International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217381

Cyber Bullying Detection on Social Media using Machine Learning

K. Mahesh, Suwarna Gothane, Aashish Toshniwal, Vinay Nagarale, Harish Gopu
Computer Science and Engineering, CMR Technical Campus (CMRTC), Hyderabad, India

Article Info
Volume 7, Issue 3
Page Number: 410-416

Publication Issue :
May-June-2021

Article History
Accepted : 25 May 2021
Published : 31 May 2021

ABSTRACT

From the day internet came into existence, the era of social networking sprouted. In the beginning, no one may have thought internet would be a host of numerous amazing services like the social networking. Today we can say that online applications and social networking websites have become a non-separable part of one's life. Many people from diverse age groups spend hours daily on such websites. Despite the fact that people are emotionally connected together through social media, these facilities bring along big threats with them such as cyber-attacks, which includes cyberbullying.

Keywords: Cyberbullying, social media, Support Vector Machine, Naïve Bayes, Test-Train Split, Classification, Detection

I. INTRODUCTION

Social networking sites are being widely used today for multiple purposes like entertainment, networking, etc. Social networking sites are a stop for multiple reasons to billions of people today. All the social media platforms require the consent of all the participating people. Communicating with people is no exception, as technology has changed the way people interact with a broader manner and has given a new dimension to communication. Many people are illegally using these communities. Many youngsters are getting bullied these days. Bullies use various services like Twitter, Facebook, and Email to bully people.

Cyberbullying is one of the most frequently happen Internet abuse and also a very serious social problem especially for teenager. Therefore, more and more researchers are devoting on how to discover and prevent the happen of cyberbullying, especially in social media. Cyberbullying is not just limited to creating a fake identity and publishing/posting some embarrassing photo or video, unpleasant rumours about someone but also giving them threats. The impacts of cyberbullying on social media are horrifying, sometimes leading to the death of some unfortunate victims.

Thus, a complete solution is required for this problem. Cyberbullying needs to stop. The problem can be tackled by detecting and preventing it by using a machine learning approach, this needs to be done using a different perspective.

Cyberbullying is a relatively new medium through which bullying occurs (e.g., chat rooms, text messages). Cyberbullying has been defined as an individual or a group wilfully using information and communication involving electronic technologies to facilitate deliberate and repeated harassment or threat to another individual or group by sending or posting cruel text and/or graphics using technological means. Many of the methods used in traditional bullying are used in cyberbullying. Direct cyberbullying can occur when one person calls another a name through an electronic message. Relational bullying can also occur online. For example, with the numerous social networking sites now online (e.g., Facebook, Myspace), 'hate groups' have become a popular approach to bullying. In a hate group, a student creates an online social group against a schoolmate, allows others to join, and collectively the group posts negative comments about the student. Fortunately, social networking sites have begun taking action against the creation of hate groups. When creating a group on Facebook, for instance, a warning is placed near the bottom of the page that reads, "Note: groups that attack a specific person or group of people (e.g., racist, sexist, or other hate groups) will not be tolerated.

II. RELATED WORK

The results of the proposed model demonstrate significant improvement in the performance of classification on all the datasets in comparison to recent existing models. The success rate of the SVM classifier with the excellent recall is 0.971 via tenfold cross-validation, which demonstrates the high efficiency and effectiveness of the proposed model. [2] Author of the Work in references Detecting Offensive Language in Social media is *Ying chen, Yilu Zhou, Sencun Zhu* and *Heng Xu* who came up with a methodology of user-level offensiveness detection seems a more feasible approach. so, the Lexical Syntactic Feature (LSF) architecture to detect offensive content and identify potential offensive users in social media. We distinguish the contribution of pejoratives/profanities and obscenities in determining offensive content, and introduce hand-authoring syntactic rules in identifying name-calling harassments. [3] Another Author *K. Jedrzejewski* and *M. Morzy* had a different methodology where The role and importance of social networks in preferred environments for opinion mining and sentiment analysis. Selected properties of social networks that are relevant with respect to opinion mining are described and general relationships between the two disciplines are outlined. The related work and basic definitions used in opinion mining is given. [4] *H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra* , Cyberbullying is a growing problem affecting more than half teens. The main goal is to study cyberbullying incidents in the social network. In this work, we have collected a sample data and their associated comments. We then designed a study and employed human contributors at the crowd-sourced Crowd Flower website to label these media sessions for cyberbullying. [5] *Kelly Reynolds, April Kontostathis, Lynne Edwards* The results of the proposed model demonstrate significant improvement in the performance of classification on all the datasets in comparison to recent existing models. The success rate of the SVM classifier with the excellent recall is 0.971 via tenfold cross-validation, which demonstrates the high efficiency and effectiveness of the proposed model.

III. PROPOSED WORK

The proposed model is introduced to overcome all the disadvantages that arises in the existing system. This system will increase the accuracy of the supervised classification results by classifying the data. An approach is proposed for detecting and preventing

Twitter cyberbullying using Supervised Binary Classification Machine Learning algorithms. Our model is evaluated on both Support Vector Machine and Naive Bayes. It enhances the performance of the overall classification results. This proposed method is supposed to have high performance while providing accurate prediction results. It also avoids sparsity problems. It is less prone to information loss. Below figure depicts the architecture of the proposed system.

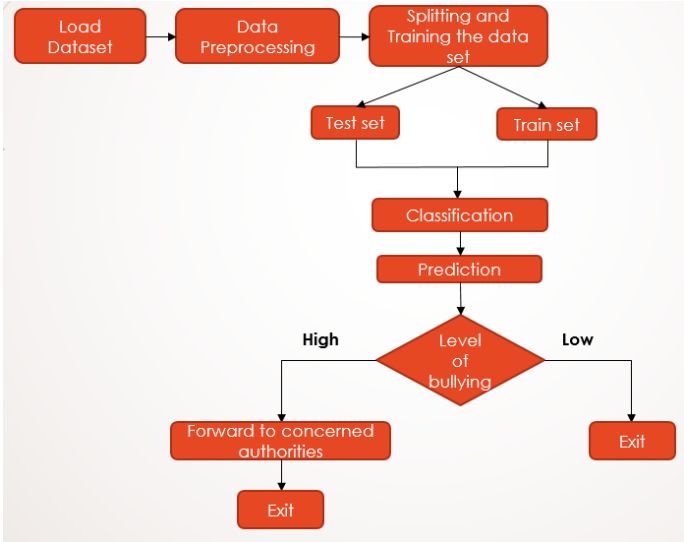

Fig 1. Proposed System Architecture

This is the project architecture where the dataset is loaded and pre-processing is done where the unwanted data is removed and then the data is split trained into test and train sets which is then classified using algorithms and then a prediction is obtained which shows us the level of bullying as high or low and then the execution of the program exits for low severity and forwards to concerned authorities and exits if High severity is present.

IV. IMPLEMENTSTION OF PROPOSED SYSTEM

The proposed system consists of six modules. The data selection is the process of selecting the data for detecting the attacks. In this project, the cyberbullying tweets dataset is used for detecting offensive and non-offensive tweets. The dataset which contains the information about the user name and tweets label.

Data pre-processing is the process of removing the unwanted data from the dataset. Missing data removal, Encoding Categorical data.

Missing data removal: In this process, the null values such as missing values are removed using imputer library.

Encoding Categorical data: That categorical data is defined as variables with a finite set of label values. That most machine learning algorithms require numerical input and output variables. That an integer and one hot encoding is used to convert categorical data to integer data.

Data splitting is the act of partitioning available data into two portions, usually for cross-validator purposes. One Portion of the data is used to develop a predictive model and the other to evaluate the model's performance. Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing.

The Supervised classification algorithm such as Naïve Bayes and Support vector machine is used in Data Mining.

Support vector machine:

For implementing this we have used Support vector machine (SVM) model is basically a representation of different classes in a hyper plane in multidimensional space. The hyper plane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyper plane.

Naive Bayes classifier:

it is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Predictive analytics algorithms try to achieve the lowest error possible by either using "boosting" or "bagging".

Accuracy – Accuracy of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

Speed – Refers to the computational cost in generating and using the classifier or predictor.

Robustness – It refers to the ability of classifier or predictor to make correct predictions from given noisy data.

Scalability – Scalability refers to the ability to construct the classifier or predictor efficiently; given large amount of data.

Interpretability – It refers to what extent the classifier or predictor understands. It's a process of predicting the offensive and non-offensive tweets from the dataset. This project will effectively predict the data from dataset by enhancing the performance of the overall prediction results.

V. RESULT GENERATION

The Final Result will get generated based on the overall classification and prediction. The performance of this proposed approach is evaluated using some measures like,

Accuracy of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

$$AC= (TP+TN)/(TP+TN+FP+FN)$$

Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives.

$$Precision=TP/(TP+FP)$$

Recall is the number of correct results divided by the number of results that should have been returned. In binary classification, recall is called sensitivity. It can be viewed as the probability that a relevant document is retrieved by the query.

$$Recall=TP/(TP+FN)$$

F measure (F1 score or F score) is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test.

$$F\text{-}measure=2TP/(2TP+FP+FN)$$

VI. OUTPUT AND SCREENSHOTS

Below are the screenshots of the output obtained from the proposed project.

Fig 2. Cyberbullying Detection Output

Here we can see how the variable explorer indicates and shows the data of the bad words out of the whole data set from the labels 1.

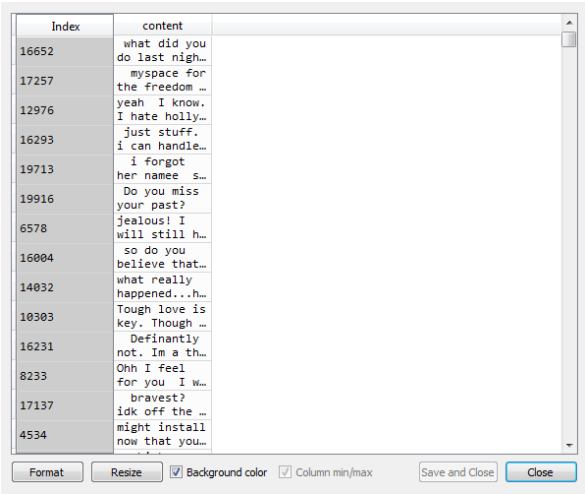

Fig 3. Sample Dataset

A sample of the data from the dataset with index in this way is generated after classifying and training the data set to separate the unwanted data and classify the important data to detect cyberbullying in the twitter database which is labelled with 1 or 0 where 1 means Positive and 0 means negative.

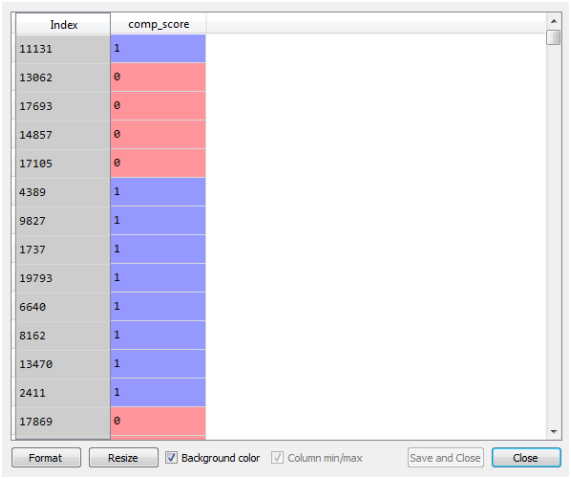

Fig 4. Labelled Values of the Dataset

As we can see above the data indexes with 0 as comp_score are not the users bullying and the indexes with 1 as value are the ones bullying. 0 and 1 indicate mainly if the data value is having abusive content which is tagged by labelling using sentiment analysis.

Number of rows in the total set: 20001

Number of rows in the training set: 15000

Number of rows in the test set: 5001

```
----------------------
Naive Bayes

------Classification Report------
              precision    recall  f1-score   support

           0       0.79      0.85      0.82      2350
           1       0.85      0.80      0.83      2651

    accuracy                           0.82      5001
   macro avg       0.82      0.82      0.82      5001
weighted avg       0.83      0.82      0.82      5001

------Accuracy------
The Accuracy Score :82


----------------------
Support vector Machine

------Classification Report------
              precision    recall  f1-score   support

           0       0.89      0.88      0.89      2538
           1       0.88      0.89      0.88      2463

    accuracy                           0.89      5001
   macro avg       0.89      0.89      0.89      5001
weighted avg       0.89      0.89      0.89      5001


------Accuracy------
The Accuracy Score :89
```

Fig 5. Accuracy of both the algorithms

As we can see now SVM has more accuracy than Naïve Bayes. Support vector machine has an accuracy of 89 whereas Naïve Bayes has an accuracy of 82.

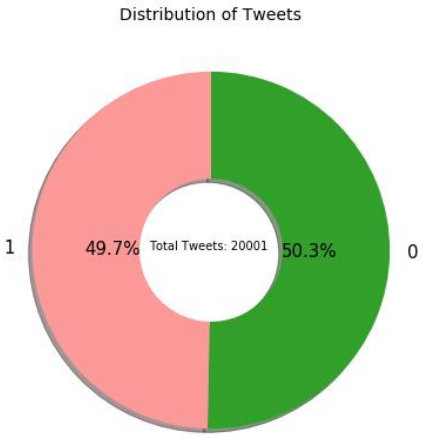

Fig 6. Distribution of Tweets

VII. FUTURE WORK

In future, it is possible to provide extensions or modifications to the proposed clustering and classification algorithms to achieve further increased performance. Apart from the experimented combination of data mining techniques, further combinations and other clustering algorithms can be used to improve the detection accuracy and to reduce the rate offensive tweets. Finally, the cyberbullying detection system can be extended as a prevention system to enhance the performance of the system.

VIII. CONCLUSION

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 both Support Vector Machine and Naive Bayes, 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 Support Vector Machine which is better than Naive Bayes. Our model will help people from the attacks of social media bullies.

IX. ACKNOWLEDGEMENT

We take this opportunity to express our gratitude and respect to all the faculty members who have guided us in this project. We take privilege to extend our profound gratitude and sincere thanks to our guide **Mr. K. Mahesh** and our PRC co-ordinator **Dr. Suwarna Gothane**, Department of computer science and Engineering, CMR Technical Campus, who constantly supported us at every stage of the project and helped us in our difficult times to make the project a success.

We are thankful to HOD Dr. K Srujan Raju. Department of Computer Science and Engineering, CMR Technical Campus, for his immense support and encouragement. We also take this opportunity to thank Dr A Raji Reddy, Director CMR Technical Campus, for providing us an encouraging environment to work with.

X. REFERENCES

[1]. Amanpreet Singh, Maninder Kaur, "Content-based Cybercrime Detection: A Concise Review", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-8, pages 1193-1207, 2019.

[2]. Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. "Detecting offensive language in social media to protect adolescent online safety". In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom), pages 71– 80. IEEE, 2012.

[3]. K. Jedrzejewski and M. Morzy, “Opinion Mining and Social Networks: A Promising Match,” 2011 Int. Conf. Adv. Soc. Networks Anal. Min., pp. 599–604, Jul. 2011.

[4]. H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, “Analysing Labelled Cyberbullying Incidents on the Instagram Social Network.”

[5]. Kelly Reynolds, April Kontostathis, Lynne Edwards, "Using Machine Learning to Detect Cyberbullying”, 2011 10th International Conference on Machine Learning and Applications volume 2, pages 241–244. IEEE, 2011.

Cite this article as :

K. Mahesh, Suwarna Gothane, Aashish Toshniwal, Vinay Nagarale, Harish Gopu, "Cyber Bullying Detection on Social Media using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 3, pp. 410-416, May-June 2021. Available at doi : https://doi.org/10.32628/CSEIT217381
Journal URL : https://ijsrcseit.com/CSEIT217381