IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 38, No. 3, June, 2024, pp. 837-846
Journal homepage: http://iieta.org/journals/ria

Neural Network-Based Cyber-Bullying and Cyber-Aggression Detection Using Twitter(X) Text

Michael Agbaje*, Oreoluwa Afolabi

Department of Computer Science, Babcock University, Illishan-Remo 121103, Ogun State, Nigeria

Corresponding Author Email: agbajem@babcock.edu.ng

https://doi.org/10.18280/ria.380310

Received: 14 November 2023
Revised: 12 March 2024
Accepted: 25 May 2024
Available online: 21 June 2024

Keywords:
cyberbullying, cyberaggression, sentiment analysis, machine learning, deep learning, CNN, RNN

ABSTRACT

Cuber-bullying and Cuber-aggression analysis and detection form a critical aspect in today's social networking research being driven by machine learning and deep learning techniques. Quite a number of several techniques have been proposed to measure and detect unpleasant or offensive content or behavior on different social platforms over the years; this includes chat analysis of speech and fuzzy logic, natural language processing (NLP) among others. In this research, sentiment analysis of tweeter text using hybrid neural network to detect cyber-bullying and cyber-aggression from users' tweets on twitter and google images was developed with RNN for the text analysis and CNN for image analysis. The proposed methodology exploits sentiment analysis features like polarity of text, specific feelings and emotions, intentions and images to determine bullying or aggression. The text data is generated from Twitter through Twitter API and images from google. Deep learning models (RNN and CNN) continue to demonstrate their potential in the task of predictive analytics to tackle data analysis and learning challenges as the size of data grows and the need for a quick and accurate result grows. The efficiency and performance of models were evaluated with RNN and CNN outperforming the other classification algorithms achieving accuracy of 0.951 and 0.911 also, F-Measure scores were 0.910 and 0.890 showing impressive performance in the analysis and detection in text and images. This research provides a substantial contribution to cyberbullying and cyberaggression analysis and detection mechanism by tapping into Twitter users' psychological features including personalities, sentiment and emotion. domain by harnessing the potentials deep learning.

1. INTRODUCTION

The availability of internet has brought about social networking. The various social media platforms existing has contributed to the world becoming a global village. People of all demographics are affected by the concerning trends of cyberbullying and cyberaggression. On a global scale more than half of the youth who use social media have experienced this kind of ongoing or coordinated online harassment. A wide range of emotions can be experienced by victims, and many of these can have detrimental effects including sadness, humiliation, and social isolation. These effects increase the likelihood of even more serious outcomes like suicide attempts [1-3]. The abusive conduct in online settings such as Cyberbullying and cyberagression describes the unfavorable effects of improper use of online communication [4]. The research on the detection of cyberbullying events has been ongoing for a decade and continues to progress. Previous studies primarily concentrate on the detection of textual-based cyberbullying, which is typically treated as a binary classification problem. Nonetheless, the field of cyberbullying research has evolved to encompass diverse classification problems. The most common type of bullying behavior tends to be name-calling. Cyber aggression is a negative impact; it is defined as the willful use of information technology to harm, threaten, slander, defame, or harass another person. Among social media companies, the platforms with the highest prevalence of cyberbullying include: Facebook (75%), Instagram (24%), Twitter (24%) and so on [5].

Empirical evidences linking users' psychological features such as personality traits and cybercrimes such as cyberbullying are many. This study deals with automatic cyberbullying detection mechanism tapping into Twitter users' psychological features including personalities, sentiment and emotion. Sentiment Analysis (SA) uses features like polarity of text, specific feelings and emotions, intentions and images to determine bullying or aggression [6]. SA computational technique can help study people's attitude, views and opinion over the web for different entities. SA uses social media as its important source of information because it continuously expands and breeds consistent and multifaceted information. Across the social networking sites, the surge for user generated content as drastically increased. And among all, twitter is the social network platform that records huge volume of opinionated information [7, 8].

Recognizing a person's true view from a piece of text, his/her speech, facial expressions, videos, or image might be challenging at times, however automatic sentiment recognition by a computer could help solved the problem. On an average, 3.6 million tweets are tweeted on twitter every

minute. So, the more volume of data poses a problem to data analysts to draw an opinion out of it. Intonation, facial expression, thought and opinion in text, video or image, can sometimes be very difficult to encode. Bottleneck in the process of extracting frames, audio, text obtained from videos, textual information and images can make analysis impossible or very tedious. Also, as a large collection of people's opinions on the internet keeps increasing, drawing conclusion from an unstructured data can be quite tasking. Cyberbullying is a heart-wrenching phenomenon because the majority of young lives that were cut short could have been saved if the relevant stakeholders, parents, schools, internet intermediaries and governments were to take appropriate steps to fight against it.

The aim of this study is to implement a users' tweet monitoring system for detecting cyber-bullying and cyber-aggression. The main contributions of the study are as follows:

·We use Sentiment Analysis (SA) features like polarity of text, specific feelings and emotions, intentions and images to determine bullying or aggression.

·This study contributes to the field by meticulously designing and evaluating a hybrid model, considering the diverse and dynamic nature of text data in social media networks.

·We use a recurrent neural network to look for word relationships and a convolutional neural network for opinion prediction in videos and images, spotting them with aggressive or bully views.

·We proposed hybridized Convolutional Neural Network with LSTM-RNN for a better accuracy and optimal performance.

The CNN enables images/videos to be digitally processed thereby giving the system to mine textual information from both videos and images, which may then be passed to the LSTM-RNN network, which is mainly built to handle textual tweets. The CNN is used to extract complex features from images, and LSTM is used as a classifier. CNN has the ability to process visual, textual, and audio data. The LSTM was added to the Recurrent Neural Network in order to solve the problem of texts with long time dependencies usually referred to as vanishing gradient. Performance evaluation shows Deep Neural Network (DNN) outperformed the other classification algorithms.

The work structure is organized as follows: Section 2 presents the literature review, Section 3 describes the methodology used, Section 4 presents the results and discussion, and finally, Section 5 concludes the study.

2. LITERATURE REVIEW

2.1 Cyberbully vs cyberaggression vs hate speech

It is argued that cyberbullying has some of the characteristics of hate speech and that many of the tools used to fight against hate may be utilised to fight against cyberbullying.

Cyberbully, a phenomenon formally identified in 2003, has been posing a severe threat to the vastly growing social media platforms in modern age of internet and digital connect. The first work on automated cyberbully detection was done in 2006 [9].

Cyberbullying is becoming common in inflicting harm on others, especially among adolescents [10]. Cyberbullying refers to the act of harming, intimidating, or attempting to coerce someone online. Unlike traditional bullying, which involves both physical and verbal abuse, cyberbullying is purely verbal due to its digital medium. Examples include harassing messages, threats, and online intimidation and so on. Victims of cyberbullying can suffer serious mental, social, and physical health consequences. Figure 1 shows the social media platform showing how people are bullied. According to tweeter statistics, 2019 was recorded to have the highest number of cyber bullying.

Cyber-aggression encompasses a wider range of online behaviors, including both repeated acts and one-off situations. Unlike cyberbullying, the intent to harm may not always be present. Examples include liking a negative comment without realizing its impact, participating in a harmful conversation, or sharing hurtful content [11].

Hate speech is defined as a bias-motivated, hostile, malicious speech aimed at a person or a group of people because of some of their actual or perceived innate characteristics [12] Hate speech expresses discriminatory, intimidating, disapproving, antagonistic and/or prejudicial attitudes toward these characteristics, which include sex, race, religion, ethnicity, colour, national origin, disability, or sexual orientation or gender. It is typically verbal abuse that goes beyond mere disagreement or criticism [13].

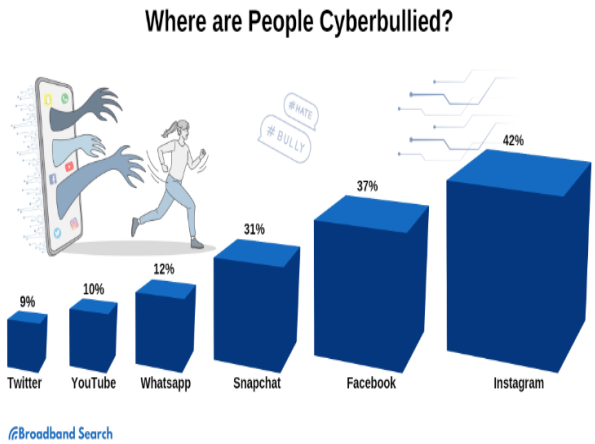

Figure 1. Where are people cyberbullied [14]

The inspiration of having biological process of neural network a supervised machine learning method, dominating all kinds of applications and proving great advantage over conventional programming, has brought about Deep learning. A deep learning technique allows direct learning from the data for all aspects of the model. It starts with the lowest level attributes, which provide an accurate representation of the data. When the amount of training data is increased, deep learning becomes more useful. Because of the growing magnitude of data and breakthroughs in the field of big data, traditional machine learning algorithms have revealed their limitations in analyzing large datasets. In the purpose of analysis, deep learning algorithms have produced superior results. It has made a distinction between machine learning techniques and traditional algorithms by taking advantage of more human brain competence [15]. Machine learning is an approach that enables algorithms to detect complex patterns and relationships from training data. By feeding these algorithms with numerous annotated text or images they can learn to recognize the distinctive features of different types The CNN

and the RNN Unit network could be used to extract the main sentiment features [16]. The neural network model is usually composed of multiple hierarchies, which can be a layer-by-layer abstraction, and the layers can be mapped by nonlinear activation functions, so it can fit very complex features and learn the hidden deep features between texts [17]. The deep learning models commonly used in the field of text sentiment analysis are CNN, Recurrent Neural Network (RNN), LSTM, and Gated Recurrent Unit (GRU) [18]. Others are Support Vector Machine (SVM), Random Forest and Naïve Bayes.

The inspiration of having biological process of neural network a supervised machine learning method, dominating all kinds of applications and proving great advantage over conventional programming, has brought about Deep learning. A deep learning technique allows direct learning from the data for all aspects of the model. It starts with the lowest level attributes, which provide an accurate representation of the data. When the amount of training data is increased, deep learning becomes more useful. Because of the growing magnitude of data and breakthroughs in the field of big data, traditional machine learning algorithms have revealed their limitations in analyzing large datasets. In the purpose of analysis, deep learning algorithms have produced superior results. It has made a distinction between machine learning techniques and traditional algorithms by taking advantage of more human brain competence [15]. Machine learning is an approach that enables algorithms to detect complex patterns and relationships from training data. By feeding these algorithms with numerous annotated text or images they can learn to recognize the distinctive features of different types The CNN and the RNN Unit network could be used to extract the main sentiment features [16]. The neural network model is usually composed of multiple hierarchies, which can be a layer-by-layer abstraction, and the layers can be mapped by nonlinear activation functions, so it can fit very complex features and learn the hidden deep features between texts [17]. The deep learning models commonly used in the field of text sentiment analysis are CNN, Recurrent Neural Network (RNN), LSTM, and Gated Recurrent Unit (GRU) [18]. Others are Support Vector Machine (SVM), Random Forest and Naïve Bayes.

2.2 Related works

This study [19] built a model for detecting Cyber-bullying and Cyber-aggression in social media using the Machine Learning Algorithm. Deep neural networks, as well as probabilistic, tree- based, and ensemble classifiers like Naive Bayes and R in terms of performance and training time. The gap of the study is that system did not provide detect abusive behaviour in real time. The study [20] provided a collaborative approach for detecting cyber-bullying in tweets using different distributed collaboration patterns. With the aid of Support Vector Machine, Naive, and Maximum Entropy (Logistics) machine learning algorithms, the model was trained to create models that were used for the classification of cyber-bullying tweets. Various experiments carried out indicate that the collaborative approach performs better than the standalone approach. There was an overhead that caused the classification time to increase as number of nodes in the network increased. This study [21] used NLP and Machine Learning in detecting Cyber-bullying and Aggression. Random Forest, strive to determine the most optimal in terms of performance and training time. The gap of the study is that system did not provide detect abusive behaviour in real time. This study [22] developed an automatic Cyber-bullying detection system in Spanish-language social network. Researchers recommended that for improved performance, adding words from other nations should be considered to the word exchange. The study [23] developed a new approach to detect cyber-bullying on Twitter using deep learning called optimized Twitter cyber-bullying detection (OCDD). Where hard task features are extracted from tweets and feed them to a classifier, and tweets are represented as a set of word vectors which captures the semantic of words and CNN which classifies tweets in a more intelligent way than traditional classification algorithms. The gap of this study is that cyberbully was not evaluated within detection contexts. This study [24] built a model to detect suspicious posts on online forums using Machine Learning in Text mining. No source of data stated and model was not adequately trained to handle a range of sentiments. This study [25] built a model to detect the presence of Cyber-bullying using Machine Learning approach. Enhancing the detection of Cyber-bullying by combining texts with videos and images was recommended by the researcher. This study [26] proposed the Multinominal Naive Bayes Classifier for detecting Cyber-bullying Tweets. The gap of this work, is that, the researcher didn't do analysis on the topic modeling. The study [27] developed a model to test the performance of machine learning approaches in Cyber-bullying detection from social media using sentiments. The researchers suggested that other models or algorithms, aside from Nave Bayes, Support Vector Machine (SVM), and k-Nearest Neighbor (k-NN), can be used in investigated in order to obtain more accurate and efficient findings. This research [28] worked on Cyber-bullying Detection by Sentiment Analysis of Tweets' Contents Written in Arabic in Saudi Arabia Society. Hybrid (i.e., both Machine Learning & Lexicon-Based) was used for the methodology.

The study [29] proposed DNN algorithm uses different symbols that can be defined as: FE: emotional features, tx: one single tweet, Pt: pre-processed text, Fm: feature embedding matrix, Wtokens: word tokens C f: combined feature. The proposed DNN model was tested on the Cyber-Troll dataset. The combination of word embedding and eight different emotional features were fed into the DNN for significant improvement in recognition while keeping the DNN design simple and computationally less demanding. When compared with the state-of-the-art models, our proposed model achieves an F1 score of 97%, surpassing the competitors by a significant margin. Their research focused on the detection of cyber aggression using deep learning models. The framework for aggression detection is based on the combination of novel emotional features and Word2Vec features.

This study [30] presents a unique Hybrid Deep Learning Architecture (HDLA)-based method for unfriendly language identification on OSNs. By combining Long Short-Term Memory (LSTM) networks with Convolutional Neural Networks (CNNs), it leverages the advantages of both approaches. The CNN part is good at hierarchically extracting spatial characteristics from text data; it can spot objectionable patterns that are typically hidden in the subtle structure details. On the other hand, the LSTM part is good at processing sequential data; it can gradually extract the contextual relationships in user posts. The effectiveness of the suggested methodology shows itself to be an appropriate approach for locating instances of cyberbullying on social media sites. Moreover, compared to all performance standards, the deep neural network that was developed performs the best, especially when it comes to identifying cases of cyberbullying.

The focus of the research is on deep learning models, specifically BiLSTM, and how efficiently they perform in this field compared to traditional machine learning algorithms. The higher performance metrics of the BiLSTM model illustrate a paradigm change in computational linguistics by emphasizing the growing importance of models that comprehend the nuances of language and context.

In this study, a stacked ensemble approach is proposed that accurately predicts Arabic sentiment by leveraging the predictive capabilities of CNN, Bi-GRU, Bi-LSTM, and hybrid DL models (CNN-Bi-GRU and CNN-Bi-LSTM) [31]. The proposed model's efficacy is evaluated using four extensive datasets: the HARD dataset, the BRAD dataset, the ARD dataset, and a real dataset composed of 71,583 Arabic reviews. Experimental results demonstrate the suitability of the proposed model for analyzing sentiments in Arabic texts. The method's first step involves feature extraction using the AraBERT model. Subsequently, five DL models are developed and trained, including CNN, Bi-GRU, Bi-LSTM, a hybrid CNN-Bi-GRU model, and a hybrid CNNLSTM model. Finally, the outputs of the base classifiers are concatenated using the multilayer perceptron algorithm. Our approach achieves an improved accuracy of 0.9256 compared to basic and hybrid deep learning methods.

For better detection performance, it proposes a hybrid strategy that makes use of the advantages of both deep learning and conventional machine learning [32]. On the image dataset, this study's state-of-the-art accuracy was 82%. In order to enhance cyberbullying detection, this study suggests a novel hybrid technique that makes use of machine learning classifiers and deep learning models as feature extractors. Understanding the intricate circumstances in which cyberbullying takes place is improved by extracting features using trained deep learning models, such as InceptionV3, ResNet50, and VGG16, and then feeding those features into classifiers like Logistic Regression and Support Vector Machines.

In order to evaluate machine learning performance, this study employs the convolutional neural network (CNN), long short-term memory (LSTM), CNN-LSTM, and LSTM-CNN models in a deep learning framework utilizing bidirectional encoder representations from transformers (BERT) data representation [33]. The effectiveness of the BERT-based deep learning model was tested in order to examine how social media users felt about Indonesia's 2024 presidential election. BERT is used with CNN, LSTM, CNN-LSTM, and LSTM-CNN models in the text representation approach. It may be inferred from the case study that the deep learning model enhances the BERT model's sentiment analysis performance.

The diverse methods currently employed for cyberbullying detection, including Natural Language Processing (NLP), Machine Learning (ML), Deep learning (DL) and Transfer Learning (TL)techniques.

The researchers suggested that for considerably better outcomes, neural networks or deep learning algorithms could be used.

3. METHODOLOGY

Our experiments were performed on a Windows 11pro computer with a Hp NoteBook 348 G5 Intel Core I5-16GB RAM/256GB SSD. Models were configured in Pycham using Keras API version 2.4.3 with Tensorflow version 2.4.

3.1 Evaluation metrics

We evaluate our model using the average accuracy, recall, precision and F1-score. Calculation of these measures was done using true positive (TP), false positive (FP), true negative (TN) and false negative (FN). Correctly classified cyberbullying tweets are called true positive(s) (TP) while FN is incorrectly as non-cyber-bullying tweets. Tweets correctly classified as non-cyber-bullying TN and those incorrectly classified as cyber-bullying are called FP.

	A	B	C	D	E	F
1	polarity	id	date	query	user	text
2	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by texting it... and might cry as a result Schoc
3	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Managed to save 50% The rest go out of bo
4	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
5	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all. I'm mad. why am i here? because I can't s
6	0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew
7	0	1467811592	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	mybirch	Need a hug
8	0	1467811594	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	COZZ	@LOLTrish hey long time no see! Yes.. Rains a bit, only a bit LOL, I'm fine thanks, how'
9	0	1467811795	Mon Apr 06 22:20:05 PDT 2009	NO_QUERY	2Hood4Hollywood	@Tatiana_K nope they didn't have it
10	0	1467812025	Mon Apr 06 22:20:09 PDT 2009	NO_QUERY	mimismo	@twittera que me muera?
11	0	1467812416	Mon Apr 06 22:20:16 PDT 2009	NO_QUERY	erinx3leannexo	spring break in plain city... it's snowing
12	0	1467812579	Mon Apr 06 22:20:17 PDT 2009	NO_QUERY	pardonlauren	I just re-pierced my ears
13	0	1467812723	Mon Apr 06 22:20:19 PDT 2009	NO_QUERY	TLEC	@caregiving I couldn't bear to watch it. And I thought the U A loss was embarrassing.
14	0	1467812771	Mon Apr 06 22:20:19 PDT 2009	NO_QUERY	robrobbierobert	@octolinz16 It it counts, idk why I did either. you never talk to me anymore
15	0	1467812784	Mon Apr 06 22:20:20 PDT 2009	NO_QUERY	bayofwolves	@smarrison i would've been the first, but i didn't have a gun. not really though, zac snyd
16	0					

Figure 2. Text attributes in tweet

$$\text{Accuracy} = ((TP+TN))/((TP+FN+TN+FP)) \quad (1)$$

$$\text{Precision} = TP/((TP+FP)) \quad (2)$$

Precision measures the number of correctly identified tweets among all tweets labelled a cyber-bullying.

The recall is the number of bullying tweets among all the

tweets in the dataset.

$$\text{Recall} = TP/((TP+FN)) \quad (3)$$

F1-score is a measure of how well your classifier balances precision and recall.

$$\text{F1-score} = (2*(P*R))/((P+R)) \quad (4)$$

3.2 Dataset

In this research, the dataset is a collection of tweets from twitter domain that have been trained and well labelled. Sentiment140 dataset was used for this project, which contains over 1 million well annotated tweets extracted from Twitter website using Twitter API, from the 2019. The dataset contains English language Tweets in Figure 2. It structurally contains 6 fields as described as follows:

1. target: The polarity of the tweet (Positive=0, Bully/Aggressive=1 and Neutral=0).
2. ids: The identification number of each tweet.
3. date: The date of the tweet.
4. flag: The query.
5. user: The particular user that tweeted.
6. text: The text contained in the tweet.

3.3 Model architecture and settings

The CNN component for image feature analysis is depicted in Figure 3. CNN allows in depth analysis of images. It consists of three layers of convolutions with respectively 32, 64, and 128 filters and allows a rich representation of features.

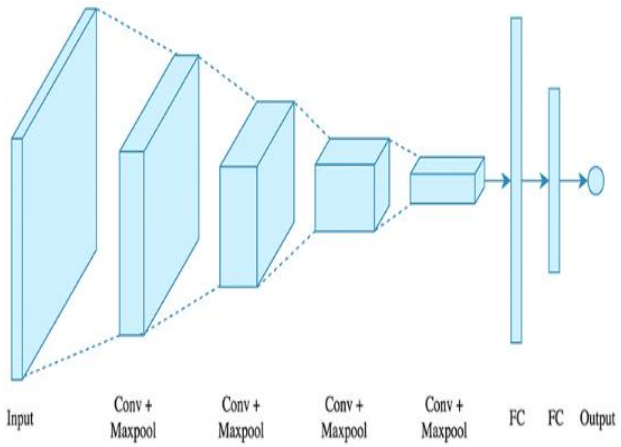

Figure 3. CNN component architecture

The LSTM architecture, like the recurrent neural network (RNN), consists of a series of repeated modules. However, the LSTM has a more intricate structure. The LSTM possesses the capacity to determine the optimal moment for replacing, updating, or discarding the information stored in individual neurons inside the cell state. The LSTM possesses the ability to effectively store and maintain information for long durations. The cell state in the LSTM experiences negligible alterations as it passes between cells via a sequence of simple linear operations. The LSTM's ability to alter the information stored in the cell state is regulated by a gate mechanism. The LSTM gate consists of three components: an input gate, a forget gate, and an output gate [34, 35].

All major open-source machine learning frameworks offer efficient, production-ready implementations of a number of RNN and LSTM network architectures. The LSTM network is a type of an RNN. This unit has the unique ability to maintain a hidden state, allowing the network to capture sequential dependencies by remembering previous inputs while processing [36]. Table 1 shows the experimental parameter for the model.

Table 1. Experimental parameters

Hyper Parameters	CNN	LSTM-RNN
Number of layers	Convolutional 32,64,128	32,64,128
Types of layers	Fully connected Pooling,1,2,3	Pooling1,2,3
Activation function on Hidden	ReLU	tanh
Activation functionon output	Softmax	Softmax
Regularization	L2	L2
Epocs	100	100
Optimizer	Adam	Adam

This work hybridized Convolutional Neural Network with Long Short-Term Memory (LSTM-RNN) for a better accuracy and optimal performance. The CNN enables images/videos to be digitally processed thereby giving the system to mine textual information from both videos and images, which may then be passed to the LSTM-RNN network, which is mainly built to handle textual tweets. The LSTM was added to the Recurrent Neural Network in order to solve the problem of texts with long time dependencies usually referred to as vanishing gradient. Figure 4 shows the architectural design of the Cyber-Aggression and Cyber-Bullying detection system developed in this project. Figure 5 shows the LSTM-RNN unit and Figure 6 shows the RNN unit. The Model architecture is made up of two main layers namely; The Output Layer also known as the user interface layer and the Hidden Layer that consists of several other sub-components. These layers with their embedded components are fully explained as follows: Data collection strategies are created to retrieve chronological tweets from Twitter endpoints. This research work makes use of Twitter Standard Search which is a platform for public information streaming. Application Programming Interface, a user interface that allows you to retrieve information in chronological sequence from twitter public domain. This API enables the system to download recent tweets that serve as input to the analyzer in real time. Image Dataset The image dataset is a collection of millions of well labeled and trained image data that serves as resources for the CNN component of the system. The main source of image data used in this project work is Google's Open Images, which is an image data million images that have been annotated with image-level labels and object bounding boxes. The reason for using this dataset was because, as the time of this write up, it is one of the largest image datasets freely available for download from Google website with no barriers, approximately 80% of the data set was used to train while the remaining 20% was for testing purposes. It is important to state that video inputs are chunked into image-like frames and treated as images too. Text Dataset Like the Image Dataset, the text dataset is a collection of tweets from twitter domain that have been trained and well labeled. Used for this project is sentiment140 dataset containing about 1.6 million well annotated tweets extracted from twitter website using twitter API. Input Identifier The

input identifier component is one that gets information from twitter Application Programming Interface (twitter(X) API) as they are being downloaded from twitter domain in real time. The main purpose of this component is that it pre-processes the input and sends the preprocessed data to the appropriate component. It has capacity to separate image/video from text, thereby sending image/video to the Convolutional Neural Network, Long Short-Term Memory and Recurrent Neural Network for further processing. In this research work, the introduction of Convolutional Neural Network is mainly to generate the textual representation of any tweeted image/video and passes counterpart for proper analysis and classification. Considering the CNN component in the architecture in Figure 3 from left to right, the input is an actual image that is scanned for features, and the light rectangle is the filter that passes over it. Stacked on top of one another in a stack is the activation map one for each used filter, which condensed via down sampling. The larger rectangle represents a patch that will be down sampled. By passing the filters over the first down sampled stack, a new collection of activation maps is formed while the 2nd down sampling condenses the 2nd group of activation maps and finally, a completely linked layer that assigns a label to each node's output. CNN applies a filter to the connection by the addition of two new kinds of layers namely; pooling and convolutional layers. The convolutional layer is the first layer in the architecture, which at first, breaks down an input image into a series of 3*3-pixel tiles square that overlap each of this will be run on a neural network with the weights left intact, resulting in the collection of tiles being transformed into an array. The output values will then be collected and placed in an array that numerically represents the content of each section content in the image, with color, width, and height channels on the axes thus leaving every tile with a 3*3*3 a three-dimensional representation. A fourth dimension for time will be added in case of videos. LSTM-RNN Regular Recurrent Neural Network hidden layers are unable to successfully store information about very long sentences, hence the need to build hidden layers with gate-operated memory unit capable of retaining the encoding done in the state for a long time. This type of RNN is called Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) capable of solving the problem of lack of long-range dependencies in the system. In this research work, LSTM-RNN is purposely built to enable easy mining of opinion from textual information derived from real-time tweets with considerations for texts with both short- and long-term dependencies usually referred to as vanishing gradient problem. The network is specially designed to accommodate phrases and sentences coming from real-time tweet, get them encoded and produces an output based on its last time step, which on purpose is then passed to a softmax activation function. The softmax function, also known as softargmax or normalized exponential function, is an activation function, which is a generalization of the logistic function to multiple dimensions. It is used as the last activation function of the LSTM-RNN to normalize the output of the network to a probabilistic distribution over predicted output class. The matcher is an output generating component by taking as an input the result of the network last activation function. It uses the result to match with the ones stored in the dataset. When a match is found, then the matcher fetches and displays the corresponding label of the softmax value. For example, consider the following dataset structure:

1. Sentence: “I will show you the stuff I am made of”
| Label: “Bully”
2. Sentence: “I will show you the place”
| Label: “Neutral”
3. Sentence: “I will show but not now!”
| Label: “Aggressive”
4. Sentence: “I will not show you because we are not mates”
| Label:” Aggressive”

Considering the above, let say an input tweet reads: “I won’t because we are not mates”, then the softmax will like produce a value close to 4 and matcher will in this case will be sentence 4 and displays the Label (i.e., Aggressive).

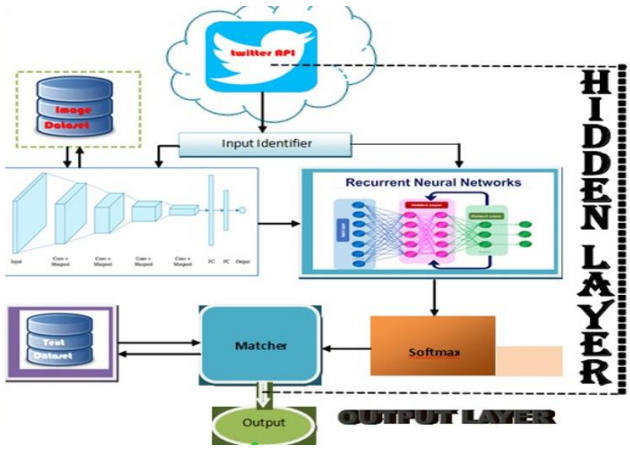

Figure 4. Model architecture

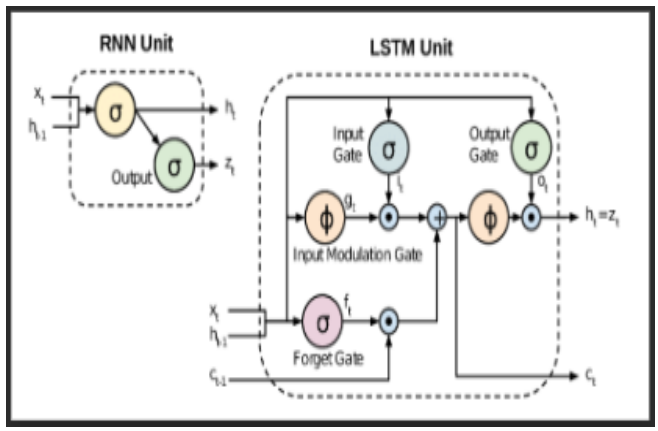

Figure 5. LSTM and RNN unit

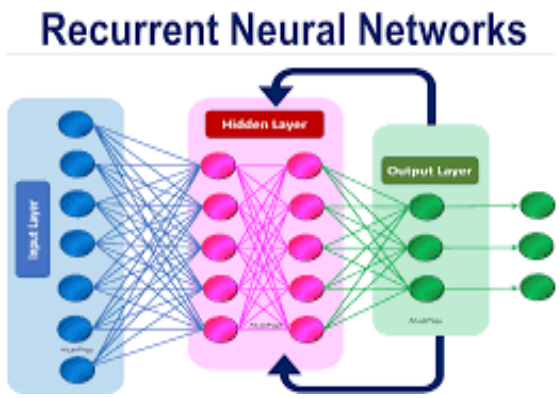

Figure 6. RNN unit

It is worthy of note that generalization can be done and the overall output can be represented in graphical form according

to specification. System Implementation The system designed in this paper was implemented as a stand-alone application that runs on desktop computer systems. The application has different functional modules and user interfaces. Development Tool The system is implemented using python programming language. Python coding was done in python Integrated Development Environment called PyCharm. Tweepy, a Python library for accessing the Twitter API was also used. The Twitter Application Programming Interface (API) enables our python code to access and use twitter data in real-time. Hypertext Mark-Up Language (HTML) and Cascading Style Sheet (CSS) were used for the interface design while JavaScript was used to pass data to/from Python and HTML elements through the use of electron.

Electron is a python plugin that enables python code to run or execute on a standard web browser. Interface Design The graphical user interfaces (GUIs) were carefully designed to be simple and user friendly. The sentiment analyzer for detecting cyberbullying and aggression developed in this work has five main pages namely; User Login page, New User Registration page, Application Welcome page, Search page, Output Display Page. The user login page provides some levels of security to the developed belief rule base expert system platform. The user registration page that enables new users to submit their information such as full name, username and password, etc. to the system's database. The information, especially the username and password, are used by the users during login. The system checks whether the supplied username and password match the one registered for the underlined user. If the login details match the one in the database, then the user is logged in successfully and taken to the Keyword Search Page, otherwise the user is denied access into the system.

4. RESULTS AND DISCUSSION

For the purpose of performance measurement, already pre-processed dataset from sentiment140 was used to train classification models such as Logistic Regression, Naive Bayes, K-Nearest Neighbors, Random Forest, Stochastic Gradient Descent (SGD) and Support Vector Machines in addition to the DNN (CNN and RNN) developed in this work. It was so apparent that Deep Neural Network (DNN) outperformed the other classification algorithms as presented in Table 2.

Table 2. Input samples

S/N	MODEL	ACCURACY	F-MEASURE
1	Logistic Regression	0.671	0.637
2	Stochastic Gradient Descent	0.655	0.601
3	Bernoulli naive Bayes	0.640	0.605
4	Random Forest	0.558	0.512
5	K- Nearest Neighbour	0.543	0.476
6	Linear Support Vector machine	0.669	0.648
7	Recurrent Neural Network	0.951	0.910
8	Convolutional Neural Network	0.911	0.890

The model built in this research work is not represented in the above table because RNN- CNN cannot be tested, for now, as a single library/plugin in python. The Figure 6 shows performance measurement graphical representation. A sample data containing 700 tweets were used, out of which 550 were used for training while the remaining 150 were used for testing as presented succinctly in Table 3.

The pictorial representation of all the tested algorithms mentioned above with respect to their performance in terms of accuracy is presented in Figure 7. From the evaluation chart, it is obvious that RNN in this context outperform other models in terms accuracy, which is one important criterion to look for when carrying out sentiment analysis of this type. It is also worth knowing that, in this research, both RNN and CNN are combined for maximum performance that cannot be competed with by any singular machine learning model. Contextually, CNN outperforms other model when it comes image/video analysis while RNN is the best in terms text mining, which was why these two deep learning models were combined in this research work for maximum performance and accuracy.

The proposed system shows a comparable improvement over the baseline methods as show in Table 4 and compared favourably with other techniques in terms of accuracy.

Moreover, the ethical implications of automated cyberbullying detection were considered, acknowledging the delicate balance between flagging harmful content and preserving user privacy and freedom of speech. The model's design required careful consideration to respect users' digital rights while maintaining its commitment to creating safer online environments. This dual commitment is reflected in the model's methodology, emphasizing user protection while striving for comprehensive detection and minimal false positives. Comparatively, the study's findings align with the current trajectory in cyberbullying research and technological advancements in artificial intelligence. They underscore the potential deep learning holds in transforming safety measures on social networking platforms. However, they also bring to light certain limitations that future studies will need to address.

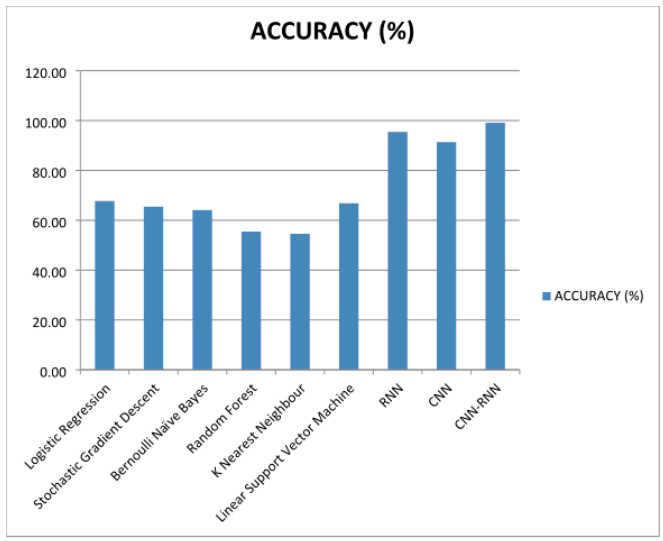

Figure 7. Performance evaluation chart

Table 3. Tested sample results

Model	STT	BOW	CDBA	ND	Accuracy %
Logistic Regression	3500	1100	735	365	66.82
Stochastic Gradient Descent	3500	1100	720	380	65.45
Bernoulli Naive Bayes	3500	1100	705	395	64.09
Random Forest	3500	1100	610	490	55.45
K Nearest Neighbour	3500	1100	600	500	54.55
Linear Support Vector Machine	3500	1100	745	355	67.73
Recurrent Neural Network	3500	1100	1050	50	95.45
Convolutional Neural Network	3500	1100	1005	95	91.36

STT-Sample tweet tested, **BOW**-Bag of words, **CDBA**-Correctly detected bullying/aggression, **ND**-Not detected

Table 4. Baseline comparison

Author	Scope	Methods	F1
Sunagar and Kanavalli [37]	Text classification	RCNN-LSTM	0.68
Sadiq et al. [38]	Cyber-Troll	TF-IDF	0.92
Model	Cyberbully	RNN-LSTM	0.91

5. CONCLUSIONS

Sentiment analysis, also known as opinion mining can be defined as the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. It is often driven by an algorithm, scoring the words used along with voice inflections that can indicate a person's underlying feelings about the topic of a discussion. Analyzing sentiments creates rooms for a more objective interpretation of factors that are otherwise difficult to compute or typically measured subjectively. This research work describes the design, implementation and applications of a Convolutional Neural Network-Recurrent Neural Network Based Cyber Bullying/Aggression detection system using Sentiment Analysis leveraging on real-Time Twitter Dataset. Combining sophisticated deep learning algorithm such as CNN and RNN provides a very robust and simplistic approach for detecting Cyber-bullying and aggression in real-time tweets using twitter Application Programming Interface thus enabling, for instance, parents to control the effects of Cyber-bullying or aggression on their kids, just in case. The increase number of children or kids being bullied online and its effects on these children cannot be overemphasized and therefore, sequel to that, there is need for a robust Cyber-bullying detection system such as the one developed in this project work. Practically speaking, the accuracy of this system is astonishingly tacit and unmatched. Therefore, this system is highly recommended for use, especially by parents and teachers, to check the level of how much a kid has been bullied, thus that can be used to destigmatize such a child from that kind of experience. The system also has the capacity to cater for identity grabbing, which can also be used to put in check the bullies that will eventually reduce the number of their bullied counterparts. Identity grabbing is the ability of the analyzer to be able to identify and display the sender or originator of a particular tweet. Future work may involve other dataset inclusion such as Instagram, WhatsApp and other social media in addition to the twitter dataset used in this project work. Thus, that would make the system more robust in power and accuracy since the power of any data mining algorithm depends solely on data. Combining DNN with other technique could improve performance prediction of models.

The system in practice will interface with Twitter API and it will only be accessible to only register user.

ACKNOWLEDGMENT

We would like to thank the anonymous reviewers for detailed comments and useful feedback. The authors declare no competing interests.

REFERENCES

[1] Despoina, C., Ilias, L., Jeremy, B., Emiliano, D.C., Gianluca, S., Anthena, V., Nicolas, K. (2019). Detecting Cyber-bullying and cyberaggression in social media. ACM. Transactions on the Web, 13(3). https://doi.org/10.1145/3343484

[2] Singh, S.P., Bhakar, S. (2019). Real time cyberbullying detection. International Journal of Engineering and Advanced Technology (IJEAT), 9(2): 5197-5201. https://www.ijeat.org/wp-content/uploads/papers/v9i2/B4253129219.pdf.

[3] Kumar, A., Sachdeva, N. (2021). Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network. Multimedia Systems, 28: 2043-2052. https://doi.org/10.1007/s00530-020-00747-5

[4] Teng, T.H., Vraathan, K.D. (2023). Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches. IEEE Access, 11: 55533-55560. https://doi.org/10.1109/ACCESS.2023.3275130

[5] Dispartilaw. https://www.dispartilaw.com/common-sites-for-social-media-cyberbullying/.

[6] Pandey A., Tiwari, N.K. (2022). Cyber crime and their detection with machine learning: Comprehensive study of phishing and cyberbullying. National Journal of Cyber Security Law, 5(1). https://lawjournals.celnet.in/index.php/njcsl/article/view/1113.

[7] Chen, W., Lin, F, Zhang, X., Li, G., Liu, B. (2022). Jointly learning sentimental clues and context incongruity for sarcasm detection. IEEE Access, 10: 48292-48300. https://doi.org/10.1109/ACCESS.2022.3169864

[8] Sherly, T.T., Jeetha, B.R. (2021). Sentiment analysis and deep learning based cyber bullying detection in twitter dataset. International Journal of Recent Technology and Engineering, 10(4): 15-25. https://doi.org/10.35940/ijrte.d6511.1110421

[9] Patchin, J.W., Hinduja, S. (2006). Bullies move beyond the schoolyard: A preliminary look at cyberbullying. Youth Violence and Juvenile Justice, 4(2): 148-169. https://doi.org/10.1177/1541204006286288

[10] Gohal, G., Alqassim, A., Eltyeb, E., Rayyani, A., Hakami, B., Faqih, A., Hakami, A., Mahfouz, M. (2023). Prevalence and related risks of cyberbullying and its effects on adolescent. BMC Psychiatry, 23(1): 39. https://doi.org/10.1186/s12888-023-04542-0

[11] Ramiandrisoa, F., Mothe, J. (2024). Aggression identification in social media: A transfer learning based approach. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, Marseille, France, pp. 26-31. https://aclanthology.org/2020.trac-1.5.pdf.

[12] Cohen-Almagor, R. (2018). Taking North American white supremacist groups seriously: The scope and the challenge of hate speech on the internet. International Journal of Crime, Justice and Social Democracy, 7(2): 38-57. https://doi.org/10.5204/ijcjsd.v7i2.517

[13] Cohen-Almagor, R. (2022). Bullying, cyberbullying and hate speech. International Journal of Technoethics, 13(1). https://doi.org/10.2139/ssrn.4033031

[14] Broadband Search (2024): All the latest cyberbullying ststistics for 2024. https://www.broadbandsearch.net/blog/cyber-bullying-statistics.

[15] Garg, V.K. (2018). Deep learning as a frontier of machine learning: A review. International Journal of Computer Applications, 182(1): 22-30. https://doi.org/10.5120/ijca2018917433

[16] Ali, H., Sana, M., Ahmad, K., Shahaboddin, S. (2018). Machine learning-based sentiment analysis for twitter accounts. Mathematical and Computational Applications, 23(1): 11. https://doi.org/10.3390/mca23010011

[17] Thirupathi, R.K., Sai, B.A., Chaitanya, V.P. (2017). Implementation of sentiment analysis on twitter data. International Journal of Pure and Applied Mathematics, (74). http://www.ijpam.eu.

[18] Chatzakou, D., Leontiadis, I., Blackburn, J., De Cristofaro, E., Stringhini, G., Vakali, A., Kourtellis, N. (2019). Detecting cyberbullying and cyberaggression in social media. ACM Transactions on the Web, 13(3): 1-51. https://doi.org/10.1145/3343484

[19] Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., Vakali, A. (2017). Mean birds: Detecting aggression and bullying on Twitter. https://doi.org/10.48550/arXiv.1702.06877

[20] Mangaonkar, A. (2017). Collaborative Detection of Cyber-bullying Behavior in Twitter Data (Master's Thesis, Purdue University, Indianapolis, Indiana). https://core.ac.uk/download/pdf/84831838.pdf.

[21] Sharma, K., Shailendra, Kshitiz, K. (2018). NLP and machine learning techniques for detecting insulting comments on social networking platforms. In 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE), Paris, France. https://doi.org/10.1109/ICACCE.2018.8441728

[22] Rosa H., Pereira, N., Ribeiro, R., Ferreira, P.C., Carvalho, J.P., Oliveira, S., Coheur, L., Paulino, P., Veiga Simão, A.M., Trancoso, I. (2018) Automatic cyberbullying detection: A systematic review. Computers in Human Behavior, 93: 333-345. https://doi.org/10.1016/j.chb.2018.12.021

[23] Fadelli, I. (2019). A deep learning-based method to detect Cyber-bullying on twitter. https://techxplore.com/news/2019-01-deep-learning-based-method-cyberbullying-twitter.html#google_vignette.

[24] Choi, R.Y., Coyner, A.S., Kalpathy-Cramer, J., Chiang, M.F., Campbell, J. (2020). Introduction to machine learning, neural networks, and deep learning. Translational Vision Science & Technology, 9(2): 14. https://doi.org/10.1167/tvst.9.2.14

[25] Arathi, H., Gabriel, S., Karina, T., Victor, M., Hector, P. Jesus, O. (2021). Social sentiment sensor in twitter for predicting cyber-attacks using ℓ1 regularization. Sensors, 18(5): 1380. https://doi.org/10.3390%2Fs18051380

[26] Sentamilselvan, K., Aneri, D., Athithiya, A.C., Kani Kumar, P. (2020). Twitter sentiment analysis using machine learning techniques. International Journal of Engineering and Advanced Technology (IJEAT), 9(3): 279-289. https://doi.org/10.1007/978-3-319-17996-4_25

[27] Chong, S., Dubey, G., Rana, A. (2017). Product opinion mining using sentiment analysis on smartphone reviews. In 2017 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, pp. 377-383. https://doi.org/10.1109/ICRITO.2017.8342455

[28] Muhammed, M., Khan, M., Bashier, E. (2016). Machine learning algorithm and application. Journal of Machine Learning: Algorithm and Application. https://doi.org/10.1201/9781315371658

[29] Khan, U., Khan, S., Rizwan, A., Atteia, G., Jamjoom, M.M., Samee, N.A. (2022). Aggression detection in social media from textual data using deep learning models. Applied Sciences, 12(10): 5083. https://doi.org/10.3390/ app12105083

[30] Kazbekova, G., Baimukhambetova, G., Bolatkyzy, D., Suleimenova, Z., Akhmetova, Z., Bekbosynov, T., Galdybekova, B., Kozhakhmetov, K. (2023). Offensive language detection on online social networks using hybrid deep learning architecture. International Journal of Advanced Computer Science and Applications, 14(11). https://doi.org/10.14569/IJACSA.2023.0141180

[31] Nouri, H., Karim, S., Habbat, N. (2023). Enhancing Arabic sentiment analysis in e-commerce reviews on social media through a stacked ensemble deep learning approach. Mathematical Modelling of Engineering Problems, 10(3): 790-798. https://doi.org/10.18280/mmep.100308

[32] Alkasassbeh, M., Almomani, A., Aldweesh, A., Al-Qerem, A., Alauthman, M., Nahar, K., Mago, B. (2024). Cyberbullying detection using deep learning: A comparative study. In 2024 2nd International Conference on Cyber Resilience (ICCR), Dubai, United Arab Emirates. https://doi.org/10.1109/ICCR61006.2024.10533166

[33] Mandhasiya, G.G., Murfi, H., Bustamam, A. (2024). The hybrid of BERT and deep learning models for Indonesian sentiment analysis. Indonesian Journal of Electrical Engineering and Computer Science, 33(1): 591-602. http://doi.org/10.11591/ijeecs.v33.i1.pp591-602

[34] Li, Y., Li, Y., Wang, J., R. S.S. (2020). Sentiment analysis for e-commerce product reviews in Chinese based on sentiment lexicon and deep learning. IEEE Access, 8: 23522-23530. https://doi.org/10.1109/ACCESS.2020.2969854

[35] Anki, P., Bustamam, A., Al-Ash, H.S., Sarwinda, D. (2020). High accuracy conversational ai chatbot using deep recurrent neural networks based on bilstm model. In 2020 3rd International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, pp. 382-387. https://doi.org/10.1109/ICOIACT50329.2020.9332074

[36] Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory

(LSTM) network. Physica D: Nonlinear Phenomena, 404: 132306. https://doi.org/10.1016/j.physd.2019.132306
[37] Sunagar, P., Kanavalli, A. (2022). A Hybrid RNN based deep learning approach for text classification. International Journal of Advanced Computer Science and Applications, 13(6). https://doi.org/10.14569/ijacsa.2022.0130636
[38] Sadiq, S., Mehmood, A., Ullah, S., Ahmad, M., Choi, G.S., On, B.W. (2021). Aggression detection through deep neural model on Twitter. Future Generation Computer Systems, 114: 120-129. https://doi.org/10.1016/j.future.2020.07.050