



IJIRCCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 5, May 2023

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.379



9940 572 462



6381 907 438



ijircce@gmail.com



www.ijircce.com

Preventing Cyberbullying in Social Networking Sites Using Deep Learning

R.DineshKumar^[1], R.Vasim Akram^[2], S.Vignesh^[3], Mrs.J.Mercy Grace,M.E^[4]

U.G Students, Department of Computer Science and Engineering, Trichy Engineering College, Konalai, Trichy, Tamilnadu, India^{1,2,3}

Assistant Professor, Department of Computer Science and Engineering, Trichy Engineering College, Konalai, Trichy, Tamilnadu, India⁴

ABSTRACT: Cyberbullying is bullying that takes place over digital devices like cell phones, computers, and tablets. Cyberbullying can occur through SMS, Text, and apps, or online in social media, forums, or gaming where people can view, participate in, or share content. Cyberbullying includes sending, posting, or sharing negative, harmful, false, or mean content about someone else. It can include sharing personal or private information about someone else causing embarrassment or humiliation. To avoid or detecting cyberbullying attacks, many existing approaches in the literature incorporate Machine Learning and Natural Language Processing text classification models without considering the sentence semantics. The main goal of this project is to overcome that issue. This project proposed a model LSTM - CNN architecture for detecting cyberbullying attacks and it used word2vec to train the custom of word embeddings. This model is used to classify tweets or comments as bullying or non-bullying based on the toxicity score. LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. A convolutional neural network (CNN) is a type of artificial neural network and it has a convolutional layer to extract information by a larger piece of text and by using this model LSTM- CNN achieve a higher accuracy in analysis, classification and detecting the cyberbullying attacks on posts and comments.

KEYWORDS: Cyberbullying Detection, Social Networking Sites, Deep learning algorithm-LSTM, BiLSTM, CNN.

PROBLEM STATEMENT

Social media networks such as Facebook, Twitter, Flickr, and Instagram have become the preferred online platforms for interaction and socialization among people of all ages. Cyber-bullying events have been increasing mostly among young people spending most of their time navigating between different social media platforms. Particularly, social media networks such as Twitter and Facebook are prone to CB because of their popularity and the anonymity that the Internet provides to abusers. In India, for example, 14 percent of all harassment occurs on Facebook and Twitter, with 37 percent of these incidents involving youngsters. Moreover, cyberbullying might lead to serious mental issues and adverse mental health effects. Most suicides are due to the anxiety, depression, stress, and social and emotional difficulties from cyberbullying events. This motivates the need for an approach to identify cyberbullying in social media messages (e.g., posts, tweets, and comments).

OBJECTIVE

This project aims to automatically detect cyberbullying from tweets by using deep learning approaches. The aim of this project is to identify the maximum number of cyberbullying related tweets from Twitter as soon as it is posted by users. The objective of our solution is to identify the bullies from raw Twitter data based on the context as well as the contents in which the tweets exist.

I. INTRODUCTION

1.1 Project description

Cyberbullying is bullying that takes place over digital devices like cell phones, computers, and tablets. Cyberbullying can occur through SMS, Text, and apps, or online in social media, forums, or gaming where people can view, participate in, or share content. Cyberbullying includes sending, posting, or sharing negative, harmful, false, or

mean content about someone else. It can include sharing personal or private information about someone else causing embarrassment or humiliation. Some cyberbullying crosses the line into unlawful or criminal behaviour.

1.2 Different Kinds of Cyberbullying:

There are many ways that someone can fall victim to or experience cyberbullying when using technology and the internet. Some common methods of cyberbullying are:

Harassment – When someone is being harassed online, they are being subjected to a string of abusive messages or efforts to contact them by one person or a group of people. People can be harassed through social media as well as through their mobile phone (texting and calling) and email. Most of the contact the victim will receive will be of a malicious or threatening nature.

Doxing – Doxing is when an individual or group of people distribute another person's personal information such as their home address, cell phone number or place of work onto social media or public forums without that person's permission to do so. Doxing can cause the victim to feel extremely anxious and it can affect their mental health.

Cyberstalking – Similar to harassment, cyberstalking involves the perpetrator making persistent efforts to gain contact with the victim, however this differs from harassment – more commonly than not, people will cyberstalk another person due to deep feelings towards that person, whether they are positive or negative. Someone who is cyberstalking is more likely to escalate their stalking into the offline world.

Revenge porn – Revenge porn, is when sexually explicit or compromising images of a person have been distributed onto social media or shared on revenge porn specific websites without their permission to do so. Normally, images of this nature are posted by an ex-partner, who does it with the purpose of causing humiliation and damage to their reputation.

Swatting – Swatting is when someone calls emergency responders with claims of dangerous events taking place at an address. People swat others with the intention of causing panic and fear when armed response units arrive at their home or place of work. Swatting is more prevalent within the online gaming community.

Corporate attacks – In the corporate world, attacks can be used to send masses of information to a website in order to take the website down and make it non-functional. Corporate attacks can affect public confidence, damaging businesses reputations and in some instances, force them to collapse.

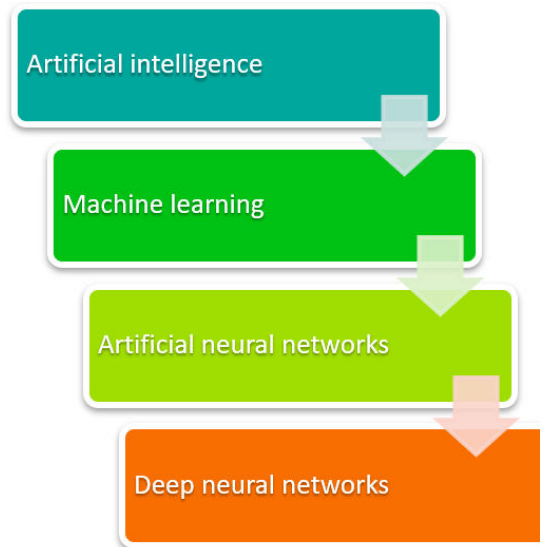
Account hacking – Cyberbullies can hack into a victim's social media accounts and post abusive or damaging messages. This can be particularly damaging for brands and public figures.

False profiles – Fake social media accounts can be setup with the intention of damaging a person or brand's reputation. This can easily be done by obtaining publicly available images of the victim and making the account appear as authentic as possible.

Slut shaming – Slut shaming is when someone is called out and labelled as a "slut" for something that they have done previously or even just how they dress. This kind of cyberbullying often occurs when someone has been sexting another person and their images or conversations become public. It is seen more commonly within young people and teenagers but anyone can fall victim to being slut shamed.

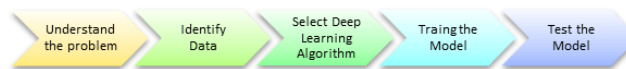
1.3 Deep Learning:

Deep Learning is a field which comes under Machine Learning and is related to the use of algorithms in artificial neural networks. It is majorly used to create a predictive model to solve the problems with just a few lines of coding. A Deep Learning system is an extensive neural network which is inspired by the function and structure of the brain. Deep learning models can be supervised, semi-supervised, or unsupervised (or a combination of any or all of the three). They're advanced machine learning algorithms used by tech giants, like Google, Microsoft, and Amazon to run entire systems and power things, like self-driving cars and smart assistants. First, machine learning had to get developed. ML is a framework to automate (through algorithms) statistical models, like a linear regression model, to get better at making predictions. A model is a single model that makes predictions about something. Those predictions are made with some accuracy. A model that learns—machine learning—takes all its bad predictions and tweaks the weights inside the model to create a model that makes fewer mistakes.



1.4 Deep Learning Process:

A deep neural network provides state-of-the-art accuracy in many tasks, from object detection to speech recognition. They can learn automatically, without predefined knowledge explicitly coded by the programmers.



II. LITERATURE SURVEY

1. M. Di Capua, et al. [1] raises the unsupervised how to develop an online bullying model based on a combination of features, based on traditional textual elements and other "social features". Features were divided into 4 categories as Syntactic features, Semantic features, Sentiment features, and Community features. The author has used the Growing Hierarchical Self Map editing network (GHSOM), with 50 x grid 50 neurons and 20 elements as the insertion layer. M. Di Capua, and others used an integration algorithm k-means to separate the input database and GHSOM in the Formspring database. The effects of this hybrid the unsupervised way surpasses the previous one results. The author then checked the youtube database at 3 p.m. Different Machine Learning Models: Naive Bayes Classifier, Decision Tree Classifier (C4.5), and Support Vector Machine (SVM) with Linear Kernel. It was saw that the combined effects of hate speech turned around so that we have lower accuracy in the youtube database compared to FormSpring tests, as in the text analysis and syntactical features work differently in on both sides. When this hybrid method is used in Twitter Database, led to weak memory and F1 Score. The model proposed by the authors can also be improved used in building constructive mitigation applications cyberbullying problems.

2. J. Yadav, et al. [2] suggests a new approach to this the discovery of internet cyberbullying on social media in using a BERT model with a single line neural and was tested on the Formspring forum and Wikipedia Database. The proposed model provided performance 98% accuracy of spring Form databases and 96% accuracy in a relatively comprehensive Wikipedia database previously used models. The proposed model provided better Wikipedia database results due to its size g without the need for excessive sampling while I The spring data form requires multiple samples.

3. R. R. Dalvi, et al. [3] suggests a way to do this detect and prevent online exploitation on Twitter using Classified supervised machine learning algorithms. In this study, the live Twitter API is used for compilation tweets and data sets. The proposed model tests both Support Vector Machine and Naive Bayes on the data sets are collected. To remove a feature, use it TFIDF vectorizer. The results show that it is accurate of an online bullying model based on Vector Support The machine is about 71.25% better than Naive Bayes was almost 52.75%.

4. Trana R.E., et al. [4] The goal was to design a machine learning model to minimize special events including text extracted from image memes. Author include a site that contains approximately 19,000 text views have been published on YouTube. This study discusses the operation of three learning machines equipment, Uninformed Bayes, Support Vector The machine, as well as the convolutional neural network used in on the YouTube website, and compare the results with existing details of the Form. They do not write continuously investigate cyber bullying algorithms sub-sections within the YouTube website. Naive Bayes beat SVM and CNN in the next four categories: race, nationality, politics, and general. SVM passed well with the inexperienced Naïve Bayes again CNN is in the same gender group, with all three algorithms show equal performance with the middle body group accuracy. The results of this study provided inaccurate data used to distinguish between incidents of abuse and non-violence. Future work can focus on the construction of a two-part separation system used to test text taken from photos to see that the YouTube website provides a better context for aggressionrelated collections.

5. G. A. León-Paredes et al. [5] they described I the development of an online bullying detection model is used Native Language Processing (NLP) and Mechanics Reading (ML). Spanish Cyberbullying Prevention The system (SPC) was developed by installing the machine learning strategies Naïve Bayes, Support Vector Machine, and Logistic Regression. Database used this study was posted on Twitter. Plurality 93% accuracy was achieved with the help of third parties techniques used. Charges of online exploitation were found with the help of this system presented the accuracy of 80% to 91% on average. Stemming and lemmatization techniques in NLP can be used continuously increasing system accuracy. Such a model can and used for adoption in English and local languages if possible.

6. P. K. Roy, et al. [6] details about creating a request for hate speech on Twitter via the help of a deep neural convolutional network. Machine learning algorithms such as Logistic Regression (LR), Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting (GB), and K-nearby Neighbors (KNN) have it used to identify tweets related to hate speech in Twitter and features have been removed using tf-idf process. The leading ML model was SVM but it managed predict 53% hate speech tweets on a 3: 1 database to be tested train. The reason behind the low forecast scale was unequal data. The model is based on the prediction of hate speech tweets. Advanced learning based methods on the Convolutional Neural Network (CNN), Long-Term Memory (LSTM), and its Content LSTM combinations (CLSTM) have the same effects as separate distributed database. 10 times cross confirmation was used in conjunction with the proposed DCNN model once you got a very good rate of memory. It was 0.88 hate speech and 0.99 non-hate speech.

III. METHODOLOGY

This project will be built with Python and web technology. Within that, we will first search for and download the dataset needed to train the model. After downloading, we will pre-process the data before trying to transfer it to Tf-Idf and Word2vec . The dataset is then trained and the model is generated separately using the LSTM(Long Short Term Memory), BLSTM (Bidirectional Long Short Term Memory), and CNN algorithms. Then, using the STREAMLIT framework, we will create a web-based application. We will retrieve realtime tweets from Twitter and then apply the generated model to these retrieved tweets to determine whether the te cyberbullying or not. We use Python as the backend, Streamlit Cloud as the database, and HTML5, CSS3, and streamlit as the frontend.

IV. EXISTING SYSTEM

In this chapter existing machine learning classifiers utilized for tweet classification will be discussed. This chapter analysed five supervised machine learning algorithms: Support Vector Machines (SVM), Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), Gradient Boosting model (GBM), Logistic Regression (LR) and Voting Classifier (Logistic Regression C Stochastic Gradient Descent classifier).

Disadvantages:

- Difficult to track.
- Most of the cyberbullying cases go unreported.
- Low accuracy.
- Time consuming process.
- Response time is slow.
- Basic features and common classifier accuracy is low.
- Data are manually labelled using online services or custom applications.

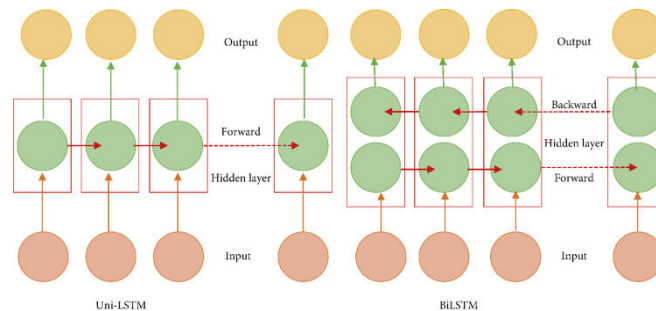
V. PROPOSED SYSTEM

In this paper, we design a model based on the bidirectional BiLSTM to detect cyberbullying in textual form.

5.1 BiLSTM:

Bidirectional LSTMs are an extension of LSTMs that can improve model performance on sequence classification problems. In problems where all time steps of the input sequence are available, Bidirectional LSTMs train two instead of one LSTMs on the input sequence. This can provide additional input context to the network and result in faster and even fuller learning on the problem. It involves duplicating the first periodic layer in the network so that there is now two layers' side-by-side, then providing the input sequence as-is as input to the first layer and providing a reversed copy of the input sequence to the second layer.

The use of sequence bi-directionally was initially justified in the domain of speech recognition because there is evidence that the input context of the whole utterance is used to interpret what is being said rather than a simple interpretation. The Use of Bidirectional LSTM may not make sense for all prediction problems but can offer benefits in terms of better results to those domains where it is appropriate.



Bidirectional LSTM (BiLSTM) is a recurrent neural network used primarily on natural language processing. Unlike standard LSTM, the input flows in both directions, and it's capable of utilizing information from both sides. It's also a powerful tool for modelling the sequential dependencies between words and phrases in both directions of the sequence. In summary, BiLSTM adds one more LSTM layer, which reverses the direction of information flow. Briefly, it means that the input sequence flows backward in the additional LSTM layer. Then we combine the outputs from both LSTM layers in several ways, such as average, sum, multiplication, or concatenation.

Advantages:

- It successfully classifies the tweets in various classes.
- Accuracy is high.
- This method detects the offensive post or messages.
- The "filtered content" is displayed at back to the page, in such a way preventing the display of explicit content.
- An automatically generate a report for each incident is also provided.

VI. IMPLEMENTATION

6.1 Modules:

Social Networking Web App
 Training Phase: Cyberbullying Tweet Classification
 Testing Phase: Prediction

6.2 Social Networking Web App

Build a social networking service is an online platform which people use to build social networks or social relationships with other people who share similar personal or career interests, activities, backgrounds or real-life connections. Social networking services vary in format and the number of features. The classification model has been exposed as a Streamlit APP which was consumed by a Web application built using Python's Streamlit framework. The main features include an Admin dashboard for visualization of cyberbullying activities, an option to search tweets, and automatic generation and emailing of reports of cyberbullying activity.



6.3 Cyberbullying Analysis APP:

In this module we developed the APP for cyberbullying analytics on chat or post user data. It focuses on keywords and analyzes chat or post according to a two-pole scale (positive and negative).

6.4 Cyberbullying Data Set Annotation:

We used cyberbullying data from Kaggle. The dataset in consisted of two labels, positive and negative, while was composed of three labels of positive, neutral, and negative. Furthermore, the dataset in was composed of five labels of positive, somewhat positive, neutral, somewhat negative, and negative.

6.5 Pre-Processing:

Datasets contain unnecessary data in raw form that can be unstructured or semi-structured. Such unnecessary data increases training time of the model and might degrades its performance. Pre-processing plays a vital role in improving the efficiency of DL models and saving computational resources. Text pre-processing boosts the prediction accuracy of the model. The preprocessing step is essential in cyberbullying detection. It consists of both cleaning of texts (e.g., removal of stop words and punctuation marks), as well as spam content removal.

6.6 Feature Extraction:

After the data pre-processing step, the next essential step is the choice of features on a refined dataset. Supervised deep learning classifiers require textual data in vector form to get trained on it. The textual features are converted into vector form using TF and TF-IDF techniques in this project. Features extraction techniques not only convert textual features into vector form but also helps to find significant features necessary to make predictions. For the most part all features do no contribute to the prediction of the target class. That is the reason feature extraction is the important part in the recognition of happy and unhappy related tweets. What actually Term Frequency(TF) means that, according to what often the term arises within the document? It's measured by TF. This will be achievable with the intention of a term would seem a lot further in lengthy documents than short documents because every document is variant in extent. Like the mode about standardization:

$$TF(t) = \frac{\text{No. of times term } t \text{ shows in a document}}{\text{Total no. of terms inside document}}$$

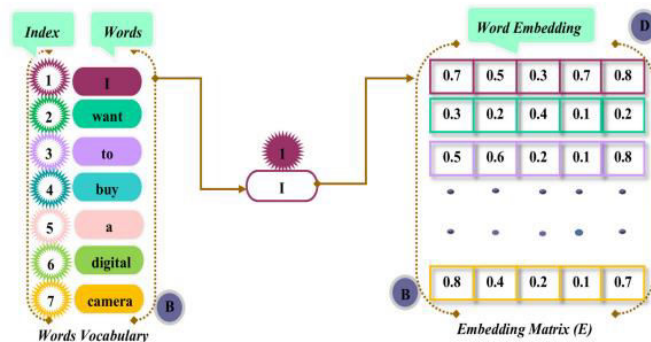
The term frequency be frequently divided with the document length (the total number of terms in the document). IDF: Inverse documents frequency proceeds to find how much a term is significant within the text. Every term is measured equally when TF is computed. Nevertheless, it is recognized that convinced terms, like "is", "of", and "that", can show much more times except contain small prominence. Therefore, frequent terms are needed to be weighed down as level up exceptional ones, through calculating following

$$IDF(t) = D \log(e) \frac{\text{Total No. of documents}}{\text{No. of documents through term } t \text{ in it}}$$

Term frequency (TF) is utilized regarding data recovery and shows how regularly an articulation (term, word) happens in a tweets.

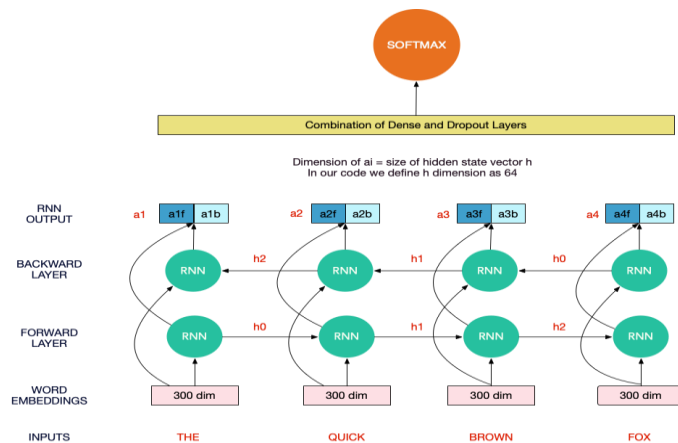
6.7 Word Embedding's:

The semantic meanings of words are provided by word embedding in this project, which is first used in the semantic word cloud generation to the best of our knowledge. We prepare the related text corpus and then train our word embedding by using the continuous bag-of-words (CBOW) model, which is implemented in the open-source toolkit word2vec. CBOW is orders of magnitude faster than the others for training datasets and yields significant gains for dependency parsing. After training, we extract the semantic meanings of all important key words from the word embedding. By using the pre-trained word embedding, each word corresponds a vector in the low dimensional space, typically 50-500.



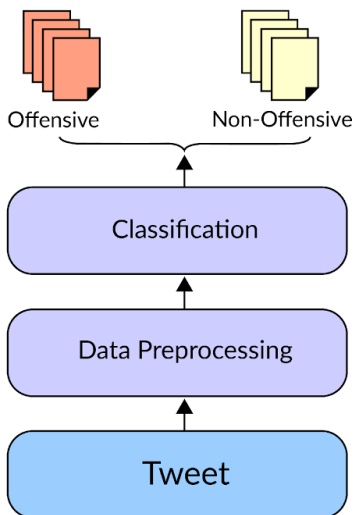
6.8 BiLSTM Classification:

For cyberbullying detection in online social media sites, four Deep Neural Network (DNN) based models (i.e., BLSTM, LSTM, CNN, and Attention-based BLSTM) were developed and the proposed model systematically detected cases of cyberbullying on numerous Social media platform. Bi-directional LSTM (BiLSTM): BiLSTM is a slightly advanced version of regular LSTM. The main difference between these two are that BiLSTM will have its input in two ways.

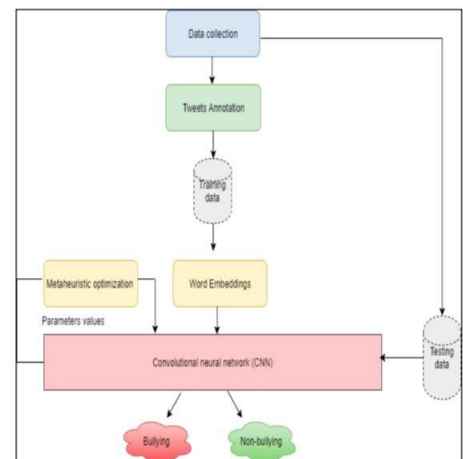


VII. SYSTEM DESIGN

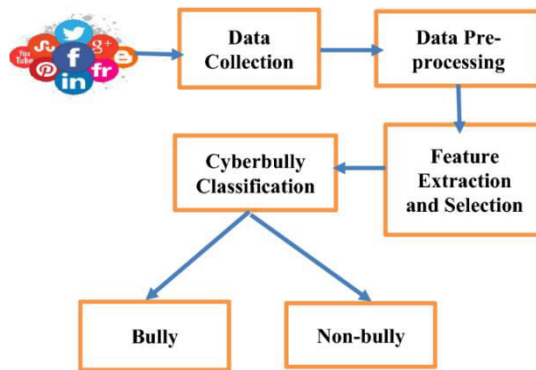
Data Flow Diagram-1:



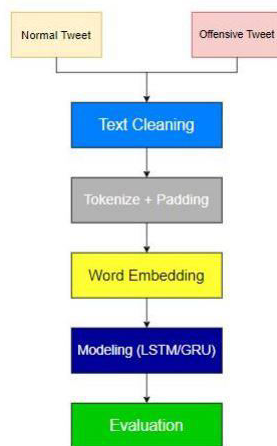
Data Flow Diagram-2:



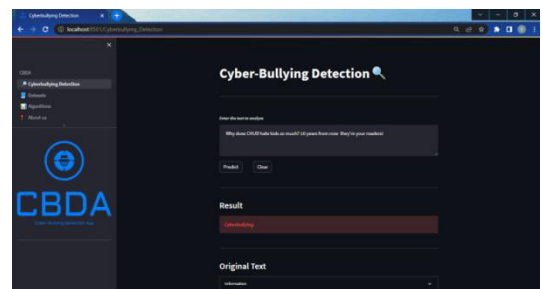
Data Flow Diagram-3:

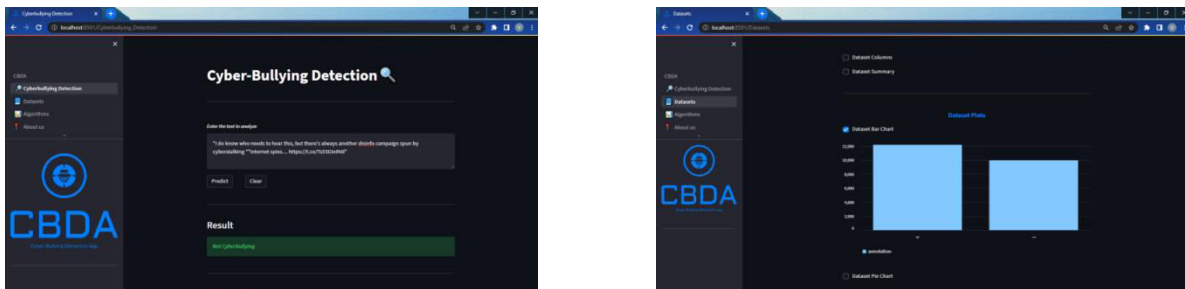


Model Layout



VIII. RESULTS





IX. CONCLUSION

Cyberbullying is the harassment that takes place in digital devices such as mobile phones, computers and tablets. The means used to harass victims are very diverse: text messages, applications, social media, forums or interactive games. One of the things that complicates these types of situations that occur through the Internet, is the anonymity this environment allows. Since this facilitates cyberbullying can cover almost all areas of the victim's life, that is: educational environment, work, social or loving life. When the identity of the harasser is not known, even if the facts are reported, in many cases it is not enough to open an investigation, identify it and pay for the crime committed. This project proposed a deep learning model Bidirectional Long Short Term Memory (BiLSTM). Thus, this project has designed a method of automatically detecting the Cyberbullying attack cases. Identifies the messages or comments or posts which the BiLSTM model predicts as offensive or negative. Experiments are conducted to test three machine learning and 2 deep learning models that are; (1) GBM, (2) LR, (3) NB, (4) LSTM-CNN and (5) BiLSTM. This project also employed two feature representation techniques word2vector. The results showed that all models performed well on tweet dataset but our proposed BiLSTM classifier outperforms by using both word2vector a among all. Proposed model achieves the highest results using word2vector with 96% Accuracy, 92% Recall and 95% F1-score.

X. FUTURE ENHANCEMENT

For the present, the App works for Twitter, so it can be extended to various other social media platforms like Instagram, Reedit, etc. Currently, only images are classified for NSFW content, classifying text, videos could be an addition. A report tracking feature could be added along with a cross-platform Mobile / Desktop application (Progressive Web App) for the Admin. This model could be implemented for many languages like French, Spanish, Russian, etc. along with India languages like Hindi, Gujarati, etc.

REFERENCES

- [1] M. Di Capua, E. Di Nardo and A. Petrosino, Unsupervised cyberbullying detection in social networks, ICPR, pp. 432-437, doi: 10.1109/ICPR.2016.7899672. (2016)
- [2] E. Englander, E. Donnerstein, R. Kowalski, C. A. Lin, and K. Parti, "Defining cyberbullying," Pediatrics, vol. 140, no. Supplement 2, pp. S148-S151, 2017.
- [3] All the latest cyber bullying statistics and what they mean in 2022. BroadbandSearch.net. (n.d.). Retrieved April 7, 2022, from <https://www.broadbandsearch.net/blog/cyber-bullying-statistics>
- [4] Canada, P. S. (2021, February 5). Government of Canada. Cyberbullying can be against the law - Canada.ca. Retrieved April 7, 2022, from <https://www.canada.ca/en/public-safetycanada/campaigns/cyberbullying/cyberbullying-against-law.html>
- [5] Hatfield, H. (n.d.). Stop school bullying and cyberbullying. WebMD. Retrieved April 7, 2022, from <https://www.webmd.com/parenting/features/prevent-cyberbullyingand-school-bullying>
- [5] J. Wang, K. Fu and C.-T. Lu, "Fine-grained balanced cyberbullying dataset", 2020.
- [6] J. Wang, K. Fu and C. -T. Lu, "SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection," 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 1699- 1708, doi: 10.1109/BigData50022.2020.9378065.
- [7] D. Poeter. (2011) Study: A Quarter of Parents Say Their Child Involved in Cyberbullying. pcmag.com. [Online]. Available: <http://www.pcmag.com/article2/0,2817,2388540,00.asp>
- [8] J. W. Patchin and S. Hinduja, "Bullies move Beyond the Schoolyard; a Preliminary Look at Cyberbullying," Youth Violence and Juvenile Justice, vol. 4, no. 2, pp. 148-169, 2006
- [9] Anti Defamation League. (2011) Glossary of Cyberbullying Terms.adl.org.[Online]. Available: <http://www.adl.org/education/curriculum-connections/cyberbullying/glossary.pdf>



- [10] N. E. Willard, Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress. Research Press, 2007.
- [11] D. Maher, "Cyberbullying: an Ethnographic Case Study of one Australian Upper Primary School Class," Youth Studies Australia, vol. 27, no. 4, pp. 50-57, 2008.
- [12] R.M. Kowalski and S.P. Limber, "Psychological, Physical, and Academic Correlates of Cyberbullying and Traditional bullying," J. Adolescent Health, 2013, vol. 53, no. 1, pp.513-520.



INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.379



ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details