Cyberbullying Detection in Twitter Using Deep Learning Model Techniques

Anu Ranjana Seetharaman and Hamid Jahankhani

Abstract The aim of the research is to develop a combined deep machine learning model (DEA_RNN) that will detect cyber bullying on Twitter. The suggested approach merges Elman-type Recurrent Neural Networks (RNNs) with a refined Dolphin Echolocation Algorithm (DEA) to improve parameter optimization and reduce the training duration. The proposed approach can handle the energetic nature of brief writings, adapt with point models for compelling extraction of trending themes and is proficient in recognizing and classifying cyberbullying tweets. The four modules in the proposed system are Dataset Collection, Data Cleaning and Preprocessing, Algorithm Implementation, and Prediction. The data cleaning and preprocessing phase involve noise removal, out-of-vocabulary cleansing, and tweet transformations to improve feature extraction and classification accuracy. The model algorithm uses DEA and RNN to generate a kind of echo similar to the behavior of dolphins during the hunting process. This research attempts to address the issue of recognizing cyberbullying on Twitter, which could be a challenging errand due to the energetic nature of brief writings and the changing nature of cyberbullying. The proposed approach beats models such as Simple Recurrent Neural Network (RNN), in terms of execution.

Keywords Deep machine learning · Twitter · Recurrent neural networks · Social media · Cyberbullying · Convolutional neural network

A. R. Seetharaman · H. Jahankhani (✉)
Northumbria University, London, UK
e-mail: Hamid.jahankhani@northumbria.ac.uk

A. R. Seetharaman
e-mail: anu.seetharaman@northumbria.ac.uk

H. Jahankhani (ed.), *Cybersecurity Challenges in the Age of AI, Space Communications and Cyborgs*, Advanced Sciences and Technologies for Security Applications,
https://doi.org/10.1007/978-3-031-47594-8_7

1 Introduction

Social media platforms have gained widespread popularity and usage across individuals of all ages. However, with this popularity comes the increasing prevalence of cyberbullying. Cyberbullying (CB) refers to hostile conduct that is purposeful, recurrent, and entails a power imbalance or asymmetry, and which occurs via electronic means or the internet. Cyberbullying on Twitter is particularly challenging to detect due to the ever-changing characteristic of brief textual content. and the changing nature of cyberbullying. It is important to detect cyberbullying on Twitter to make social media safe. Content alignment based on administered computerized learning models is commonly utilized for putting tweets into offensive and non-offensive tweets. The directed classifiers have moo execution. We cannot change the course names and it is very noteworthy to used scenarios. Too, they may be appropriate as it were for a foreordained group pertaining occasions but cannot deal with tweets that change over time. Topic modeling ways have moreover been used to extricate crucial points from a set of information to make designs or classes within the whole dataset. Whereas the concept is comparable, point modeling has impediments in taking care of the energetic nature of tweets and in recognizing unused occasions and scenarios [11].

The present study recommends the usage of a hybrid deep learning approach termed as DEA-RNN to surmount the challenges in identifying cyberbullying on the Twitter social network. The proposed model is an amalgamation of Elman-based Recurrent Neural Networks (RNNs) and a fine-tuned Dolphin Echolocation Algorithm (DEA) that addresses the problems of parameter optimization and training time reduction. This approach offers numerous benefits over existing methods as it can handle the ever-changing characteristic of short texts, extract popular topics through topic modeling, and classify cyberbullying tweets with high accuracy. Cyberbullying affects to use technology and social medias to bully and hurt other's mental health. This mainly affects the youngsters, and the research are made on its impact especially on the children [2]. Cyberbullying can have genuine impacts on their passionate and mental wellbeing, counting expanding their levels of self-destructive ideation, sadness, and uneasiness.

In expansion, cyberbullying can also have long-term impacts on an individual's mental wellbeing. It is vital to note that cyberbullying not as it were influencing the casualties but moreover the culprits, who may encounter negative mental results such as uneasiness, disgrace, and blame. Cyberbullying may be a genuine issue that can have obliterating impacts on youthful casualties and their families. It is crucial to address this issue and provide support for those who have been affected by cyberbullying [1].

The contributions of this thesis include the proposed DEA-RNN model for cyberbullying detection, a new Twitter dataset collection for evaluation, and thorough experimental results that reveal its superior performance over other competing models. The proposed DEA-RNN model is a promising approach for detecting cyberbullying on Twitter and has the potential to be extended to other social media platforms as well.

2 Cyberbullying and Its Impact

Among the youths, cyberbullying is one of the growing concerns, and its effect on adolescents and children's well-being is of specific interest to the researchers. It can be defined as any aggressive behaviour which is also carried out through electronic devices or the internet [3]. Cyberbullying has serios impact on the mental health as well as on the emotional health of youngsters and children. Because of this, their suicidal thoughts will be developed, depression, and anxiety level will be increased.

According to the research outcome, it is shown that the victims of cyberbullying experience high depression and anxiety compared to the non-anxiety ones. Also, its impact is more dangerous compared to the traditional bullying. Its victims experience feelings of isolation, shame, and helplessness that lead to mental health problems.

As per the study it has also been seen that cyberbullying has significant effects on an individual's academic performance that result in low grades and desired not to go to school. In addition, their social relationships get affected in a negative way that led to low self-esteem and high social isolation.

Cyberbullying also shows long-term effects on a person's mental health that includes PTSD (Post-Traumatic Stress Disorder), depression, and anxiety [9]. The victims of cyberbullying also experience physical symptoms like sleep disturbances, stomach pains, and headaches. Its impact is not limited to the victims only, perpetrators also experience negative psychological consequences like anxiety, shame, and guilty.

Cyberbullying's effects could be damaging individuals physically. Cyberbullying's physical effects include sleeping issues, headaches, and stomach aches. In addition, all psychological effects involve suicidal thoughts, self-harm, depression, and anxiety. Cyberbullying's victims might have also trouble to form relationships with all other individuals. Such victims might be afraid of trusting others and mightn't want in socializing with all others. It could lead to them to feel isolated as well as lonely. Parents must be aware of all such effects along with taking appropriate actions if they sense the teen is cyberbullied. Cyberbullying could have some serious consequences for all bullies as well as the victims. Hence, both teenagers along with parents must know all effects of preventing cyberbullying from taking place or dealing with this if this happens.

3 Neural Network

A neural network is an AI-based algorithm which can process natural language, recognise speech and image. This algorithm is effective because of its ability to generalize and learn from the examples. This ability helps it to make predictions even in noisy and complex environments [14].

As per the study, the advancement in neural network algorithms has led to development of deep learning which mainly refers to the neural network having several layers. The use of deep learning is successful in many applications like speech and image recognition, autonomous driving, and natural language processing [7].

Neural networks are machine learning's subset and depend upon training data for learning and improving their accuracy with time. although once every learning algorithm is fine-tuned, each of these algorithms is a powerful tool in artificial intelligence and computer science that allow for classifying and clustering data at higher velocity. In addition, tasks in image recognition as well as speech recognition could take hours versus minutes if compared to manual identification [6]. Neural networks could be divided into different types that are utilized for different purposes. Multilayer perceptron (MLP) includes output layer, input layer along with one hidden layer. Such neural networks are developed of sigmoid neurons, in place of perceptron, as most issues are non-linear.

In Natural Language Processing (NLP), the trained data from the models are used. This trained data is also used for computer vision and other similar neural networks. The MLP and the Convolutional Neural Network (CNN) are similar, but the difference is the CNNs are used for image and pattern recognition and computer vision. Principles of linear algebra, especially matrix multiplication, are harnessed by such networks for identifying patterns in any image. Also, a recurrent neural network (RNN) can be identified using feedback loops.

Neural networks work quite the same as the neural network of any human brain. Any neuron within any neural network is the mathematical function which collects as well as classifies information as per a particular architecture. A strong and efficient resemblance is bear by the network to different statistical methods: regression analysis as well as curve fitting. Any neural network includes all interconnected nodes' layers. In addition, all nodes are termed perception and are the same as any multiple linear regression. The signal is fed by the perception created by the multiple linear regression into the activation function [15, 16]. Neural networks are used widely, with a different application for many financial operations, product maintenance, business analytics, enterprise planning, and trading. Neural networks evaluate price data along with unearthing opportunities to make trade decisions depending upon data analysis.

3.1 Neural Network Algorithm to Detect Cyberbullying

As per the study, the detection of cyberbullying is said to be a challenging task because of the generation of a huge amount of data on social media platforms. However, neural network algorithms have shown promise in identifying cyberbullying behaviour. CNN (Convolutional Neural Network) is one of the algorithms which achieved success in the detection of cyberbullying in social media.

CNNs work by processing input data in a series of convolutional layers, where each layer learns specific features from the data. Such features are then used to make predictions about the input data. In the context of cyberbullying detection, CNNs can learn to identify specific patterns of language, sentiment, and context that are associated with cyberbullying behaviour [15].

CNN is the algorithm which advances all current work to detect cyberbullying with the adoption of deep learning's principles in place of classical machine learning. The architecture of CNN includes four layers, such as embedding, max pooling, convolutional and dense. It can be achieved by generating word embeddings for all words in any tweet along with feeding these to the CNN directly. Word embedding is techniques' class used for generating textual material's numerical representation. Word embedding's striking feature is that these generate the same type of representations for similar semantic words. Such a remarkable feature also enables the machine in understanding what any text means in place of dealing with this as random numbers' string.

Another neural network algorithm which has been used to detect cyberbullying is long short-term memory (LSTM) network. Each of these LSTM networks is the kind of recurrent neural network which can learn and remember long-term dependencies in sequential data [6]. This makes them well-suited for detecting cyberbullying in text-based social media data, where the context of a post may be spread out over multiple sentences or comments.

LSTM network can be used as the ensemble model of deep learning with different layers that can outperform single-layer models of neural networks. For further optimizing these models, two-word embeddings' stacking could be done for enhancing the performance of these models. It works on all data in real-time and it assesses if any text contains content of cyberbullying by running this through the model. This LSTM network is utilized within different activities of deep learning and artificial intelligence that include text mining, sequence mining, image processing, and NLP.

As per the study [10] ensemble of CNN and LSTM models is used to detect cyberbullying in tweets. Their results showed also that this ensemble model has outperformed individual models, achieving 0.79 score. Authors noted that the use of neural network algorithms for cyberbullying detection is a promising approach, but more research is needed to develop models that are robust to variations in language and context.

These models make the detection of cyberbullying one fully automated approach along without any human involvement or expertise while guaranteeing also better results. In addition, comprehensive experiments have also proved that algorithms of deep learning have outperformed classical approaches of machine learning within the issue of cyberbullying. These models can improve the detection of cyberbullying in future and could be used for analysing along with rooting out online bullying's instances perpetrated by users of social media. Classifiers of deep learning to detect hateful texts as well as discussion contents that precede cyberbullying are developed along with might be applied in different platforms [13]. These models have demonstrated some promising performance and will improve much more in the future. Solving the issue is vital to control materials of social media in different languages along with protecting all users from any negative impact of any toxic comment such as offensive language or verbal assaults.

3.2 Naïve Bayes-Based Algorithm to Perform Content-Based Analysis

Naive Bayes (NB) is a probabilistic machine learning algorithm that has been successfully applied in text classification tasks. In the context of cyberbullying detection, NB-based algorithms have shown promise in performing content-based analysis to measure the presence and severity of cyberbullying in social media data [4].

NB algorithms work by computing the probability of each feature in the input data belonging to each class. The class with highest probability then gets assigned to the input data (Andrade et al. 2013). In the context of cyberbullying detection, the input data may be a post, comment, or message, and the classes may be cyberbullying or non-cyberbullying.

To use NB-based algorithms for cyberbullying detection, researchers have developed various features that can capture the linguistic and semantic characteristics of cyberbullying behaviour. These features may include word frequency, sentiment, topic modelling, and other language-based features [12].

The algorithm of Naïve Bayes is the supervised algorithm that is based upon Bayes theorem along with utilized to solve classification issues. This algorithm is used within text classification that includes one high-dimensional set of data to train. Naïve Bayes classifier is an efficient algorithm of classification that helps develop fast models using machine learning which could make some fast predictions. This algorithm assumes that a specific feature's occurrence is independent of other features' occurrence. Each feature of this algorithm contributes individually to identification. Any Naïve Bayes model is easy to develop and quite useful for huge data sets. This algorithm is also known for outperforming every highly sophisticated method of classification. This algorithm performs better for categorical input variables in place of numerical variables.

There exists training data for training the model along with making this functional. Then, there is a need in validating all data to evaluate the model as well as make new predictions. With the use of conditional probability, the evidence's probability can be calculated from all given outputs. The primary goal is computing the output's probability-based upon all current attributes. This algorithm of Naïve Bayes is utilized as the technique of probabilistic learning for classification of texts. It is a well-known algorithm used for classification of documents of different classes. This algorithm is also used for analysing feelings or sentiments, whether positive, negative, or neutral.

Although the algorithm of Naïve Bayes has numerous limitations, still this is the most selected algorithm to solve issues of classification due to its simplicity. This algorithm works properly upon spam filtering as well as document classification. This algorithm possesses the highest success rate when compared to all other algorithms due to its efficiency as well as speed. In this Naïve Bayes classifier, all features are independently used, and each feature is independent of one another.

3.3 Content-Based Analysis Using Icons and Hieroglyphs

The use of icons and hieroglyphs as selection features for training and testing machine learning algorithms for cyberbullying detection has gained attention in recent years. These features can capture the visual and symbolic characteristics of cyberbullying behaviour that may not be adequately captured by language-based features.

To use icons and hieroglyphs as selection features, researchers have developed various techniques to extract and represent them as numerical features that can be used in machine learning algorithms. These techniques may include colour-based feature, texture-based feature, shape-based feature, and other visual features.

Studies have shown that using icons and hieroglyphs as selection features can improve the accuracy of machine-learning algorithms in cyberbullying detection. For example, as per the study [5] Support Vector Machine (SVM) algorithms are used with visual features to detect cyberbullying in Arabic tweets. Their results showed that the SVM-based algorithm achieved an accuracy of 85.71% in detecting cyberbullying, which is much higher when compared to accuracy achieved by language-based features alone [8].

Visual information, like hieroglyphs, images, and icons, capture imagination. Several hieroglyphs analysis needs epigraphers in spending much time to browse existing catalogues for identifying individual graphs. Technical advances within digitization, information management, and automatic analysis of images are enabling this possibility in analysing, organizing, and visualizing huge cultural datasets. As a visual cue, shapes are used in different tasks of image analysis. Techniques of data visualization that organize as well as present data within one structured graphical format, could be applied for visualizing glyph images along with enabling efficient browsing as well as a search of hieroglyph datasets. These systems could enable mining along with the discovery of semantic as well as visual patterns of hieroglyphs. However, in multimedia computing systems' design to realize cross-cultural computing, an issue is searching and analysing modern data.

Algorithms of computer vision have shown some potential in providing new insights into the digital humanities' realm. Different systems are proposed for aiding the analysis of artistic, historical, and cultural materials that could efficiently facilitate scholars' daily life in this field. Visual data retrieval is used to search technique drawings, clip-arts, hand-drawn images, 3D objects, trademark images, and natural image datasets. All traditional retrieval systems involve grid-based matching, curve fitting, and point-to-point matching. Such methods either don't scale well upon luge datasets or offer only limited flexibility upon shape variations [17]. However, local shape descriptors have been proposed recently and used for retrieval based on shapes. These methods could scale sub-linearly with proper search structures.

Human speech's emergence can be considered as a coincidental which largely depends upon both non-verbal as well as verbal forms. However, every writing system is different, as writing could not be classified to be a language; this is simply the way to record spoken languages using visible signages. Writing's pictorial style is accepted as the means to communicate. Pictographic writing systems' study has introduced many implicate to develop cultures as well as languages. Such systems include ideas, objects, images, and symbols. Such icons also function as the noun as they explain one word which means one point; however, to construct the meaning, this pictographic work needs the syllables' hierographic structure. Syllabifications are quite like constructing linguistic words as units' combinations within a language. In this new millennium, every emoji icon goes through a similar construction.

3.4 Decision Tree Algorithms to Detect Harmful Words

In the context of cyberbullying detection, decision tree algorithms can be used to classify harmful words into different categories based on their severity and context. Researchers have developed various decision tree-based algorithms for this purpose, such as C4.5, CART, and ID3. These algorithms use different criteria for feature selection and splitting nodes in the decision tree.

The social media platform has become public expression's modern channel for all individuals irrespective of every socio-economic boundary. Due to this pandemic situation, the common people also have begun looking to platforms of social media as the formal manner for being connected with different masses. Several researchers have also published their work upon the automated detection of offensive content as well as hate speech.

Sentiment analysis is a process to classify entities into neutral, negative, or positive categories with the use of NLP. Several models of deep learning and machine learning are used for tackling the issue of the detection of hate speech. Random Forrest Classifier is the ensemble learning approach which uses multiple decision trees for regression along with classification. Such classifiers aggregate decision trees' results for improving the whole accuracy. For any provided data input, there is the creation of one decision tree for all samples within the dataset. Each tree's prediction is voted upon as well as the prediction having the highest votes is also determined as end prediction. Sentiment analysis is performed on the dataset for obtaining each sentence's polarity, along with testing if the analysis will be enough for classifying all meanings.

Accomplishing the detection of harmful words is a tough task as this includes processing text along with understanding that context. The datasets of harmful words are not clean usually; hence, they should be pre-processed before algorithms of classification could detect harmful words in them. In addition, different models of machine learning have also different strengths which make them better than all others for specific tasks like detecting harmful words. A few models are much more accurate while the remaining others are much more efficient. Also, it is crucial in using different models along with comparing their performance for finding the most appropriate one for the detection of harmful words. Some pre-training approaches have also become popular in the last couple of years, and this is crucial in testing if they work properly with detection algorithms of harmful words.

4 Research Methods

Social media stages have picked up broad notoriety and utilization over people of all ages. Be that as it may, with this notoriety comes the expanding predominance of cyberbullying. Cyberbullying (CB) is characterized as forceful behavior that's purposefulness, rehashed, and includes a lopsidedness of control or quality, which takes put through electronic gadgets or the web. Cyberbullying on Twitter is especially challenging to identify due to the energetic nature of brief writings and the changing nature of cyberbullying. Topic modeling techniques have also been employed to extricate crucial subjects from a set of information to make designs or classes within the whole dataset. Whereas the concept is comparative, point modeling has confinements in dealing with the energetic nature of tweets and in recognizing unused occasions and scenarios [11].

The aim of this investigation is to create a hybrid deep learning framework, named DEA-RNN, for identifying instances of cyberbullying on social media platform Twitter. The research is aimed to assess the efficiency of the DEA-RNN model in comparison to other established algorithms.

The proposed model will be a good way to detect bullying on Twitter and has the potential to be extended to other social media platforms. Moreover, as with any machine learning model, there are potential limitations and ethical considerations that need to be addressed. Like the proposed model might fail to detect minor cyberbullying or might consider something like hard management as cyberbullying. This may cause consequences and it is very important to propose a transparent model and to investigate privacy very seriously while developing the model to avoid potential negative impacts. The proposed DEA-RNN model has the capability to aid in the identification and prevention of cyberbullying on social media platforms. However, further research and careful consideration of ethical principles are necessary to ensure its effectiveness and minimize any unintended consequences.

4.1 DEA-RNN Model

The DEA-RNN model is a combination of two main components: a Deep Embedding Attention (DEA) model and RNN. The DEA model is used to get information from the text data. It consists of multiple layers of BiLSTM and Attention Mechanism. The BiLSTM layers help to capture the sequential dependencies in the text data, while the Attention Mechanism helps to highlight the most relevant parts of the text. The output of the DEA model is a fixed-length vector that summarizes the input text.

The RNN component is used to model the temporal dependencies in the Twitter data. Specifically, the model uses a Bi-GRU) to capture the temporal dynamics of the Twitter data. The output of the RNN component is also a fixed-length vector that summarizes the temporal dynamics of the Twitter data. Finally, the outputs of the DEA and RNN components are fed in the model.

To train the DEA-RNN model, a large, annotated dataset of cyberbullying tweets is required. Supervised learning technique is used to train the DEA model on the annotated dataset and the unsupervised machine learning techniques is used to train RNN on the unannotated dataset. The DEA-RNN model combines deep learning and recurrent neural networks. So, it is a powerful approach for detecting cyberbullying on Twitter and it captures both the textual and temporal aspects of the Twitter data.

Current worldwide consideration to the challenges of natural assurance is constraining firms and governments to assess, rank, and select eco-efficient innovations. Innovations may devour inputs to create both alluring and undesirable yields. It appears that the data envelopment analysis (DEA) could be an appropriate strategy to assess eco-efficient advances. There are a few DEA expansions for dealing with undesirable yield, and now and then it is troublesome to select an appropriate demonstrate to assess the innovations. The challenge gets to be much more complex when the results of models are not comparable. In such a condition, subjective determination of elective DEA models may lead to deviation from an ideal choice.

4.2 RNN

RNN is a recurrent neural network, which is an artificial neural network that was developed to explicitly handle sequential data, such as time-series data or jobs involving natural language processing. These types of data can be processed in different methods. In RNN, a feedback loop is included. This feature can take the previous input and utilize it to inform the processing of the current input in a network. This is achieved by re-introducing the network's output from a time step earlier in the process as an additional input at the time step that is now being processed. Because of this, the RNN can construct a memory of the preceding inputs and the order in which they occurred. Because RNNs can capture long-term dependencies between the inputs, they are particularly useful for simulating the progression of sequential data. Because of this, they are ideally suited for activities such as music composition, speech recognition, and the processing of natural languages.

However, RNNs are susceptible to the vanishing gradient problem, which manifests itself when the gradients in the network become very insignificant as they propagate back through the feedback loop. This issue can only be solved by avoiding feedback loops entirely. Because of this, it may be challenging for the network to understand its long-term dependencies. In order to solve this issue, a variety of RNNs, such as the LSTM and (GRU networks, have been developed. These networks make use of gating methods to restrict the flow of information throughout the network and avoid the vanishing gradient problem.

In general, RNNs have shown themselves to be an effective tool in the field of machine learning, and this area of research and development remains very active.

In neural networks, recurrent systems have been considered difficult to study because the maximization process is NP-Hard, which applies to all neural networks. The problem is particularly challenging for recurrent systems because of the overall reliance issue. This issue refers to the problems of disappearing and bursting gradients that arise when propagating errors over numerous temporal phases.

Bidirectional RNNs (BRNN), which use data from both the past and the future to determine the outcome at any given time t, were introduced in the second article by Schuster and Paliwal. This differs from earlier networks where the outcome was only influenced by the input. BRNN has been successfully applied, among other things, to series labeling tasks in natural language processing.

One approach to further enhance the capabilities of RNNs is by combining them with Data Envelopment Analysis (DEA) for machine learning. DEA is a technique that is used to evaluate the relative efficiency of decision-making units, which can be adapted to work with RNNs in a number of ways. In addition, DEA can also help to identify the most important input variables, which can be useful for feature selection and data visualization. Another way of combining RNNs with DEA is to use DEA as a post-processing step, where the outputs of the RNN are evaluated using DEA to identify the most efficient outputs. This can be useful for problems where there are multiple output variables, as DEA can help to identify which outputs are most important for decision-making.

In general, the combination of RNNs and DEA has the potential to improve the performance of RNNs in learning solutions, by helping to minimize the dimensionality of the input space, identify important input and output variables, and evaluate the efficiency and performance of the RNN.

5 Data Analysis and Critical Discussion

5.1 Data Gathering

The input dataset consists of tweets that were spitting approximately 32 cyberbullying catchphrases and were obtained through the Twitter API. Some of the catchphrases recommended in brain research writing include bonehead, LGBTQ (gender ambiguous, transgender, and odd), prostitute, pussy, faggot, shit, sucker, skank, ass, live, on edge, simpleton, extortion, ambush, fuck, fucking, disgusting, bitch, ass, whale, etc. Other watchwords like "boycott," "kill," "kick the bucket," "beastly," "loathe," and "attack," "fear-based oppressor," "peril," "biased," and "dim" were suggested inside. The initial collection consists of 430,000 records, with watchwords based on extremism, indignation, swearing, and sexism contributing about 125,000 tweets each. The tweets in this dataset combine different exemptions. Because tweets in the English language are necessary, tweets including other tongue-in-cheek phrases are deleted, and retweets are modified. All of these shapes are created as part of the subsequent pre-processing system. The remaining essential pre-processing activities are then carried out. Information collection from Twitter is the method of recovering information from the social media stage Twitter. Twitter gives an endless source of information that can be utilized for different purposes, counting inquire about, assumption investigation, and promoting examination. There are a few ways to gather information from Twitter, and the strategy utilized depends on the particular needs of the analyst or investigator. One way to gather information from Twitter is to utilize the Twitter API. To utilize the Twitter API, one ought to apply for get to and get an API key, which permits get to the Twitter API. Once get to be allowed, the engineer can use the API to recover information such as tweets, retweets, likes, adherents, and more.

Another way to gather information from Twitter is to utilize web scratching instruments. Web scratching apparatuses are software programs that can consequently extricate information from web pages. To gather information from Twitter using web scratching apparatuses, one should recognize the net page containing the information and utilize the net scratching device to extricate the information from the internet page. This strategy requires a few programming aptitudes and may damage Twitter's terms of benefit, so caution is prompted. Information can too be collected from Twitter utilizing third-party devices and services. These apparatuses and administrations give get to Twitter information through an API or web interface, making it less demanding for analysts and investigators to gather and analyze Twitter information. A few well-known third-party devices and administrations for collecting Twitter information incorporate Hootsuite, Grow Social, and TweetDeck. Ready to conclude that information collection from Twitter requires an understanding of the accessible strategies and instruments and the needs of the analyst or investigator. With appropriate information collection methods, Twitter can be a fabulous source of information for a wide run of applications.

5.2 Data Annotation

This section devotes the majority of its attention to the process of annotating and classifying the tweets that were chosen from the initial dataset obtained from Twitter. To get things rolling, a random selection of the gathered tweets yielded the first 10,000 tweets. Following the selection of these tweets, a group of three human annotators difficultly doled out labels to each of them by hand over the course of one and a half months. The tweets were classified as either non-cyberbullying (category "0") or cyberbullying (category "1") depending on their substance. The human annotators made their labeling choices by carefully considering the point by point rules that were given to them.

The creators arbitrarily chosen 10,000 tweets for this reason, and each tweet was physically labeled into one of two categories: non-cyberbullying ("0") or cyberbullying ("1"). Three human annotators were included within the labeling prepare, which endured for one and a half months. The choice on whether a tweet included cyberbullying was based on particular rules is surveyed. Any errors found between the two annotators' names were settled by a third annotator. After settling the inconsistencies and cleaning up the information, the creators gotten a last dataset of 10,000 labeled tweets, out of which 6,508 (65%) were non-cyberbullying, and 3,492 (35%) were cyberbullying tweets. Be that as it may, the creators watched that the labeled Twitter dataset was imbalanced, with an altogether bigger number of non-cyberbullying tweets compared to cyberbullying tweets. To address this awkwardness, the creators utilized adjusting procedures such as oversampling or under-sampling.

DEAR_RNN (Deep Evolutionary Attention-based Recurrent Neural Network) is a machine learning model that is designed to identify cyberbullying in tweets. Data annotation is a crucial step in training the DEAR_RNN model. In data annotation, human annotators manually label the selected tweets as cyberbullying or non-cyberbullying. In the case of DEAR_RNN, the annotators were instructed to label the tweets into two categories, "0" for non-cyberbullying and "1" for cyberbullying. The process of data annotation for DEAR_RNN began with randomly selecting 10,000 tweets from the Twitter dataset. These tweets were then labeled manually by three human annotators over a period of one and a half months.

A third annotator was brought in to reconcile any inconsistencies that were found during the initial phase of the annotation process. Following the resolution of the inconsistencies, the final dataset had a total of 10,000 tweets with associated labels. Out of these, 6,508 (65%) were categorized as tweets which did not include cyberbullying, while 3,492 (35%) were classified as tweets that included cyberbullying. The labeled tweets from Twitter had an uneven conveyance, with a more noteworthy number of tweets that did not include cyberbullying than tweets that did include cyberbullying. Oversampling was utilized in arrange to address the issue of unequal representation of classes, and the Engineered Minority Oversampling Strategy (Destroyed) was connected in arrange to oversample the minority course (cyberbullying tweets). In order to ensure that the dataset was representative of the whole, the oversampling method consisted of repeatedly reproducing the cyberbullying samples. Following the application of the oversampling technique, the total number of tweets that were examined was 13,016. After that, the DEAR_RNN model was trained with the help of this balanced dataset.

5.3 Data Pre-processing

After the information collection and comment stages, the following step is to pre-process and clean the information to get ready it for include extraction and determination. Pre-processing and cleaning are basic steps in any machine learning assignment as they guarantee that the information is in a reasonable organize and free from mistakes, irregularities, and commotion. The DEAR_RNN data pre-processing and cleaning phase involve several sub-tasks, including:

Tokenization—The first step in data pre-processing is to split the tweets into individual words or tokens. The tokens can be obtained using various methods such as whitespace separation, punctuation separation, and regular expressions.

Spell checking—This is important because misspelled words can negatively affect the accuracy of the subsequent tasks.

Removal of URLs, user mentions, and hashtags:—URLs, user mentions, and hashtags are not relevant to the analysis of the text and can be removed to reduce noise in the dataset.

Cleaning of special characters and punctuation:—Special characters and punctuation such as emojis, quotation marks, and brackets can also be removed to simplify the data and reduce noise.

Encoding—Finally, the pre-processed and cleaned data is encoded into a numerical format that can be used for feature extraction and selection. This is typically done using techniques such as one-hot encoding, word embedding, or bag-of-words.

In general, the pre-processing and cleaning phase in DEAR_RNN is crucial for obtaining accurate and reliable results in subsequent tasks.

In DEAR_RNN, the data pre-processing and cleaning phase can be split into three sub-phases to make sure that the raw tweet dataset is cleaned up and turned into a final dataset. In the first sub-phase, noise like URLs, hashtags, mentions, punctuation, symbols, and emoticons are taken out. OOV words are cleaned up by doing things like spell checking, expanding acronyms, changing slang, and removing characters that are too long. The final sub-phase is made up of changes to tweets, such as changing the text to lowercase, stemming, word segmentation (tokenization), and taking out stop words. These smaller steps are needed to improve the quality of the tweets, get more accurate feature extraction and classification, and improve the quality of the tweets. By cleaning and pre-processing the data in this way, the DEAR_RNN model can work with it better, which leads to more accurate predictions.

5.4 Feature Extraction and Selection

Feature extraction and selection is a crucial step in the DEAR_RNN model as it aims to extract relevant features from the pre-processed tweet data that will aid in the classification of cyberbullying and non-cyberbullying tweets. During this phase of the process, which is known as the feature extraction phase, several kinds of features are retrieved from the twitter data that has previously been processed. Word-level features, character-level features, and semantic-level features are all included in these features. The use of n-grams and character embeddings are both utilized in the process of character feature extraction. Techniques such as topic modeling and word embeddings are utilized in the process of extracting information on a semantic level.

After the features have been retrieved, the next step is to use feature selection techniques in order to determine which characteristics are the most important and how they contribute to the classification process. Techniques for selecting features have the goals of reducing the dimensionality of the feature space, cutting down on noise. The method known as Data Envelopment Analysis (DEA) is employed inside the framework of the DEAR_RNN model in order to carry out the process of feature selection. In the DEAR_RNN model, each tweet is treated as a DMU, and the features are treated as inputs and outputs.

Extracting features in a Twitter dataset involves transforming the raw text data into numerical features that can be used for machine learning models. This process involves several steps:

Tokenization—This is the process of splitting the text into individual words, called tokens. This is usually done by splitting the text on whitespace, punctuation, and other delimiters. For example, the sentence “Hello, world!” would be tokenized into two tokens, “Hello” and “world”.

Stop word removal—Stop words are common words that do not carry much meaning, such as “the”, “and”, and “of”. These words are often removed to reduce the dimensionality of the feature space and improve performance.

Vectorization—This is the process of converting the tokenized and preprocessed text into a numerical representation that can be used for machine learning. There are several methods for vectorizing text, including bag-of-words, TF-IDF, and word embeddings.

Feature selection—This is the process of selecting a subset of the extracted features that are most relevant for the machine learning task. This can be done using various techniques, such as chi-squared test, mutual information, and feature importance scores. In a Twitter dataset, the above steps are applied to the tweets in the dataset to extract relevant features for machine learning models. The features can include various aspects of the tweets, such as the presence of certain words or hashtags, the sentiment of the tweet, the user who posted the tweet, and so on. These features can be used to train machine learning models to perform tasks such as sentiment analysis, topic classification, and cyberbullying detection.

In the context of a Twitter dataset, feature extraction refers to the process of identifying and selecting relevant pieces of information that can be used to classify tweets as cyberbullying or non-cyberbullying.

In addition to the above, POS tags can be used to identify specific types of words (e.g., nouns, verbs, adjectives) and their relationships within a sentence. There are various feature selection methods available in the literature, and one of the most commonly used methods is the Information Gain (IG) method. This method ranks the features based on their ability to provide the most significant amount of information about the classification problem. The most prominent features are then selected and fed into the DEA-RNN classifier.

Generally, feature extraction in a Twitter dataset involves identifying and selecting relevant pieces of information from text data using NLP tools. The selected features can be further enhanced by considering POS tags, function words, and content word features. Finally, feature selection methods such as Information Gain can be used to identify the most significant features for the classification task.

5.5 Data Analysis

In this section, we will detail how we evaluated the efficacy of the DEA-RNN model by using datasets acquired from Twitter. Specifically, we will focus on the evaluation process. We utilized a number of different models.

We additionally detail the methods that were utilized in order to annotate the data that was included in the input dataset. We chose two deep learning-based models (RNN and Bi-LSTM) and three other models from the current research on cyberbullying detection in social media to compare the performance of our proposed model with that of existing models. We found that our proposed model performed better than the existing models. We used the configuration settings of these baseline models exactly as they were described in the original papers they came from.

Python version 3.7.4 and the Pycharm IDE version 2020.2.3 were utilized for the purposes of our studies, in addition to a number of different libraries.

A number of criteria are used to assess the proposed DEA-RNN model, and it is compared to other deep learning and machine learning-based models. On a personal computer, the tests are carried out using common libraries and tools. Preprocessing operations are performed on the input dataset using the NLTK Python module in accordance with the methodology proposed in this study. Following that, the dataset is divided into training and testing subsets. The dataset is divided into three scenarios: 65:435% (Scenario 1), 72:28% (Scenario 2), and 93:7% (Scenario 3), in order to assess the effectiveness of tweet classification using each approach. The best performance of each approach is shown by the evaluation measures that are chosen. Each approach is used 20 times, the results are averaged, and then each evaluation metric is given a more accurate average value. The fivefold cross-validation method is also used.

An overview of the assessment measures that were utilized to gauge the effectiveness of the proposed model is given in this section. Each approach in this study is tried 20 times across all tests to ensure accurate results, and the average of the outcomes for each assessment metric is calculated.

In this segment, the DEA-RNN classifier's experimental findings are appeared, at the side comparisons to other profound learning models just like the Bi-LSTM and RNN, as well as machine learning models just like the MNB, RF, and SVM. In the tests, three alternative input dataset scenarios—Scenario 1 (65:35%), Scenario 2 (72:28%), and Scenario 3 (93:7%)—were used to assess the classifier's capacity to predict cyberbullying. In order to generate an average of the results, experiments were ran 20 times for each classifier and dataset input scenario. Each model's performance was assessed using a variety of criteria.

The study compared the DEA-RNN model's accuracy to that of other current models. The authors calculated the average accuracy for each scenario and discovered that Scenario 3's DEA-RNN had the highest average accuracy, 90%. The accuracy ratings for the other models were, respectively, 83.45%, 80%, 77%, 64%, and 75%. In scenario 1, the DEA-RNN demonstration achieved an accuracy rate of 82%, outperforming the taken into consideration current models. Among all the models, Bi-LSTM had the second-highest score, however MNB had the execution result that was the worst possible.

Therefore, the accuracy of the proposed model and alternative approaches was best in Scenario 3.

Figure 1 shows the performance evaluation based on accuracy.

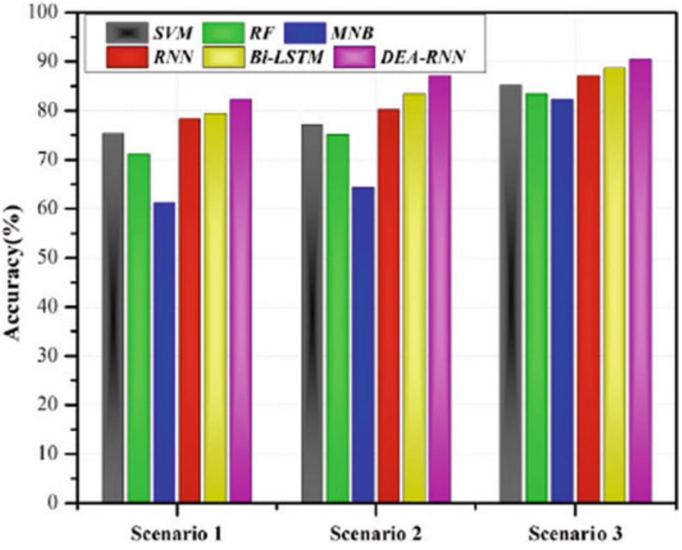

Fig. 1 Accuracy based evaluation

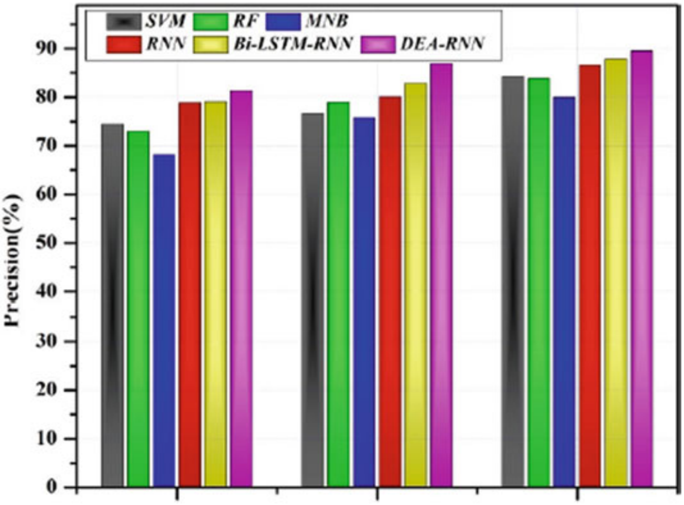

Fig. 2 Precision based evaluation

Comparing the DEA-RNN model's performance against that of other models already in use. In scenario 3, the accuracy of the DEA-RNN was 89.52%, compared to 87.9%, 86.62%, 84.25%, 80.01%, and 83.87% for others. In scenario 2, the DEA-RNN outperformed the other existing models, which attained precision values of 82.88%, 80.09%, 76.6%, 75.78%, and 78.96%, respectively. The DEA-RNN's best precision value was 87.02%. Bi-LSTM achieved the second-highest precision score among all the models in scenarios 2 and 3, while MNB had the lowest performance outcomes (Fig. 2).

In this part we tested a new model called DEA-RNN to see if it could do a better job than other computer programs at detecting cyberbullying on Twitter. They compared it to other models and tested it with different sets of data, using measures like accuracy, recall, precision, F-measure, and specificity to see how well it did. They found that the DEA-RNN model was the most accurate, with an average accuracy of 90.45% in Scenario 3, and it did better than other models like Bi-LSTM and RNN. The deep learning models worked better than the machine learning models, and the MNB model did the worst. Overall, the DEA-RNN model looked really good for detecting cyberbullying on Twitter.

6 Conclusion and Future Work

In light of the discoveries of this consideration, there are a number of distinctive suggestions and regions for future investigate that may well be taken into thought. The investigate of designs of cyberbullying on other social media stages, such as

Instagram, Facebook, YouTube, and Flickr, is one conceivable road for advance request that might be sought after. This will provide a more comprehensive picture of the patterns of cyberbullying over an assortment of stages, and it'll help find commonalities and refinements.

Utilizing information from various sources in arrange to distinguish occurrences of cyberbullying is another curiously range of request. In expansion, the substance of tweets isn't the as it were thing that future inquire about might center on; it can too examine the behavior of Twitter clients. This will give a more comprehensive understanding of the marvel of cyberbullying on Twitter and permit for the creation of cures that are more successful. In expansion, the inquire about found that the DEA-RNN demonstrate was fruitful in distinguishing occurrences of cyberbullying on Twitter based exclusively on the substance of the tweets themselves. Be that as it may, cyberbullying can moreover take put through other shapes of media, such as sound and video recordings and still photos. As a result, the recognizable proof of cyberbullying in these distinctive media sorts should be the subject of encourage investigate within the future.

The investigate proposes that real-time checking and intercession may be a strategy that has potential within the battle against cyberbullying. Tweets that constitute cyberbullying can be categorized and distinguished in genuine time with the assistance of the show that was proposed, which clears the way for mediations that are incite and effective. The experiments had been performed on distinctive input dataset situations and evaluated the use of several metrics such as accuracy, don't forget, precision, F-degree, and specificity. The consequences showed that the DEA-RNN model outperformed other fashions in terms of accuracy, attaining the best average accuracy of ninety.45% in scenario 3, followed by Bi-LSTM and RNN. The deep studying fashions completed higher than the gadget getting to know fashions, and MNB had the worst overall performance among all the fashions. The proposed version and different methods finished optimally in state of affairs three in terms of accuracy. General, the DEA-RNN model showed promising results for cyberbullying detection on Twitter.

From this study we can see that, Twitter clients can annoy one another, call one other names, spread rumors, or by and large lies, and send messages of terrorizing. These are all illustrations of cyberbullying. Cyberbullying can be coordinated at anybody, counting open figures, writers, activists, lawmakers, and indeed normal clients and celebrities.

The usage of mysterious accounts or imposter profiles may be a drift that has been watched in connection to cyberbullying on Twitter. These accounts are made in arrange to conceal the user's personality, which makes it much less difficult to irritate and bully other individuals without fear of the results of one's activities. Since of this namelessness, there's a more prominent potential for the spread of despise discourse, prejudice, and other sorts of bias.

Another later advancement is the utilize of hashtags, or "hashtags," to spread despise discourse or single out certain people or organizations. Individuals have utilized hashtags like #CancelCulture and #MeToo to irritate and scare others online, which has driven to a poisonous climate on Twitter.

In expansion, the utilize of bots and other shapes of mechanized account can intensify the impacts of cyberbullying that happens on Twitter. These accounts are pre-set to spread messages of contempt and amplify the reach of cyberbullying, both of which make it more troublesome to screen and combat the marvel. Twitter has, in response to these advancements, made an assortment of arrangements and instruments to avoid cyberbullying. These approaches and devices incorporate channels for announcing injurious substance, channels, and calculations implied to distinguish and expel such substance. However, cyberbullying proceeds to be an unavoidable issue on the stage, and on the off chance that it is to be successfully tended to, it'll require a concerted exertion effort exertion on the portion of clients, administrators, and social media firms.

References

1. Al Duhayyim M, Mohamed HG, Alotaibi SS, Mahgoub H, Mohamed A, Motwakel A, Eldesouki M (2022) Hyperparameter tuned deep learning enabled cyberbullying classification in social media. Comput Mater Contin 73:5011–5024
2. Al-Garadi MA, Hussain MR, Khan N, Murtaza G, Nweke HF, Ali I, Mujtaba G, Chiroma H, Khattak HA, Gani A (2019) Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges. IEEE Access 7: 70701–70718
3. Altay EV, Alatas B (2018) Detection of cyberbullying in social networks using machine learning methods. In: 2018 international congress on big data, deep learning and fighting cyber terrorism (IBIGDELFT). IEEE, pp 87–91
4. Dhyani M, Kumar R (2021) An intelligent chatbot using deep learning with bidirectional RNN and attention model. Mater Today: Proc 34:817–824
5. Haidar B, Chamoun M, Serhrouchni A (2019) Arabic cyberbullying detection: Enhancing performance by using ensemble machine learning. In: 2019 international conference on internet of things (ithings) and IEEE green computing and communications (greencom) and IEEE cyber, physical and social computing (cpscom) and IEEE smart data (smartdata). IEEE, pp 323–327
6. Jeevan RK, Venu Madhava Rao SP, Kumar PS, Srivikas M (2019) EEG-based emotion recognition using LSTM-RNN machine learning algorithm. In: 2019 1st international conference on innovations in information and communication technology (ICIICT). IEEE, pp 1–4
7. Liu H, Lang B, Liu M, Yan H (2019) CNN and RNN based payload classification methods for attack detection. Knowl-Based Syst 163:332–341
8. Muaad AY, Kumar GH, Hanumanthappa J, Benifa JB, Mourya MN, Chola C, Pramodha M, Bhairava R (2022) An effective approach for Arabic document classification using machine learning. Global Trans Proc 3(1): 267–271
9. Muller HA, Zavolokina L (2021) Artificial intelligence for combating cyberbullying on social media platforms: a systematic review. IEEE Trans Comput Soc Syst 8(4):984–999
10. Murnion S, Buchanan WJ, Smales A, Russell G (2018) Machine learning and semantic analysis of in-game chat for cyberbullying. Comput Secur 76:197–213
11. Murshed BAH, Abawajy J, Mallappa S, Saif MAN, Al-Ariki HDE (2022) DEA-RNN: a hybrid deep learning approach for cyberbullying detection in Twitter social media platform. IEEE Access 10:25857–25871
12. Paullada A, Raji ID, Bender EM, Denton E, Hanna A (2021) Data and its (dis) contents: a survey of dataset development and use in machine learning research. Patterns 2(11):100336
13. Perez C, Karmakar S (2023) An NLP-assisted Bayesian time-series analysis for prevalence of Twitter cyberbullying during the COVID-19 pandemic. Soc Netw Anal Min 13(1):51

14. Raiyani K, Gonçalves T, Quaresma P, Nogueira V (2018) Fully connected neural network with advance preprocessor to identify aggression over Facebook and Twitter
15. Yang C, Jiang W, Guo Z (2019) Time series data classification based on dual path CNN-RNN cascade network. IEEE Access 7:155304–155312
16. Yang L, Shami A (2020) On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing 415:295–316
17. Zeng Z, Li Y, Li Y, Luo Y (2022) Statistical and machine learning methods for spatially resolved transcriptomics data analysis. Genome Biol 23(1):1–23