ISSN 2063-5346

TOWARDS PRIVACY PROTECTION FOR USERS OF SOCIAL MEDIA BY CYBERBULLYING PREDICTION USING MACHINE LEARNING

Samreen Sultana[1] Mr. Syed Ahmeduddin[2]

Article History: Received: 10.05.2023 Revised: 29.05.2023 Accepted: 09.06.2023

Abstract

Cyberbullying is a major problem encountered on internet that affects teenagers and also adults. It has lead to mishappenings like suicide and depression. Regulation of content on Social media platorms has become a growing need. The following study uses data from two different forms of cyberbullying, hate speech tweets from Twittter and comments based on personal attacks from Wikipedia forums to build a model based on detection of Cyberbullying in text data using Natural Language Processing and Machine learning. Three methods for Feature extraction and four classifiers are studied to outline the best approach. Increasing internet use and facilitating access to online communities such as social media have led to the emergence of cybercrime. Cyberbullying, a new form of bullying that emerged recently with the development of social networks, means sending messages that include slanderous statements, or verbally bullying other people in front of rest of the online community. The characteristics of online social networks enable cyberbullies to access places and countries that were previously unattainable for using SVM we are going to identify cyberbullying in twitter. Objectives of this implementation written in objective section. Image character with the help of OCR will be done by us to find image - based cyberbullying the impact on individual basis thus will be checked on dummy system. Machine learning and natural language processing techniques to identify the characteristics of a cyberbullying exchange and automatically detect cyberbullying by matching textual data to the identified traits. On the basis of our extensive literature review, we categorise existing approaches into 4 main classes, namely supervised learning, lexicon-based, rule-based, and mixed-initiative approaches. Supervised learning-based approaches typically use classifiers such as SVM and Naïve Bayes to develop predictive models for cyberbullying detection. We are using natural language processing techniques and machine learning methods namely, Bayesian logistic regression, random forest algorithm, support vector machines have been used to determine cyberbullying. The user on social media has saw a boom in recent time with increase in number of users of the net and emerges as the major networking platform of our era. But it also has its own repercussions on society and the mental health of a person such as online abuse, harassment, scamming, private information leaks and trolling. Cyberbullying affects a person both physically and mentally, particularly for girls and students, and sometimes escalated to their suicide. Online harassment has a huge bad impact on society. No of cases have occurred in different parts due to online bullying, such as sharing private information, abusing someone online, and racial discrimination. So, there is a need for identification of bullying on social apps and this has become a major concern all over the world. Our motive for this research is to compare different techniques to find out the most effective technique to detect online harassment by merging NLP (natural language processing) with ML (machine learning).

Index Terms — Support vector machines, Deep learning, Computational modeling, Blogs, Cyberbullying, Pressing, Intelligent systems.

[1]Research Scholar, Dept. of Computer Science and Engineering, Lords Institute of Engineering & Technology, Hyderabad, Telangana

[2]Asst Professor, Dept. of Computer Science and Engineering, Lords Institute of Engineering & Technology, Hyderabad, Telangana

DOI:10.48047/ecb/2023.12.9.29

I. INTRODUCTION

Modern young people ("digital natives") have grown in an era dominated by new technologies where communications are pushed to quite a real-time level, and pose no limits in establishing relationships with other people or communities. The fast growing use of social networking sites among the teens have made them vulnerable to get exposed to bullying. Comments containing abusive words effect psychology of teens and demoralizes them. In this work we have devised methods to detect cyberbullying using supervised learning techniques. Cyberbullying is the use of technology as a medium to bully someone. Although it has been an issue for many years, the recognition of its impact on young people has recently increased. Through machine learning, we can detect language patterns used by bullies and their victims, and develop rules to automatically detect cyberbullying content. Moreover, they are a place where people engage in social interaction, offering the existing friendships. On the negative side however, social establish new relationships and maintain existing friendships. On the negative side however, social media increase the risk of children being confronted with threatening situations including grooming or sexually transgressive behaviour, signals of depression and suicidal thoughts, and cyberbullying. Users are reachable 24/7 and are often able to remain anonymous if desired: this makes social media a convenient way for bullies to target their victims outside the school yard. The detection of cyberbullying and online harassment is often formulated as a classification problem. Techniques typically used for document classification, topic detection, and sentiment analysis can be used to detect electronic bullying using characteristics of messages, senders, and the recipients. It should, however, be noted that cyberbullying detection is intrinsically more difficult than just detecting abusive content. Additional context may be required to prove that an individual abusive message is part of a sequence of online harassment directed at a user for such a message to be labelled as cyberbullying. The growth of cyberbullying activities is increasing as equally as the growth of social networks. Cyberbullying activities poses a significant threat to mental and physical health of the victims. Project about detection of bullying is present but implementation for monitoring social network to detect cyberbullying activities is less. Hence, the proposed system focuses on detecting the presence of cyberbullying activity in social networks using natural processing language.

Now over ever technology has become associate integral a region of our life. With the evolution of the net. Social media is trending recently. but as all the alternative things misusers will come out usually late sometime early but there will be clearly. Now Cyberbullying is common recently. Sites for social networking ar terrific tools for communication at intervals individuals. Use of social networking has become widespread over the years, though, usually people notice immoral and unethical ways that during which of negative stuff. we've got a bent to visualize this happening between teens or usually between young adults. one of the negative stuffs they're doing is bullying each other over the net.

II. SYSTEM ANALYSIS

Problem Statement:

Social networks Networks give us great opportunities to communicate, and also increase the vulnerability of young people to threatening situations on the Internet. Cyberbullying on social media is a global phenomenon due to its large number of active users. The trend shows that social network cyberbullying is increasing rapidly day by day. Recent research shows that cyberbullying is a growing problem

among young people. Successful prevention depends on the proper detection of potentially harmful messages, and the information overload of the Internet requires intelligent systems to automatically detect potential hazards. Therefore, in this project, we will focus on creating a model to automatically detect cyberbullying in social media text by simulating messages created by social media bullying.

AIM OF THE PROJECT

The main aim of the detecting the cyberbullying model will help to improve manual monitoring for cyberbullying on social networks. In this project we fetch the tweets from twitter accounts and preprocess the tweets and images and applying generated model will detect the cyberbullying or not. The objectives of the systems development and event management are: Collect the data set of bullying words and preprocess it and apply natural language processing and then machine learning algorithms Generate different machine learning algorithm model. Fetch the tweets from twitter account and preprocess it. Apply generated model on the fetched tweets and get final output cyberbullying or not.

SCOPE OF THE PROJECT

Cyberbullying is the use of electronic communication to bully a person by sending harmful messages using social media, instant messaging or through digital messages. Cyberbullying can be very damaging to adolescents and teens. It can lead to anxiety, depression, and even suicide. Also, once things are circulated on the Internet, they may never disappear, resurfacing at later times to renew the pain of cyberbullying. Cyberbullying can be very damaging to adolescents and teens. It can lead to anxiety, depression, and even suicide. Also, once things are circulated on the Internet, they may never disappear, resurfacing at later times to renew the pain of cyberbullying. So overcome these issues detecting the cyberbullying is very important in now a days which will help to stop cyberbullying on social media networks. search and find the the dataset and download it for train the model. After downloading first we will pre process the data and then transferred to Tf Idf. It then trains the dataset using Naive Bayes, Support Vector Machine (SVM), and DNN algorithms and creates models separately. Next, we will develop a web application using the FLASK framework. It imports live tweets from Twitter, then applies the generated model to the imported tweets and checks whether text or images are cyberbullying. For this purpose, we use python as backend, Mysql as database and html, css, javascript etc as frontend.

METHODOLOGY

We will develop this project with the help of python and web technology. Using html and CSS, we will design and develop the web interfaces for the project. Then after preparing the web interfaces, we will search and download the dataset that we need to classify. After downloading the dataset, we will pre-process the data and then transfer to Tf-Idf. Then we will generate codes for the machine learning algorithms (Naive Bayes, Decision Tree, Random Forest, SVM, DNN Model) using python. So here, we are using python as backend and for frontend html, CSS etc.

The real-world posts or text contain number of unnecessary symbols or texts. For instance, emojis and symbols are not needed to detect cyber bullying. Hence, first they are removed and then machine learning algorithms are applied for the identification of bullying text. In this phase, the task is to remove unnecessary characters like symbols, emojis, numbers, links etc. And after those two important features of the text is prepared:

Bag-of-Word: The machine learning algorithms are not going to work directly with texts. So, we have to convert them into some other form like numbers or vectors before applying machine learning algorithm to them. In this way the data is converted by Bag-of-Words (BOW) so that it can be ready to use in next round.

TF-IDF: One of the important features to be considered is this. TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure to know the importance that a word carries in a document.

Proposed System:

Cyber bullying detection is solved in this project as a binary classification problem where we are detecting two majors form of Cyber bullying: hate speech on Twitter and Personal attacks on Wikipedia and classifying them as containing Cyber bullying or not.

We may need made an effort to identify cyberbullying in regional language using text classification algorithms. Though we have used various text based classification algorithms such as SVM, Logistic Regression, Passive Aggressive and Random forest but for future purposes, other machine learning models or practices such as CNN and even NLP can be used for the given dataset that we have to work on.

Advantages:

- Accurate review prediction
- Application that will be created uses web frame Flask which brings flexible and wide use access control prediction system

III. PROPOSED MODULAR IMPLEMENTATION

Algorithm/ Technique Used:

Data Preparation for common classification algorithms:

The data needs to prepared in the following way before feeding it to the common classification algorithms

1. Read the dataset Cyberbullying_Tweets.csv
2. Clean the dataset

a. Remove punctuation marks

b. Remove URLs

c. Remove special characters

d. Convert the text to lower case

e. Apply Lemmatization technique to remove tenses from texts

f. Apply Stemming to remove prefixes or suffixes and get the root words

g. Remove stop words

h. Drop duplicate rows

i. Convert categorical columns to numerical columns using label encoding

3. Creating Train and Test Split arrays
4. Ensure that the datasets are balanced
5. Apply TF-IDF vectorization on Train and Test Data
6. Feed the Train and Test Data to various machine Learing Algorithms
7. The results are as follows:

Models	**AccuracyScores**
Random Forest	0.933961
XGBoost	0.931914
LightGBM	0.931299
Support Vector Machine	0.921880
Gradient Boosting	0.921470
Decision Tree	0.913484
Multilayer Perceptron	0.905703
AdaBoost	0.898741
Naive Bayes	0.819187

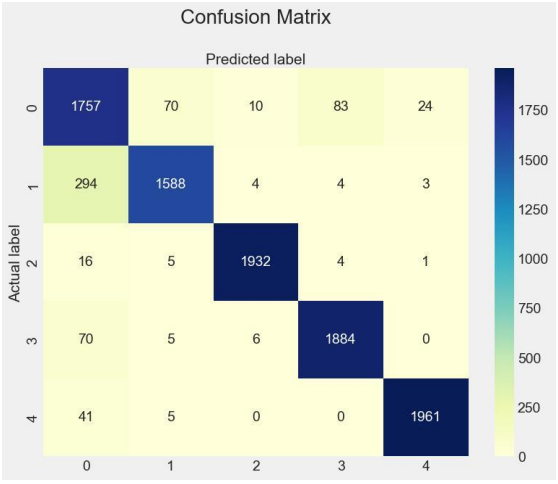
Hence we create our final classifier using Random Forest Algorithm.

Final Classifier:

RF Accuracy: 0.9339612982492065

Training Accuracy Score: 100.0%

Validation Accuracy Score: 93.4%



RF Classification Report:

	precision	recall	f1-score	support
0	0.81	0.90	0.85	1944
1	0.95	0.84	0.89	1893
2	0.99	0.99	0.99	1958
3	0.95	0.96	0.96	1965
4	0.99	0.98	0.98	2007
accuracy		0.93	9767	
macro avg	0.94	0.93	0.93	9767
weighted avg	0.94	0.93	0.93	9767

Below is the proposed modular implementation of the project.

Admin Module:

The admin of the system is responsible for the activities like:

1. Login
2. Upload cyber bullying tweets dataset that was downloaded from Kaggle
3. Exploratory Data Analysis
4. Data Preprocessing

a. Clean the dataset

i. Remove punctuation marks
ii. Remove URLs
iii. Remove special characters
iv. Convert the text to lower case
v. Apply Lemmmatization technique to remove tenses from texts
vi. Apply Stemming to remove prefixes or suffixes and get the root words
vii. Remove stop words
viii. Drop duplicate rows
ix. Convert categorical columns to numerical columns using label encoding

b. Creating Train and Test Split arrays
c. Ensure that the datasets are balanced
d. Apply TF-IDF vectorization

5. Feeding the dataset to multiple classification algorithms

a. Random Forest
b. SVM
c. Multinomial Naïve Baiyes
d. XGBoost
e. Light GBM
f. Adaboost
g. Gradient Boosting
h. Decision Trees
i. Multilayer perceptron

6. Creation of model using the Random Forest Algorithm

IV. PROJECT EXECUTION

Admin Login:

This is the login page for the admin module. The admin need to login into the system with his credentials in order to perform operations like uploading the dataset, Training the dataset, Exploratory data Analysis of the dataset, Feeding the dataset to different Machine learning Algorithms to find the Algorithm that can meet the best accuracy and Create a model that can be hosted on the Flask Application to be used by the users.

Fig: Admin Login

Upload Dataset:

On this page, the administrator of the system can upload datasets that are used for training the machine learning models. The admin has to select the file by clicking on the Choose file button and click on the upload button to upload the file to the server. Once the upload is complete, a success message would be displayed that the file is successfully uploaded. For this project we are using cyberbullying_tweets.csv as a dataset.

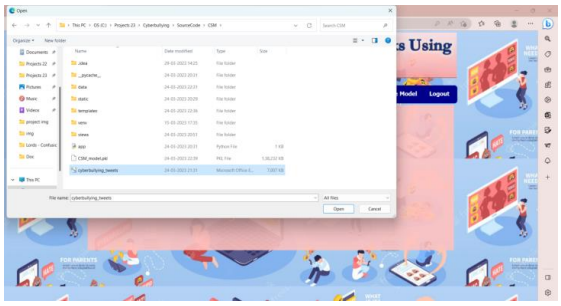


Fig: Upload Dataset & File Uploaded Successfully.

Data Analysis:

Exploratory Data Analysis is performed on the dataset in order to clean the dataset for any missing data, identify patterns, identify the relationships of various parameters of the outputs with the help of graphs, statistics etc.

Fig: Data Analysis

Count Analysis:

The below graph shows the Count Analysis over data present in the dataset.

Fig: Count Analysis

Tweet Length analysis of Normal Tweets:

The below graph shows the Tweet Length analysis of Normal Tweets over data present in the dataset.

Fig: Tweet Length analysis of Normal Tweets

Tweet Length analysis of Gender Tweets:

The below graph shows the Tweet Length analysis of Gender Tweets over data present in the dataset.

Fig: Tweet Length analysis of Gender Tweets

Tweet Length analysis of Religion Tweets:

The below graph shows the Tweet Length analysis of Religion Tweets over data present in the dataset.

Fig Tweet Length analysis of Religion Tweets

Tweet Length analysis of Electricity Tweets:

The below graph shows the Tweet Length analysis of Electricity Tweets over data present in the dataset.

Fig Tweet Length analysis of Electricity Tweets

Word Cloud Analysis of Normal Tweets:

The below graph shows the Word Cloud Analysis of Normal Tweets over data present in the dataset.

Fig Word Cloud Analysis of Normal Tweets

Word Cloud Analysis of Gender Tweets:

When the dataset is feed to various algorithms for Word Cloud Analysis of Gender Tweets

Fig: Word Cloud Analysis of Gender Tweets

Word Cloud Analysis of Religion Tweets:

When the dataset is feed to various algorithms for Word Cloud Analysis of Religion Tweets

Fig: Word Cloud Analysis of Religion Tweets

Word Cloud Analysis of Electricity Tweets:

When the dataset is feed to various algorithms for Word Cloud Analysis of Electricity Tweets

Fig: Word Cloud Analysis of Electricity Tweets

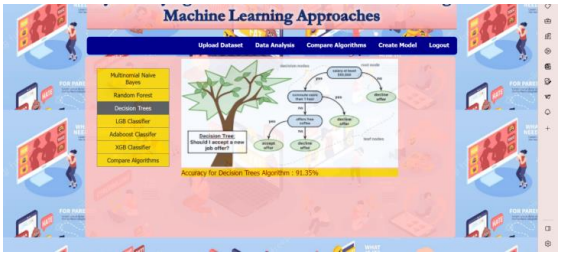


Fig: Decision Trees

Compare Algorithms:

On this page, the admin can feed the dataset to various Algorithms to train them and get the test accuracy for each algorithm. When the dataset is feed to various algorithms to evaluate the situation with some parameters like Accuracy, F1-Score , Recall…

Random Forest:

When the dataset is feed to Random Forest algorithm we observe that the test accuracy is 93.4%.

Fig: Random Forest

Decision Trees:

When the dataset is feed to Decision Trees algorithm we observe that the test accuracy is 91.35%.

LGB Classifier:

When the dataset is feed to Light Gradient Boosting algorithm we observe that the test accuracy is 93.13%.

Fig: LGB Classifier

AdaBoost Classifier:

When the dataset is feed to AdaBoost Classifier algorithm we observe that the test accuracy is 89.87%.

Fig: AdaBoost Classifier

XGB Classifier:

When the dataset is feed to XGB Classifier algorithm we observe that the test accuracy is 89.87%.

Fig: XGB Classifier

Compare Algorithm Summary:

On this page, the admin can feed the dataset to various Algorithms to train them, get the test accuracy for each algorithm and their accuracies are summarized here.

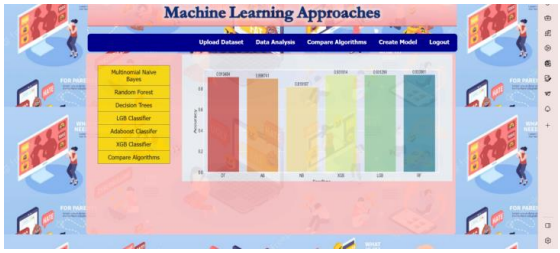


Fig: Compare Algorithm Summary

Create Model:

This screen shows the Training Accuracy of the Model is 100% and Test Accuracy of the Model is 93.4%.

Fig: Create Model

CONCLUSION

In particular, cyberbullying has become more common and has begun to raise significant social issues with the rising prevalence of social media sites and increased social media use by teenagers. There needs to design automatic cyberbullying detection method to avoid bad consequences of cyber harassment. Considering the significance of cyberbullying detection, in this study, we investigated the automated identification of posts on social media related to cyberbullying by considering two features BoW and TF-IDF. Four machine learning algorithms are used to identify bullying text and Random Forest for both BoW and TF-IDF.

We have developed an approach towards the detection of cyberbullying behaviour. If we are able to successfully detect such posts which are not suitable for adolescents or teenagers, we can very effectively deal with the crimes that are committed using these platforms. An approach is proposed for detecting and preventing Twitter cyberbullying using Supervised Binary Classification Machine Learning algorithms. Our model is evaluated on multiple classification algorithms and also for feature extraction, we used the TFIDF vector. As the results show us that the accuracy for detecting cyberbullying content has also been great for Random Forest classifier with 100% training accuracy and 93.4% testing accuracy . Our model will help people from the attacks of social media bullies.

REFERENCES

[1] Jason Brownlee, "How to use Word Embedding Layers for Deep Learning with Keras" in Deep Learning for Natural Language Processing.

[2] Justin W. Patchin, "Summary of Our Cyberbullying Research (2019)" , Cyberbullying Research Centre, July 10, 2019.

[3] Rui Zhao, Kezhi Mao, "CyberBullying Detection based on SemanticEnhance Marginalize Denoising Autoencoders" IEEE Transaction on Affective Computing.

[4] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Houng Wei, Haobo Xu "Attention-based Bi-directional Long Short Term Memory Network for Relation Classification" proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 207-212.

[5] MS. Snehal Bhoir, Tushar Ghorpade, Vanita Mane "Comparative Analysis of Different Word Embedding Models" IEEE.

[6] V. Banerjee, J. Telavane, P. Gaikwad and P. Vartak, "Detection of Cyberbullying Using Deep Neural Network", 2019 5th International Conference on Adavnced Computing & Communication System (ICACCS), Coimbatore, India.

[7] Agrawal S., Awekar A. (2018) "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms", In: Pasi G., Piwowarski B., Azzopardi L., Hanbury A. (eds) Advances in Information Retrieval. ECIR. Lecture Notes in Computer Science, vol 10772. Springer, cham.

[8] Brown, E. Clery and C. Ferguson, "Estimating the prevalence of young people absent from school due to bullying", National Center for Social Research.

[9] Monirah A., Al-Ajlan, Mourad Ykhlef, "Optimized Twitter Cyberbullying Detection based on Deep Learning", 978-1-5386-4110-1, IEEE.

[10] Vandana Nanda Kumar, Binsu C, Kovoor, Sreeja M.U., "CyberBullying Revelation in Twitter Data using Naïve-Bayes Classifier Algorithm" International Journal of Advanced Research in Computer Science. Volume 9, No.