iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45358

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Digital Tormenting Identification ON Web-based Entertainment Utilizing Machine Learning

Sowjanya G Savanth[1], Geetha M[2]
[1, 2]*Department of Master of Computer Application, BIET, Davangere*

Abstract: ***Cyber bullying is a major problem encountered on internet that affects teenagers and also adults. It has led to mis-happenings like suicide and depression. Regulation of content on Social media platforms has become a growing need. The following study uses data from two different forms of cyber bullying, hate speech tweets from Twittter and comments based on personal attacks from Wikipedia forums to build a model based on detection of cyber bullying 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. For Tweet data the model provides accuracies above 90% and for Wikipedia data it gives accuracies above 80%.***

Keywords: Wikipedia, Twitter, Natural Language Processing and Cyber bullying.

I. INTRODUCTION

Introduction With the rapid growth of social media, users especially adolescents are spending significant amount of time on various social networking sites to connect with others, to share information, and to pursue common interests. It has been found that 70% of teens use social media sites on a daily basis and nearly one in four teens hit their favorite social media sites 10 or more times a day. 19% of teens report that someone has written or posted mean or embarrassing things about them on social networking sites. As adolescents are more likely to be negatively affected by biased and harmful contents than adults, detecting online offensive contents to protect adolescent's online safety becomes an urgent task. To address concerns on children access to offensive contents over Internet, administrators of social media often manually review online contents to detect and delete offensive materials; however, manually reviewing are labour intensive and time consuming.

Some automatic contents filtering software packages, such as Appen, eblaster, iambigbrother, Internet Security Suite etc, have been developed to detect and filter online offensive contents. Most of them simply blocked web pages and paragraphs that contained dirty words. These word based approaches not only affect the readability and usability of web sites, but also fail to identify subtle offensive messages.

Sites for social networking are excellent tools for communication within individuals. Use of social networking has become widespread over the years, though, in general people find immoral and unethical ways of negative stuff. We see this happening between teens or sometimes between young adults. One of the negative stuffs they do is bullying each other over the internet. In online environment we cannot easily said that whether someone is saying something just for fun or there may be other intention of him. Often, with just a joke, "or don't take it so seriously," they'll laugh it off. Cyber bullying is the use of technology to harass, threaten, embarrass, or target another person. Often this internet fight results into real life threats for some individual. Some people have turned to suicide. It is necessary to stop such activities at the beginning. Any actions could be taken to avoid this for example if an individual's tweet/post is found offensive then maybe his/her account can be terminated or suspended for a particular period.

A. Proposed System

Cyber bullying detection is solved in this project as a binary classification problem where we are detecting two major form of Cyber bullying: hate speech on Twitter and Personal attacks on Wikipedia and classifying them as containing Cyber bullying or not. The proposed system uses: Support Vector Machine (SVM) for Twitter Hate Speech and Random Forest Classifier for Personal attacks.

SVM is basically used to plot a hyper plane that creates a boundary between data points in number of features (N)-dimensional space.

To optimize the margin value hinge function is one of best loss function for this. Linear SVM is used in the following case which is optimum for linearly separable data. In case of 0 misclassification, i.e. the class of data point is accurately predicted by our model, we only have to change the gradient from the regularization arguments.

A random forest consists of many individual decision trees which individually predict a class forgiven query points and the class with maximum votes is the final result. Decision Tree is a building block for random forest which provides a prediction by decision rules learned from feature vectors. An ensemble of these uncorrelated trees provides a more accurate decision for classification or regression.

II. LITERATURE SURVEY

In recent years, users are widely intended to express and share their opinions over the Internet. However, due to the characters of social media, it appears negative use of social media. Cyberbullying is one of the abuse behaviors in the Internet as well as a very serious social problem.

Under this background and motivation, it can help to prevent the happen of cyberbullying if we can develop relevant techniques to discover cyberbullying in social media. Thus, in this paper we propose an approach based on social networks analysis and data mining for cyberbullying detection. In the approach, there are three main techniques for cyberbullying discovery will be studied, including keyword matching technique, opinion mining and social network analysis. In addition to the approach, we will also discuss the experimental design for the evaluation of the performance.

The use of new technologies along with the popularity of social networks has given the power of anonymity to the users. The ability to create an alter-ego with no relation to the actual user, creates a situation in which no one can certify the match between a profile and a real person.

This problem generates situations, repeated daily, in which users with fake accounts, or at least not related to their real identity, publish news, reviews or multimedia material trying to discredit or attack other people who may or may not be aware of the attack. These acts can have great impact on the affected victims' environment generating situations in which virtual attacks escalate into fatal consequences in real life. In this paper, we present a methodology to detect and associate fake profiles on Twitter social network which are employed for defamatory activities to a real profile within the same network by analyzing the content of comments generated by both profiles.

Accompanying this approach we also present a successful real life use case in which this methodology was applied to detect and stop a cyberbullying situation in a real elementary school.

As the size of Twitter© data is increasing, so are undesirable behaviors of its users. One of such undesirable behavior is cyberbullying, which may even lead to catastrophic consequences. Hence, it is critical to efficiently detect cyberbullying behavior by analyzing tweets, if possible in real-time. Prevalent approaches to identify cyberbullying are mainly stand-alone and thus, are time-consuming.

This research improves detection task using the principles of collaborative computing. Different collaborative paradigms are suggested and discussed in this paper. Preliminary results indicate an improvement in time and accuracy of the detection mechanism over the stand-alone paradigm.

With the increasing use of social media, cyberbullying behavior has received more and more attention. Cyberbullying may cause many serious and negative impacts on a person's life and even lead to teen suicide.

To reduce and stop cyberbullying, one effective solution is to automatically detect bullying content based on appropriate machine learning and natural language processing techniques. However, many existing approaches in the literature are just normal text classification models without considering bullying characteristics. In this paper, we propose a representation learning framework specific to cyberbullying detection. Based on word embedding's, we expand a list of pre-defined insulting words and assign different weights to obtain bullying features, which are then concatenated with Bag-of-Words and latent semantic features to form the final representation before feeding them into a linear SVM classifier. Experimental study on a twitter dataset is conducted, and our method is compared with several baseline text representation learning models and cyberbullying detection methods. The superior performance achieved by our method has been observed in this study.

Innovation is developing quickly today. This headway in innovation has changed how individuals cooperate in an expansive way giving communication another dimension.

But despite the fact that innovation encourages us in numerous parts of life, it accompanies different effects that influence people in a few or the other way. Cyberbullying is one of such effects. Cyberbullying is a wrongdoing in which a culprit focuses on an individual with online provocation and loathes which has antagonistic emotional, social and physical effects on the victim. So as to address such issue we proposed a novel cyberbullying detection method dependent on deep neural network. Convolution Neural Network is utilized for the better outcomes when contrasted with the current systems.

.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

III. SYSTEM DESIGN

A. System Architecture

We used SVM for twitter data and Random forest classifier for Wikipedia data for identifying cyber bullying. SVM offers very high accuracy compared to other classifiers such as logistic regression, and decision trees. It is known for its kernel trick to handle nonlinear input spaces. It is used in a variety of applications such as face detection, intrusion detection, classification of emails, news articles and web pages, classification of genes, and handwriting recognition.

SVM is an exciting algorithm and the concepts are relatively simple. The classifier separates data points using a hyperplane with the largest amount of margin. That's why an SVM classifier is also known as a discriminative classifier. SVM finds an optimal hyperplane which helps in classifying new data points. Random forests are a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

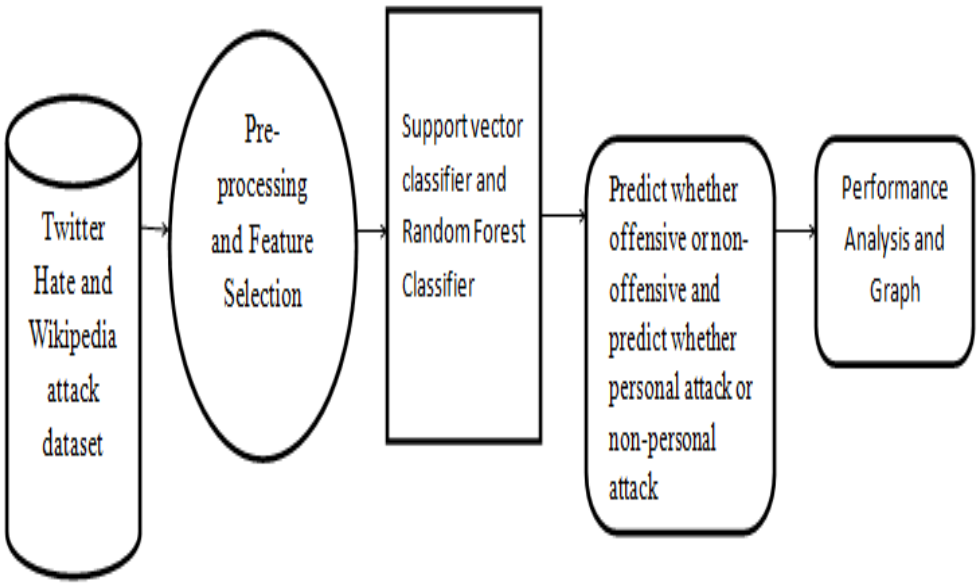


Fig. 1 Architecture of our proposed System

B. Dataflow Diagram

Figure 2 depicts DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.

Training phase: Dataset is pre-processed by removing null values and error values. The training phase can be summarized as follows:

1) Extract features from the pre-processed dataset
2) Train a SVM for twitter datatset and Random forest classifier for Wikipedia dataset using this feature set.
The output of the training phase is a trained classifier capable of predicting classification personal attacks and offensive comments.

The performance of the trained classifier can be evaluated using measures like accuracy, sensitivity and specificity.

Classification: This phase can be summarized as follows:

a) Take as comment as input.
b) Pre-process the uploaded comment
c) Extract the required features from it.
d) Use the trained classifier to predict the comment.
The output of this phase is a classification label.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

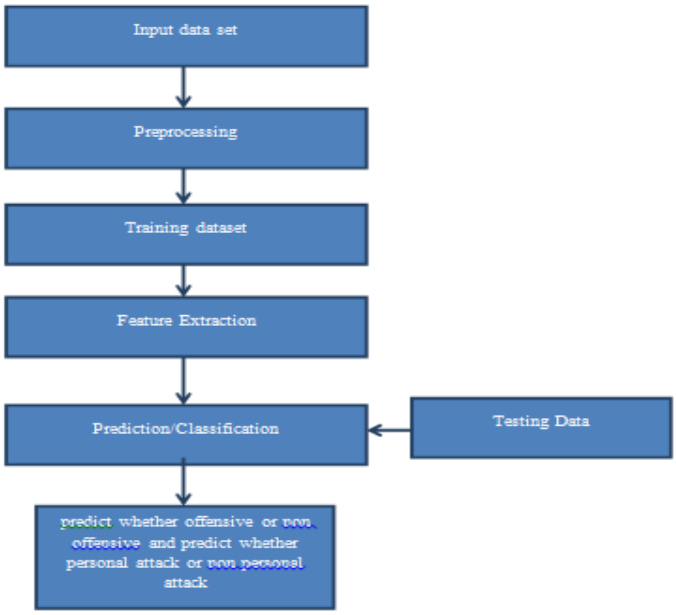


Fig. 2 Dataflow Diagram

C. Activity Diagram

Figure 3 represents Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modelling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

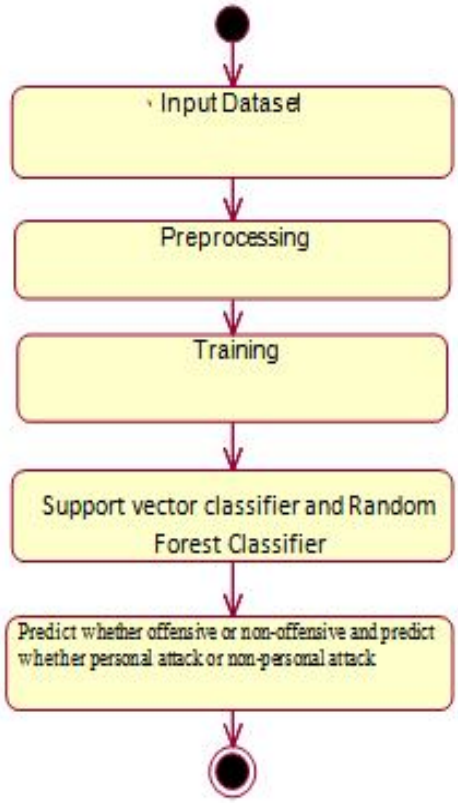


Fig. 3 Activity Diagram

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

IV. IMPLEMENTATION

A. *Modules*

1) Data Collection
2) Dataset
3) Data Preparation
4) Model Selection
5) Analyse and Prediction
6) Accuracy on test set
7) Saving the Trained Model

B. *Twitter Hate Detection*

1) *Data Collection:* This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.
There are several techniques to collect the data, like web scraping, manual interventions and etc.
Detection of Cyberbullying on Social Media Using Machine learning
Dataset: The dataset consists of 31962 individual data. There are 3 columns in the dataset, which are described below

a) Id: unique id
b) Labels :
1: offensive
0: non offensive
c) Tweet : comment

2) *Data Preparation:* We will transform the data. by getting rid of missing data and removing some columns. First we will create a list of column names that we want to keep or retain. Next we drop or remove all columns except for the columns that we want to retain.
Finally we drop or remove the rows that have missing values from the data set. Steps to follow:

a) Removing extra symbols
b) Removing punctuations
c) Removing the Stopwords
d) Stemming
e) Tokenization
f) Feature extractions
g) TF-IDF vectorizer
h) Counter vectorizer with TF-IDF transformer

3) *Model Selection:* We used SVC algorithms

C. *Support Vector Machines*

Generally, Support Vector Machines is considered to be a classification approach, it but can be employed in both types of classification and regression problems. It can easily handle multiple continuous and categorical variables. SVM constructs a hyperplane in multidimensional space to separate different classes. SVM generates optimal hyperplane in an iterative manner, which is used to minimize an error. The core idea of SVM is to find a maximum marginal hyperplane(MMH) that best divides the dataset into classes.

1) *Accuracy on test set:* We got an accuracy of 96.02% on test set.
2) *Saving the Trained Model:* Once you're confident enough to take your trained and tested model into the production-ready environment, the first step is to save it into a .h5 or . pkl file using a library like pickle .
Make sure you have pickle installed in your environment. Next, let's import the module and dump the model into . pkl file

D. Wikipedia Attack

1) *Data Collection:* This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.
There are several techniques to collect the data, like web scraping, manual interventions and etc.
Detection of Cyberbullying on Social Media Using Machine learning
2) *Dataset:* The dataset consists of 115864 individual data. There are 4 columns in the dataset, which are described below

a) Review Id: unique id
b) comment : comment about wikipedia titles
c) year : year of comment
d) attack : Personal attack or non personal attack

3) *Data Preparation:* We will transform the data. by getting rid of missing data and removing some columns. First we will create a list of column names that we want to keep or retain.
Next we drop or remove all columns except for the columns that we want to retain.
Finally we drop or remove the rows that have missing values from the data set. Steps to follow:

a) Removing extra symbols
b) Removing punctuations
c) Removing the Stopwords
d) Stemming
e) Tokenization
f) Feature extractions
g) TF-IDF vectorize
h) Counter vectorizer with TF-IDF transformer

4) *Model Selection:* We used RandomForestClassifier algorithms

Let's understand the algorithm in layman's terms. Suppose you want to go on a trip and you would like to travel to a place which you will enjoy. So what do you do to find a place that you will like? You can search online, read reviews on travel blogs and portals, or you can also ask your friends.

Let's suppose you have decided to ask your friends, and talked with them about their past travel experience to various places. You will get some recommendations from every friend. Now you have to make a list of those recommended places. Then, you ask them to vote (or select one best place for the trip) from the list of recommended places you made. The place with the highest number of votes will be your final choice for the trip.

In the above decision process, there are two parts. First, asking your friends about their individual travel experience and getting one recommendation out of multiple places they have visited. This part is like using the decision tree algorithm. Here, each friend makes a selection of the places he or she has visited so far.

The second part, after collecting all the recommendations, is the voting procedure for selecting the best place in the list of recommendations. This whole process of getting recommendations from friends and voting on them to find the best place is known as the random forests algorithm.

It technically is an ensemble method (based on the divide-and-conquer approach) of decision trees generated on a randomly split dataset. This collection of decision tree classifiers is also known as the forest. The individual decision trees are generated using an attribute selection indicator such as information gain, gain ratio, and Gini index for each attribute. Each tree depends on an independent random sample. In a classification problem, each tree votes and the most popular class is chosen as the final result. In the case of regression, the average of all the tree outputs is considered as the final result. It is simpler and more powerful compared to the other non-linear classification algorithms.

How does the algorithm work?

It works in four steps:

a) Select random samples from a given dataset.
b) Construct a decision tree for each sample and get a prediction result from each decision tree.
c) Perform a vote for each predicted result.
d) Select the prediction result with the most votes as the final prediction.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

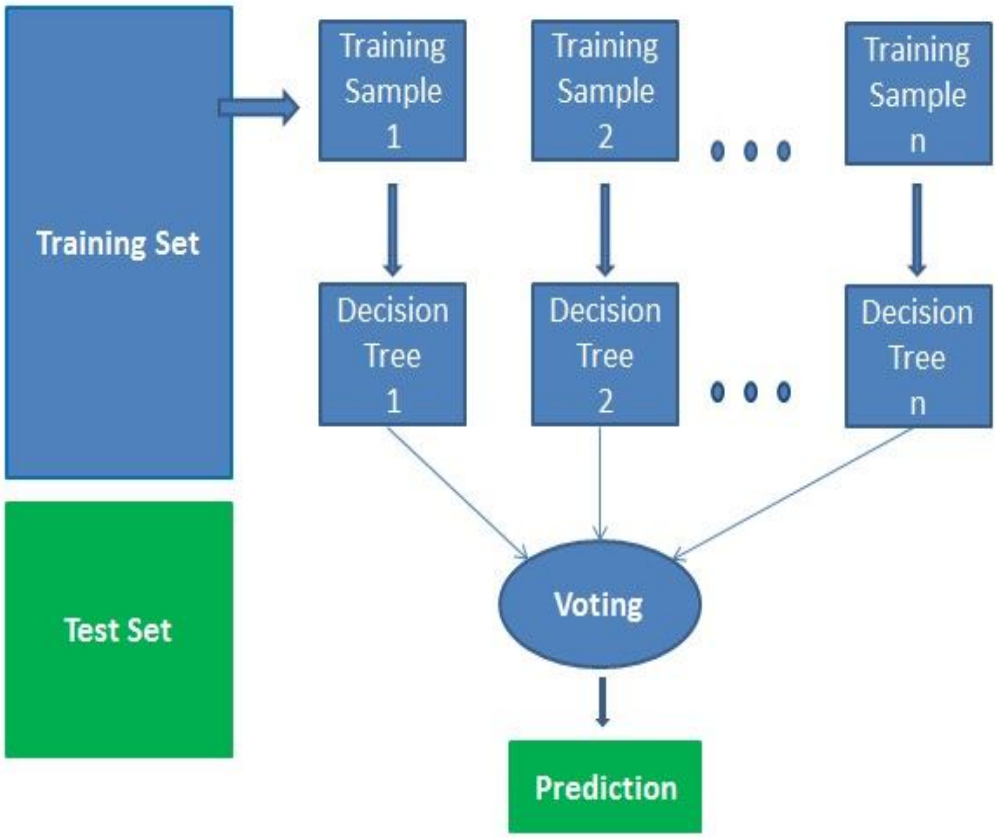


Fig. 4 Randomforest classifier working Diagram

The model is difficult to interpret compared to a decision tree, where you can easily make a decision by following the path in the tree.

E. Finding Important Features

Random forests also offers a good feature selection indicator. Scikit-learn provides an extra variable with the model, which shows the relative importance or contribution of each feature in the prediction. It automatically computes the relevance score of each feature in the training phase. Then it scales the relevance down so that the sum of all scores is 1.

This score will help you choose the most important features and drop the least important ones for model building.

Random forest uses gini importance or mean decrease in impurity (MDI) to calculate the importance of each feature. Gini importance is also known as the total decrease in node impurity. This is how much the model fit or accuracy decreases when you drop a variable. The larger the decrease, the more significant the variable is. Here, the mean decrease is a significant parameter for variable selection. The Gini index can describe the overall explanatory power of the variables.

F. Analyze and Prediction

In the actual dataset, we chose only 2 features :

Text: the tweets

Labels :

1: personal attack

0: non personal attack

1) *Accuracy on test set:* We got a accuracy of 99.02% on test set.
2) *Saving the Trained Model:* Once you're confident enough to take your trained and tested model into the production-ready environment, the first step is to save it into a .h5 or . pkl file using a library like pickle .

   Make sure you have pickle installed in your environment. Next, let's import the module and dump the model into . pkl file

V. RESULTS AND SCREENSHOTS

The figure 5 depicts the output screen of the proposed system. It is the home page in which we get login button. We use django framework for fron-end. Once we start the django server we get this page in our local host port.

Fig. 5 Home Page

Home page got two separate buttons for logging twitter dataset and Wikipedia dataset cyberbullying identification.

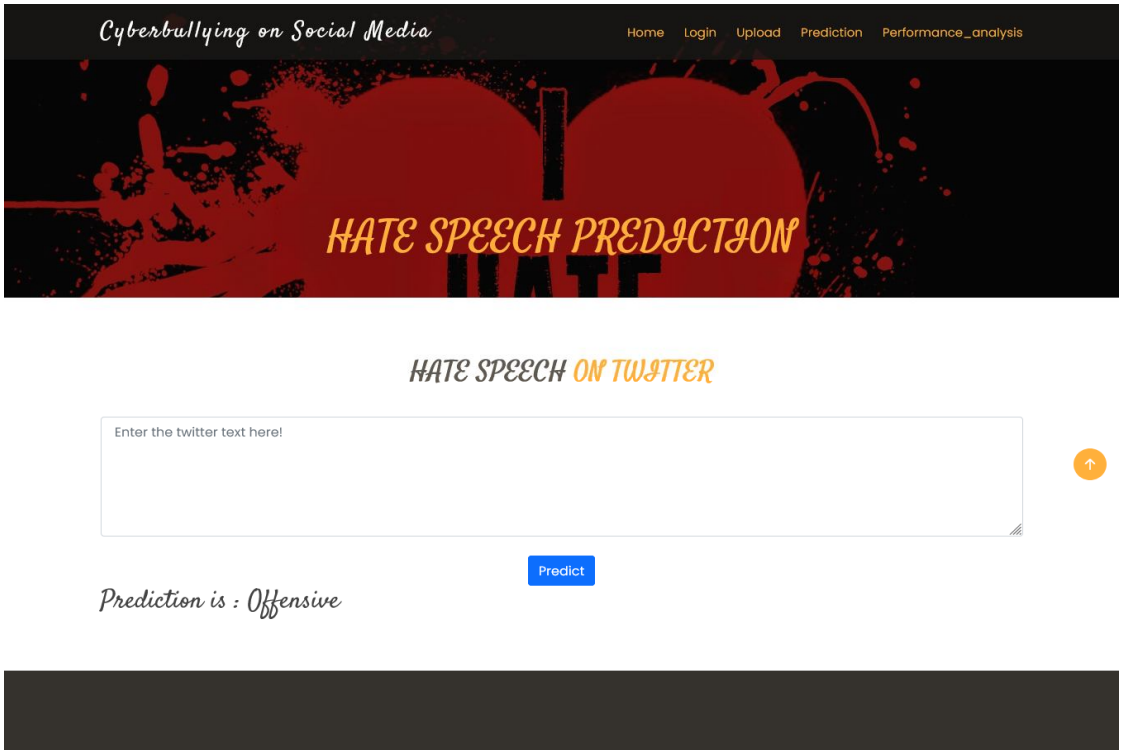


Fig. 6 Predicted result of the input comment.

Figure 6 shows the predicted result for the twitter comment. Its shows the comment is offensive.

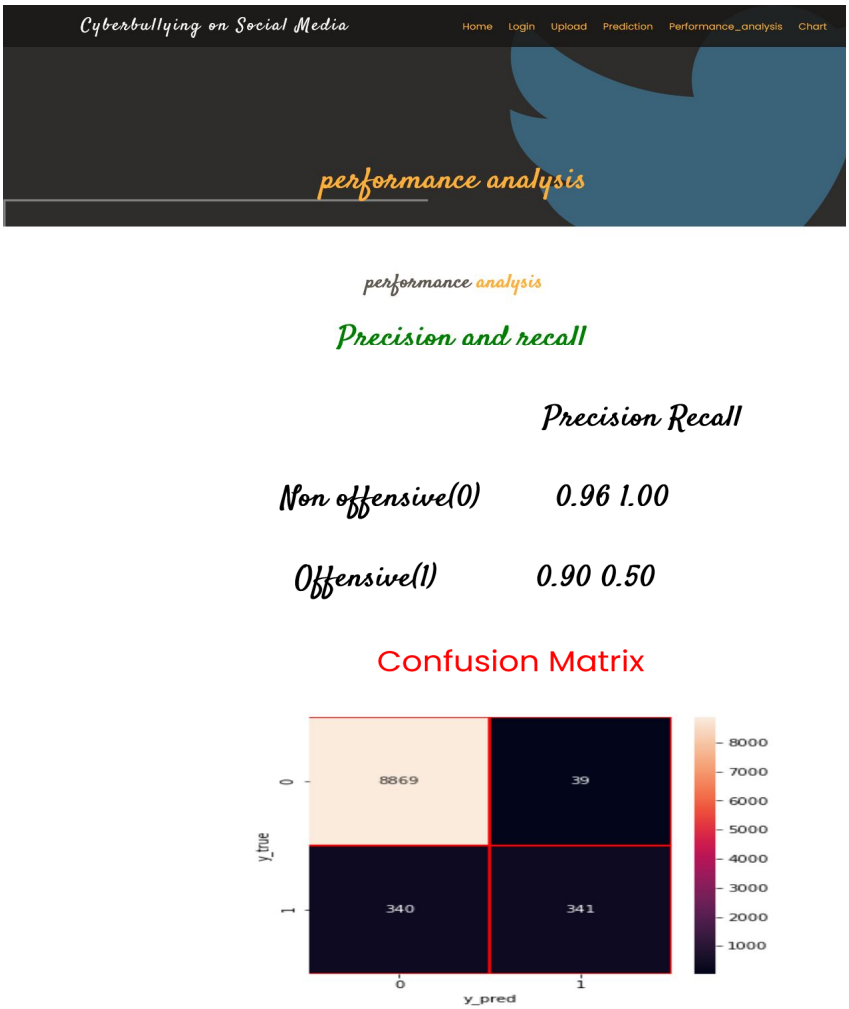
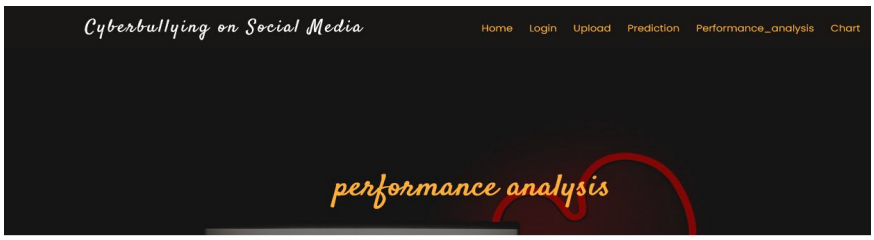


Fig. 7 Performance analysis of SVM for twitter data cyberbullying attack identification

Figure 7 portrays the perfromance of our created cycberbullying discovery for twitter information. Our model has exactness of 96% with SVM calculation. Likewise it shows Derived Confusion Matrix of the System for SVM

Fig. 8 Predicted outcome of the input wikipedia comment



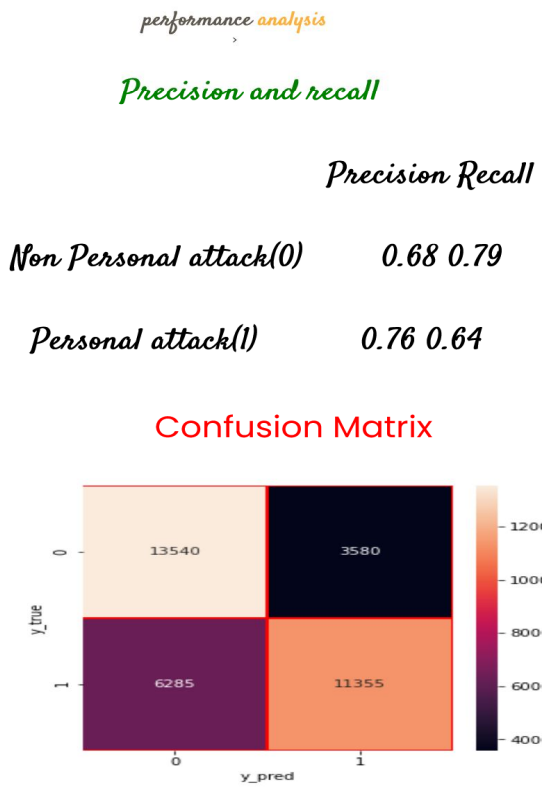


Fig. 9 Performance analysis of Random forest for Wikipedia data cyberbullying attack identification

Figure 9 depicts the perfromance of our developed cycberbullying detection for Wikipedia data. Our model has accuracy of 76% with Random forest algorithm. Also it shows Derived Confusion Matrix of the System for RF algorithm.

VI. CONCLUSION

Cyber bullying across internet is dangerous and leads to mis-happenings like suicides, depression etc and therefore there is a need to control its spread.

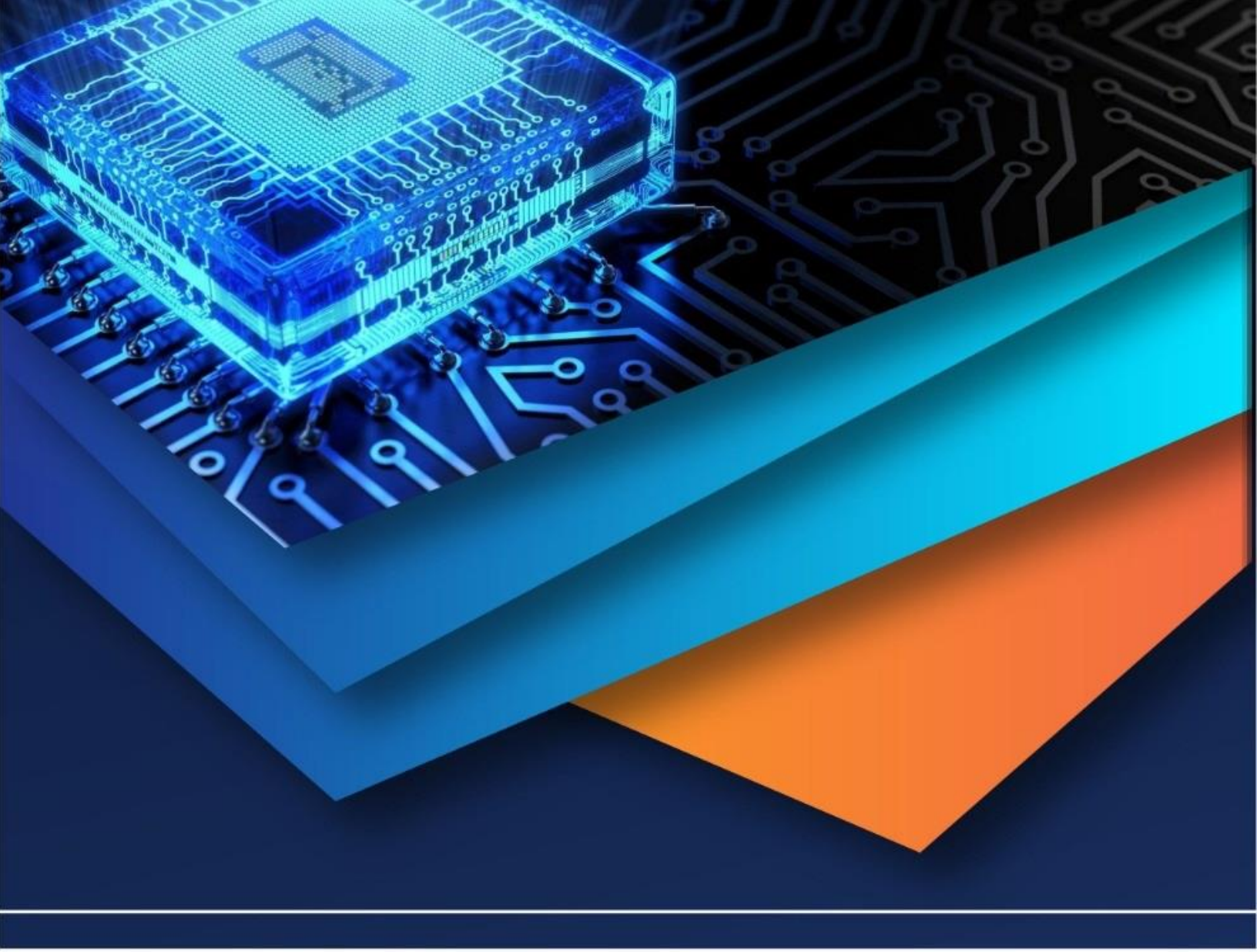
Therefore cyber bullying detection is vital on social media platforms. With viability of more data and better classified user information for various other forms of cyber-attacks Cyberbullying detection can be used on social media websites to ban users trying to take part in such activity In this paper we proposed an architecture for detection of cyber bullying to combat the situation.

We discussed the architecture for two types of data: Hate speech Data on Twitter and Personal attacks on Wikipedia. For Hate speech Natural Language Processing techniques proved effective with accuracies of over 90 precent using basic Machine learning algorithms because tweets containing Hate speech consisted of profanity which made it easily detectable. Due to this it gives better results with Bow and Tf-Idf models rather than Word2Vec models However, Personal attacks were difficult to detect through the same model because the comments generally did not use any common sentiment that could be learned however the three feature selection methods performed similarly.

Word2Vec models that use context of features proved effective in both datasets giving similar results in comparatively less features when combined with Multi Layered Perceptron..

REFERENCES

[1] H. Ting, W. S. Liou, D. Liberona, S. L. Wang, and G. M. T. Bermudez, "Towards the detection of cyberbullying based on social network mining techniques," in Proceedings of 4th International Conference on Behavioral, Economic, and Socio Cultural Computing, BESC 2017, 2017, vol. 2018-January, doi: 10.1109/BESC.2017.8256403.
[2] P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying," 2014, doi: 10.1007/978-3-319-01854-6_43.
[3] A. Mangaonkar, A. Hayrapetian, and R. Raje, "Collaborative detection of cyberbullying behavior in Twitter data," 2015, doi: 10.1109/EIT.2015.7293405.
[4] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," 2016, doi: 10.1145/2833312.2849567.
[5] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, "Detection of Cyberbullying Using Deep Neural Network," 2019, doi: 10.1109/ICACCS.2019.8728378.
[6] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," 2011, doi: 10.1109/ICMLA.2011.152.
[7] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying Detection using Pre-Trained BERT Model," 2020, doi: 10.1109/ICESC48915.2020.9155700.
[8] M. Dadvar and K. Eckert, "Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study," arXiv. 2018.
[9] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," arXiv. 2018.
[10] 10. Y. N. Silva, C. Rich, and D. Hall, "BullyBlocker: Towards the identification of cyberbullying in social networking sites," 2016, doi: 10.1109/ASONAM.2016.7752420.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY