Cyber Bullying Detection And Prevention In Social Networking Sites Using Deep Learning Model

Ms R. Sudha[1], Abrose Banu.A[2], Akshaya.R[3], Yuvarani.G[4]
[1, 2, 3, 4] Dept of Information Technology
[1, 2, 3, 4] PSNA College of Engineering and technology, Dindigul, Tamilnadu

Abstract- *Social networking site is being rapidly increased in recent years, which provides platform to connect people all over the world and share their interests. However, Social Networking Sites is providing opportunities for cyberbullying activities. Cyberbullying is harassing or insulting a person by sending messages of hurting or threatening nature using electronic communication. Cyberbullying poses significant threat to physical and mental health of the victims. Hence, it is essential to monitor user's posts and filter the cyberbullying related post before it is spread. Detection of cyberbullying and the provision of subsequent preventive measures are the main courses of action to combat cyberbullying. This project proposed a system for automatic detection and prevention of cyberbullying tweets and bully from social networks considering the main characteristics of cyberbullying such as Intention to harm an individual, repeatedly and over time and using abusive curl language or cyberbullying using BiLSTM algorithm. The proposed model is capable to detect cyberbullying content on Twitter automatically. This approach is based on a bag of words and TFIDF (term frequency-inverse document frequency) approach. These features are used to train deep learning classifiers. The bully level encapsulates the level of hate in a given digital environment. We present methods to automatically determine the bullylevel, and block post and users of the post by the utilizing transfer learning on pre-trained language models with annotated data to create automated cyberbullying detectors. We evaluate our approaches on a set of websites and discussion forums.*

I. INTRODUCTION

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.

The most common places where cyberbullying occurs are:

- Social Media, such as Facebook, Instagram, Snapchat, and Tik Tok
- Text messaging and messaging apps on mobile or tablet devices
- Instant messaging, direct messaging, and online chatting over the internet
- Online forums, chat rooms, and message boards, such as Reddit
- Email
- Online gaming communities

II. LITERATURE REVIEW

1.Andrea Pereraa ,Pumudu Fernando," Accurate Cyberbullying Detection and Prevention on Social Media",2021

Objective:

The aim of this project isa system for automatic detection and prevention cyberbullying considering the main characteristics of cyberbullying such as Intention to harm an individual, repeatedly and over time and using abusive curl language or cyberbullying using supervised machine learning.

Methodology:

The usage of digital/social media is increasing day by day with the advancement of technology. People in the twenty-first century are being raised in an internet-enabled world with social media. Communication has been just one button click. Even though there are plenty of opportunities with digital media people tend to misuse it. People spread hatred toward a person in social networking. Cyberbullying affects people in different aspects. It doesn't affect only for health, there are more different aspects which will lead life to a threat. Cyberbullying is a worldwide modern phenomenon which humans cannot avoid hundred percent but can be prevented. Most existing solutions have shown techniques/approaches to detect cyberbullying, but they are

not freely available for end-users to use. They haven't considered the evolution of language which makes a big impact on cyberbullying text. This article proposed a TF-IDF (Term Frequency, Inverse Document Frequency) by using TFIDF which can measure the importance of words in a document and Common words such as "is", "am" do not affect the results due to IDF. This article used Support Vector Machines (SVM), A well-known efficient binary classifier to train the model. Logistic regression was used to select the best combination of features. SVM algorithm, training data is used to learn a classification function. It can classify new data not previously seen in one of the two categories. It separates the training data set into two categories using a large hyperplane. Logistic regression is a linear classifier that predicts the probabilities.

Merits:

- Accuracy is high.
- It can classify new data not previously seen.
- Efficiency is high.

Demerits:

- High cost.
- Long and tedius job.

2.Jalal Omer Atoum," Cyberbullying Detection Through Sentiment Analysis",2020

Objective:

The aim of this project is a SA model for identifying cyberbullying texts in Twitter social media.

Methodology:

In recent years with the widespread of social media platforms across the globe especially among young people, cyberbullying and aggression have become a serious and annoying problem that communities must deal with. Such platforms provide various ways for bullies to attack and threaten others in their communities. Various techniques and methodologies have been used or proposed to combat cyberbullying through early detection and alerts to discover and/or protect victims from such attacks.This article proposed an approach to detect cyberbullying from Twitter social media platform based on Sentiment Analysis that employed machine learning techniques; namely, Naïve Bayes and Support Vector Machine. The data sets used in this article is a collection of tweets that have been classified into positive, negative, or neutral cyberbullying. Before training and testing such machine learning techniques, the collected set of tweets have gone through several phases of cleaning, annotations, normalization, tokenization, named entity recognition, removing stopped words, stemming and n-gram, and features selection. The results of the conducted experiments have indicated that SVM classifiers have outperformed NB classifiers in almost all performance measures over all language models. Specifically, SVM classifiers have achieved an average accuracy value of 92.02%, while, the NB classifiers have achieved an average accuracy of 81.1 on the 4-gram language model.

Merits:

- It has better performance measures than NB classifiers on such tweets.
- Low cost and low time consuming.

Demerits:

- Suffer from an inability to detect indirect language harassment.
- Accuracy is low.

3.MdManowarul Islam; Md Ashraf Uddin; Linta Islam; ArnishaAkter; Selina Sharmin; Uzzal Kumar Acharjee," Cyberbullying Detection on Social Networks Using Machine Learning Approaches",2020

Objective:

The aim of this project is to design and develop an effective technique to detect online abusive and bullying messages by merging natural language processing and machine learning.

Methodology:

The use of social media has grown exponentially over time with the growth of the Internet and has become the most influential networking platform in the 21st century. However, the enhancement of social connectivity often creates negative impacts on society that contribute to a couple of bad phenomena such as online abuse, harassment cyberbullying, cybercrime and online trolling. Cyberbullying frequently leads to serious mental and physical distress, particularly for women and children, and even sometimes force them to attempt suicide. Online harassment attracts attention due to its strong negative social impact. Many incidents have recently occurred worldwide due to online harassment, such as sharing private chats, rumours, and sexual remarks. Therefore, the identification of bullying text or message on social media has

gained a growing amount of attention among researchers. The purpose of this article is to design and develop an effective technique to detect online abusive and bullying messages by merging natural language processing and machine learning. Two distinct features, namely Bag-of - Words (BoW) and term frequency-inverse text frequency (TFIDF), are used to analyse the accuracy level of four distinct machine learning algorithms.

Merits:

- The words that occur more frequently should be given more importance as they are more useful for classification.
- Classifier is a supervised learning model which provides accurate result because several decision trees are merged to make the outcome.

Demerits:

- Weak-supervision loss.
- Classifier accuracy is low.

4.Jamal,Alasadi; RamanathanArunachalam; Pradeep K. Atrey; Vivek K. Singh," A Fairness-Aware Fusion Framework for Multimodal Cyberbullying Detection,2020

Objective:

The aim of this project is a fairness-aware fusion framework that ensures that both fairness and accuracy remain important considerations when combining data coming from multiple modalities.

Methodology:

Recent reports of bias in multimedia algorithms (e.g., lesser accuracy of face detection for women and persons of colour) have underscored the urgent need to devise approaches which work equally well for different demographic groups. Hence, here posit that ensuring fairness in multimodal cyberbullying detectors (e.g., equal performance irrespective of the gender of the victim) is an important challenge**.** This article describes one of the first attempts at a Bayesian fusion framework that not only optimizes for accuracy but also considers fairness. The framework takes into account the accuracy and the fairness score for each modality to assign them weights. The weights of each modality and the agreement between them is used to come up with optimal decisions that balance accuracy and fairness. The results of applying the framework to a multimodal (visual + textual) cyberbullying detection problem demonstrate the efficacy of the approach in yielding high levels of both accuracy and bias. The results pave way for a more accurate and fair approach for cyberbullying detection, which would provide equitable opportunities to different groups in improving their quality of life.

Merits:

- It ensuring both accuracy and fairness.
- Efficiency is high.
- Speed is high to detect the offensive words.

Demerits:

- Algorithms that perform differently for different groups.
- Many issue of fairness for not detecting the cyberbullying cases.
- Time consuming process.

5.Zehua Zhao; Min Gao; Fengji Luo; Yi Zhang; QingyuXiong," LSHWE: Improving Similarity-Based Word Embedding with Locality Sensitive Hashing for Cyberbullying Detection" 2020.

Objective:

The aim of this project is a word embedding method called LSHWE to solve this limitation, which is based on an idea that deliberately obfuscated words have a high context similarity with their corresponding bullying words.

Methodology:

This articleproposes a similarity-based word embedding method LSHWE to solve the “deliberately obfuscated words” problem in cyberbullying detection task. LSHWE has two steps. Firstly, for a given corpus, it generates: (a) a co-occurrence matrix C; (b) a rare word list R; (c) a nearest neighbour list NL obtained by locality sensitive hashing; and (d) a nearest neighbour matrix N. Secondly, an LSH-based auto encoder is used to learn the word vectors according to C and N. The proposed embedding method has two characteristics: (1) LSHWE can represent well on rare words. LSHWE is a global similarity-based word embedding method thus the representations of rare words learnt through LSHWE can be as close as possible to their corresponding words’ representations; and (2) LSHWE is a highly efficient algorithm. This method uses an approximate nearest neighbour search method to search the top-k nearest neighbours instead of exact nearest neighbour search methods, which can greatly reduce the running time. This article design experiments from three aspects: effectiveness of LSHWE on

cyberbullying detection task, algorithm efficiency and parameter sensitivity. Experiment results demonstrate that LSHWE can alleviate the “deliberately obfuscated words” problem and is highly efficient on large-scale datasets.

Merits:

- It can alleviate the “deliberately obfuscated words” problem.
- Highly efficient on large-scale datasets.

Demerits:

- Running time is long.
- High Cost.

III. EXSITING 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).

Random Forest

RF is a tree based classifier in which input vector generated trees randomly. RF uses random features, to create multiple decision trees, to make a forest. Then class labels of test data are predicted by aggregating voting of all trees. Higher weights are assigned to the decision trees with low value error. Overall prediction accuracy is improved by considering trees with low error rate.

Support Vector Machine

The Support vector machine (SVM) is understood that executes properly as sentiment analysis. SVM typifies preference, confines and makes usage of the mechanisms for the assessment and examines records, which are attained within the index area. Arrangements of vectors for every magnitude embody crucial details. Information (shown in form of vector) has been arranged in type to achieve this target. Next, the border is categorized in two training sets by stratagem. This is a long way from any area in the training samples. Support-vector machines in machine learning includes focused learning models connected to learning evaluations which inspect material that is exploited to categorize, also revert inspection.

Naive Bayes

Ordering approach, Naive Bayes(NB), with sturdy (naive) independent assumptions among stabilities, depends on Bayes' Theorem. NB classifier anticipates that the proximity of a specific element of class that is confined to the closeness of a couple of different variables. For instance, a natural organic product is presumably viewed as an apple, if its shading is dark red, if type of it is round and it is roughly 3 creeps in expansiveness. In machine learning, Naive Bayes classifiers are a gathering of essential "probabilistic classifiers" considering applying Bayes' speculation with gullible opportunity assumptions between the features. They are considered as the minimum problematic Bayesian network models.

Gradient Boosting Machine

GBM is a ML based boosting model and is widely being used for regression and classification tasks, which works by a model formed by ensemble of weak prediction models, commonly decision trees. In boosting, weak learners are converted to strong learners. Every new generated tree is a modified form of previous one and use gradient as loss function. Loss calculate the efficiency of model coefficients fitting over underlying data. Logically loss function is used for model optimization.

Logistic Regression

In LR class probabilities are estimated on the basis of output such as they predict if the input is from class X with probability x and from class Y with probability y. If x is greater than y, then predicted output class is X, otherwise Y. Insight, a logistic approach used for demonstrating the probability of a precise group or else, occurrence is obtainable, e.g., top/bottom, white/black, up/down, positive/negative or happy/unhappy. This is able to stretch out and to show a small number of classes about events, for example, to make a decision if an image includes a snake, hound, deer, etc., every article being famous in the image would be appointed a probability wherever in the series of 0 and 1 with whole addition to one.

Stochastic Gradient Descent

Gradient Descent's types include Stochastic Gradient Descent (SGD). SDGD is an iterative strategy for advancing a target work with appropriate perfection properties (for example differentiable or sub differentiable). Degree of advancement is calculated by it in light of development of alternative variables. It is very well, may be viewed as a

stochastic guess of inclination plummet advancement, since it replaces the genuine angle (determined from the whole informational index) by a gauge thereof (determined from an arbitrarily chosen subset of the information).

Voting Classifier

Voting Classifier(VC) is a cooperative learning which engages multiple individual classifiers and combines their predictions, which could attain better performance than a single classifier. It has been exhibited that the mixture of multiple classifiers could be more operative compared to any distinct ones. The VC is a meta classifier for joining tantamount or hypothetically exceptional ML classifiers for order through greater part throwing a voting form. It executes "hard" and "soft" casting a ballot. Hard voting gives the researcher the chance to foresee the class name in place of the last class mark that has been anticipated often through models of characterization.

Disadvantages

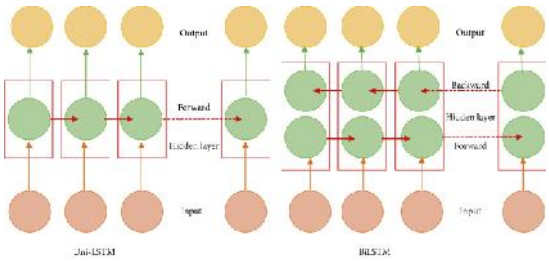
- Process of reporting such cases is long, tedious job.
- Difficult to track.
- Most of the cyberbullying cases go unreported.
- Low accuracy.
- Time consuming process.
- Problem is not automatically detected and not promptly report bullying message.
- Response time is slow.
- Basic features and common classifier accuracy is low.
- Data are manually labelled using online services or custom applications.
- Usually data limited only to a small percentage.

IV. PROPOSED SYSTEM

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

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 LSTMs may not make sense for all prediction problems but can offer benefits in terms of better results to those domains where it is appropriate.



BiLSTM

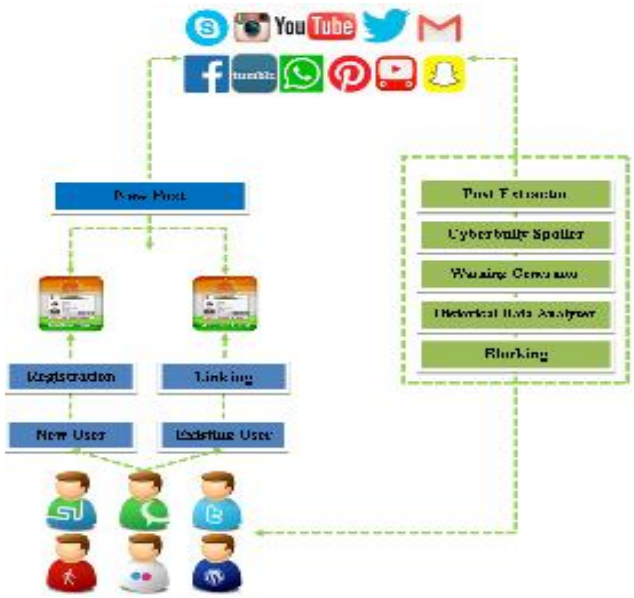
Bidirectional LSTM (BiLSTM) is a recurrent neural network used primarily on natural language processing. Unlike standard LSTM, the input flows in both directions, and it's capable of utilizing information from both sides. It's also a powerful tool for modelling the sequential dependencies between words 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.
- Auto report generator generates a simple report for probable accusers.
- Several analytics and report can be sent to the crime department.
- Accuracy is high.
- Foul language on any given page, removes it, and can highlight words as well,
- This method detects the offensive post or messages it block that user id.
- 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.

V. METHODOLOGY

In the proposed framework shown in Figure 5.1., the process of detecting cyberbully activities begins with input dataset from social network. Input is text conversation collected from twitter. Input is given to data pre-processing which is applied to improve the quality of the project data and subsequent analytical steps, this includes removing stop words, extra characters and hyperlinks. After performing pre-processing on the input data, it is given to Feature Extraction. Feature Extraction is done to obtain features like Noun, Adjective and Pronoun from the text and statistics on occurrence of word (frequency) in the text. The cyberbullying words are given as training dataset. With the training dataset the preprocessed online social network conversation is tested for bullying word presence. Feature Vector distance algorithm detects the cyberbully words present in the conversation and displays it. For cyberbully Classification, BiLSTM is used.



Dataset Description

Twitter dataset used in this experiment is scrapped from Kaggle repository. Twitter is used excessively, mainly because it has an easy-to-access API to collect data. All these datasets are manually labelled and publicly available. From this microblogging platform, 49,692 tweets were collected and manually annotated. The tweets were collected by search of terms which refer to religious, sexual, gender, and ethnic minorities. The dataset is divided into two categories: bullying and non-bullying.

Non-bullying Text: This type of comments or posts are non-bullying or positive comments. For example, thecomment like" This photo is very beautiful" is positive and non-bullying comments.

Bullying Text: This type belongs to bully type comments or harassment's. For example, "go away bitch" is a bullying text or comment and we consider as negative comment.

In total 39747 were labelled as bullying and 7945 were labelled as non - bullying. The remaining tweets were labelled as neither.

No.	Category	Tweets	Percentile
1	Bullying	39747	82%
2	Non-Bullying	7945	18%

Cyberbullying Tweets Description

Threat/Blackmail: expressions containing physical or psychological threats or indications of blackmail.

Insult: expressions meant to hurt or offend the victim.

- General insult: general expressions containing abusive, degrading or offensive language
- that are meant to insult the addressee.
- Attacking relatives: insulting expressions towards relatives or friends of the victim.
- Discrimination: expressions of unjust or prejudicial treatment of the victim. Two types
- of discrimination are distinguished (i.e., sexism and racism). Other forms of discrimination
- should be categorized as general insults.

Curse/Exclusion: expressions of a wish that some form of adversity or misfortune will befall the victim and expressions that exclude the victim from a conversation or a social group.

Defamation: expressions that reveal confident or defamatory information about the victim to a large public.

Sexual Talk: expressions with a sexual meaning or connotation. A distinction is made between innocent sexual talk and sexual harassment.

Defense: expressions in support of the victim, expressed by the victim himself or by a bystander.

- Bystander defense: expressions by which a bystander shows support for the victim ordiscourages the harasser from continuing his actions.
- Victim defense: assertive or powerless reactions from the victim.

Encouragement to the harasser: expressions in support of the harasser.

Other: expressions that contain any other form of cyberbullying-related behavior than the ones described here.

Role-allocation in cyberbullying

- **Harasser or Bully**: person who initiates the bullying.
- **Victim**: person who is harassed.
- **Bystander-defender**: person who helps the victim and discourages the harasser from continuing his actions.
- **Bystander-assistant**: person who does not initiate, but helps or encourages the harasser.

Modules Description

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 REST API which was consumed by a Web application built using Python's Flask 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.

Aadhar User Account Management

- **New User**

Create user account with aadhar number.

- **Existing User**

The existing users of Facebook will also have to upload a scanned copy of their Aadhar Card. If they fail to do so, their profile will be suspended within the next 15 days.

Cyberbullying Analysis API

In this module we developed the API 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).

VI. EXPERIMENTAL RESULTS AND DISCUSSION

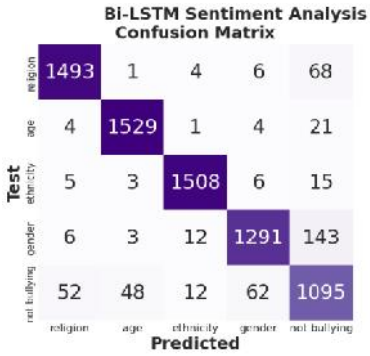
6.1. Evaluation Metrics

6.1.1. Confusion Matrix

The detection of spam emails can be evaluated by different performance measures. Confusion Matrix is being used to visualise the detection of the emails for models Several measurements are used for performance evaluation of classifiers like accuracy, precision, recall, and f-score. These measurements are computed by a confusion matrix, which is composed of four terms. Confusion matrix can be defined as below:

- True positive (TP): are the positive values correctly classified as positive.
- True Negative (TN): are the negative values correctly classified as negative.
- False Positive (FP): are the negative values incorrectly classified as positive.
- False Negative (FN): are the positive values incorrectly classified as negative.

For the performance evaluation of our proposed model, we use the following metrics.

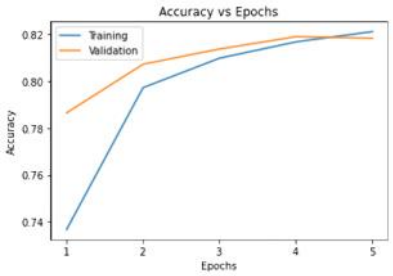


6.1.2. Accuracy

The accuracy measure is the ratio of the number of bully users detected to the total number of bullies. It does not perform well with imbalanced data sets

$$\text{AccuracyCM} = \frac{\text{\# of detected bullies}}{\text{total number of bullies}}$$

training score :0.991249719542293
testing score :0.979372197309417



6.1.3. Precision

Precision is evaluation metrics used in binary classification tasks.Precision is the measure of exactness.

$$\text{Precision} = \frac{\text{\# of true bullies detected}}{\text{total number of detected users}}$$

In simple terms, high precision means that an algorithm returned substantially more bully users

6.1.4. Recall

The recall is a fraction of the predicted correctly classified applications to the total number of applications classified correctly or incorrectly. Recall is the measure of completeness.

$$\text{Recall} = \frac{\text{\# of true bullies detected}}{\text{total number of true bullies}}$$

whereas high recall means that an algorithm returned most of the bullies.

6.1.5.F-Score

F-score is the harmonic mean of precision and recall. It symbolizes the capability of the model for making fine distinctions. F1 Measure is the harmonic mean between precision and recall. The range for F1 is [0, 1]. It measures how many bullies are identified correctly and how robust it is. Mathematically, it can be expressed as

F1-Score=2*(Precision*Recall)/ (Precision + Recall)

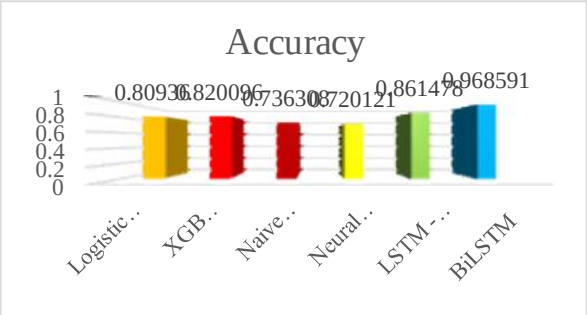
F1 Measure attempts to find a balance between precision and recall. The greater the F1 Measure, the better is the performance of our approach.

6.2. Results and Analysis

The results and comparisons of different classifiers after data training and testing are presented in this section. We gathered 49799tweets from the online resource ‘kaggle’ and translated them into English using the python library Googletrans, which uses the Google Translate Ajax API. 42797tweets were used to train various ML and DL models. One seven thousand tweets were used for testing in order to quantify accuracy and assessment metrics. As explained about evaluation measures in chapter 9, we have evaluated accuracy, precision, recall, and f-measures that are evaluation measures measured using LR, XGBM and Naive Bayes, LSTM-CNN and BiLSTM. Finally, using various graphs, a comparison of models is presented below. The findings in Table 4 show that the deep learning algorithm (BiLSTM) is a stronger method for detecting cyberbullying tweet classification, with high accuracy of 98.4%.

No	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.809360	0.816488	0.809360	0.812462
1	XGB Classifier	0.820096	0.829113	0.820096	0.823191
2	Naive Bayes Classifier	0.736308	0.723307	0.736308	0.708342
3	Neural Network	0.720121	0.808795	0.752411	0.778392
4.	LSTM - CNN	0.861478	0.895267	0.827423	0.852364
5.	BiLSTM	0.968591	0.986842	0.958647	0.984567

In the mentioned Table 4, we have compared the accuracy of four different ML and DL models. We can see that the DL model (BiLSTM) is the most accurate among all the models, but it takes a long time to train. ML models like LR, XGB and Naive Bayes are around the same accuracy percentage lower than LSTM/CNN and BiLSTM, which is also a DL model and has the lowest accuracy percentage. Figure shows accuracy comparison of ML and DL models.



VII. 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 projecthasdesigned 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 then it blocks that person id,then the admin can create automated reports and send to the concern department. 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 Tfand TF-IDF. The results showed that all models performed well on tweet dataset but our proposed BiLSTM classifier outperforms by using both TF and TF-IDF among all. Proposed model achieves the highest results using TF-IDF with 96% Accuracy, 92% Recall and 95% F1-score.

VIII. FUTURE ENHANCEMENT

For the present, the bot 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] A. S. Srinath, H. Johnson, G. G. Dagher and M. Long, "BullyNet: Unmasking cyberbullies on social networks", IEEE Trans. Computat. Social Syst., vol. 8, no. 2, pp. 332-344, Apr. 2021.

[2] Z. L. Chia, M. Ptaszynski, F. Masui, G. Leliwa and M. Wroczynski, "Machine learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection", Inf. Process. Manage., vol. 58, no. 4, Jul. 2021.

[3] N. Yuvaraj, K. Srihari, G. Dhiman, K. Somasundaram, A. Sharma, S. Rajeskannan, et al., "Nature-inspired-based approach for automated cyberbullying classification on multimedia social networking", Math. Problems Eng., vol. 2021, pp. 1-12, Feb. 2021.

[4] R. R. Dalvi, S. B. Chavan and A. Halbe, "Detecting a Twitter cyberbullying using machine learning", Ann. Romanian Soc. Cell Biol., vol. 25, no. 4, pp. 16307-16315, 2021.

[5] N. Yuvaraj, V. Chang, B. Gobinathan, A. Pinagapani, S. Kannan, G. Dhiman, et al., "Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification", Comput. Electr. Eng., vol. 92, Jun. 2021.

[6] A. Al-Hassan and H. Al-Dossari, "Detection of hate speech in Arabic tweets using deep learning", Multimedia Syst., Jan. 2021.

[7] Y. Fang, S. Yang, B. Zhao and C. Huang, "Cyberbullying detection in social networks using bi-GRU with self-attention mechanism", Information, vol. 12, no. 4, pp. 171, Apr. 2021.

[8] B. A. Talpur and D. O'Sullivan, "Multi-class imbalance in text classification: A feature engineering approach to detect cyberbullying in Twitter", Informatics, vol. 7, no. 4, pp. 52, Nov. 2020.

[9] A. Agarwal, A. S. Chivukula, M. H. Bhuyan, T. Jan, B. Narayan and M. Prasad, "Identification and classification of cyberbullying posts: A recurrent neural network approach using under-sampling and class weighting" in Neural Information Processing, Cham, Switzerland:Springer, vol. 1333, pp. 113-120, 2020.

[10] C. Iwendi, G. Srivastava, S. Khan and P. K. R. Maddikunta, "Cyberbullying detection solutions based on deep learning architectures", Multimedia Syst., 2020.

[11] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, "XBully: Cyberbullying detection within a multi-modal context," in Proc. 12th ACM Int. Conf. Web Search Data Mining, Jan. 2019, pp. 339–347.

[12] C. Van Hee et al., “Automatic Detection of Cyberbullying in Social Media Text.” 2018.
[13] M. Rezvan, S. Shekarpour, L. Balasuriya, K. Thirunarayan, V. Shalin, and A. Sheth, “A Quality Type-aware Annotated Corpus and Lexicon for Harassment Research,” pp. 33–36, 2018.
[14] A. H. Alduailej and M. B. Khan, “The challenge of cyberbullying and its automatic detection in Arabic text,” 2017 Int. Conf. Comput. Appl. ICCA 2017, pp. 389–394, 2017.
[15] A. Power, A. Keane, B. Nolan, and B. O. Neill, “A lexical database for public textual cyberbullying detection,” Rev. Lenguas Para Fines Específicos, vol. 2, pp. 157–186, 2017.
[16] M. Drahošová and P. Balco, “ScienceDirect The analysis of advantages and disadvantages of use of social media the analysis of advantages and disadvantages of use of social media in European Union in European Union,” Procedia Comput. Sci., vol. 109, pp. 1005–1009, 2017.
[17] R. Zhao, A. Zhou, and K. Mao, “Automatic detection of cyberbullying on social networks based on bullying features,” in Proc. 17th Int. Conf. Distrib. Comput. Netw., Jan. 2016, pp. 1–6.
[18] V. K. Singh, Q. Huang, and P. K. Atrey, “Cyberbullying detection using probabilistic socio-textual information fusion,” in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM), Aug. 2016, pp. 884–887.
[19] A. Squicciarini, S. Rajtmajer, Y. Liu, and C. Griffin, “Identification and characterization of cyberbullying dynamics in an online social network,” in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining, Aug. 2015, pp. 280–285.