, pp. 24–47, 2022.
[45] A. S. Srinath, H. Johnson, G. G. Dagher and M. Long, "BullyNet: Unmasking cyberbullies on social networks," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 2, pp. 332–344, 2021.
[46] C. Kumar, T. S. Bharati and S. Prakash, "Online social network security: A comparative review using machine learning and deep learning," *Neural Processing Letters*, vol. 53, no. 1, pp. 843–861, 2021.
[47] M. Zhu, A. H. Anwar, Z. Wan, J. H. Cho, C. A. Kamhoua *et al.,* "A survey of defensive deception: Approaches using game theory and machine learning," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, pp. 2460–2493, 2021.
[48] H. Sun and R. Grishman, "Employing lexicalized dependency paths for active learning of relation extraction," *Intelligent Automation & Soft Computing*, vol. 34, no. 3, pp. 1415–1423, 2022.
[49] B. Omarov, A. Batyrbekov, K. Dalbekova, G. Abdulkarimova, S. Berkimbaeva *et al.,* "Electronic stethoscope for heartbeat abnormality detection," in *5th Int. Conf. on Smart Computing and Communication (SmartCom 2020)*, Paris, France, pp. 248–258, 2020.
[50] F. Bozyiğit, O. Doğan and D. Kilinç, "Categorization of customer complaints in food industry using machine learning approaches," *Journal of Intelligent Systems: Theory and Applications*, vol. 5, no. 1, pp. 85–91, 2022.
[51] B. Omarov, S. Narynov, Z. Zhumanov, A. Gumar and M. Khassanova, "A skeleton-based approach for campus violence detection," *Computers, Materials & Continua*, vol. 72, no. 1, pp. 315–331, 2022.