International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Reducing Side Effects of Cyber Bullying Using Machine Learning

Mr. Mohammed Afzal,
Computer Science & Engineering(AIML)
(Assistant Professor)
Sphoorthy Engineering College
(JNTUH)
mdafzal.aiml@gmail.com

Dr. Subba Rao Kolavennu ,
Computer Science & Engineering
(JNTUH)
Sphoorthy Engineering College
(JNTUH)
profrao99@gmail.com

I. Venu Gopal ,
Computer Science & Engineerng
(B.Tech, JNTUH)
Sphoorthy Engineering College
(JNTUH)
venugopalinampudi1408@gmail.com

S. Hari Haran,
Computer Science & Engineering
(B.Tech, JNTUH)
Sphoorthy Engineering College
(JNTUH)
singamhariharan18@gmail.com

J. Shyam Sundar Lal,
Computer Science & engineering
(B.Tech, JNTUH)
Sphoorthy Engineering College
(JNTUH)
Jaruplashyamsundarlal@gmail.com

A. Pavan Kalyan
Computer Science & engineering
(B.Tech, JNTUH)
Sphoorthy Engineering College
(JNTUH)
Kalyankrishna157@gmail.com

Abstract

Public shaming on online social networks and associated online public forums such as Twitter has increased. These occurrences have a disastrous effect on the victim's social, political, and economic well-being. Despite the obvious negative consequences, nothing has been done to address this in popular online social media, with the justification that the vast volume and diversity of such remarks necessitates an infeasible number of human moderators to complete the work. We automate detecting public shaming via Twitter from victims' perspective in this research, focusing on two aspects: incidents and shamers. Abusive, comparison, passing judgment, religious/ethnic, sarcasm/joke and whataboutery are the six sorts of shameful tweets, and each tweet is categorized into one of these categories as non-shaming. It has been shown that most people who submit comments in a shaming event are likely to humiliate the victim. Surprisingly, shamers' Twitter follower counts grow quicker than those of non shamers.

Keywords - Cyberbullying, Social Media, NLP, Supervised learning, Twitter API.

1. Introduction

Millions of young people spend their time on social networking, and the sharing of information is online. Social networks have the ability to communicate and to share information with anyone, at any time, and in the number of people at the same time. There are over 3 billion social media users around the world. According to the National Crime Security Council (NCPC), cyberbullying is available online where mobile phones, video game apps, or any other way to send or send text, photos, or videos deliberately injure or embarrass another person. Cyberbullying can happen at any time all day, week and you can reach anyone anywhere via the internet. Text, photos, or videos of cyberbullying may be posted in an undisclosed manner. It can be difficult, and sometimes impossible, to track down the source of this post. It was also impossible to get rid of these messages later. Several social media platforms such as Twitter, Instagram, Facebook, YouTube, Snapchat, Skype, and Wikipedia are the most common bullying sites on the internet.

Some of the social networking sites, such as Facebook, and the provision of guidance on the prevention of bullying. It has a special section that explains how to report cyber-bullying and to prevent any blocking of the user. On Instagram, when someone shares photos and videos made by the user to be uncomfortable, so the user can monitor or block them.

As the social lifestyle exceeds the physical barrier of human interaction and contains unregulated contact with strangers, it is necessary to analyze and study the context of cyberbullying. Cyberbullying makes the victim feel that he is being attacked everywhere as the internet is just a click away. It can have mental, physical, and emotional effects on the victim. Cyberbullying mainly takes place in the form of text or images on social media. If bullying text can be distinguished from non-bullying text, then a system can act accordingly. An efficient cyberbullying detection system can be useful for social media websites and other messaging applications to counter such attacks and reduce the number of cyberbullying cases. The objective of the cyberbullying detection system is to identify the cyberbullying text and also take its meaning into consideration. One first analyzes the various aspects of a particular text and then applies the previous information or visuals to find the context of the text. There is a need to create a personalized system that can access such a text effectively and efficiently.

2. Literature Survey

We have surveyed the existing projects and finally thought of making necessary modifications for getting the latest edition.

Existing System:-

There are several such systems deployed and made with different approaches like counting the number of positive and negative words and rate the sentence based on that. Or to check if the sentence contains any bad or abusive words. Several algorithms like SVM has been deployed for this purpose.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

Some of the examples are :-

1) BULLY BLOCKER : It is a machine learning-based system that uses natural language processing(NLP) techniques to detect cyberbullying in social media posts. The system analyzes text data to identify patterns of abusive language and flags posts that are likely to contain cyber bullying content.

2) CYBER SHIELD : It is another machine learning-based system that uses natural language processing(NLP) to detect cyberbullying. This system is designed to analyze text data from social media platforms and can identify instances of cyber bullying based on the tone, sentiment and content of the messages.

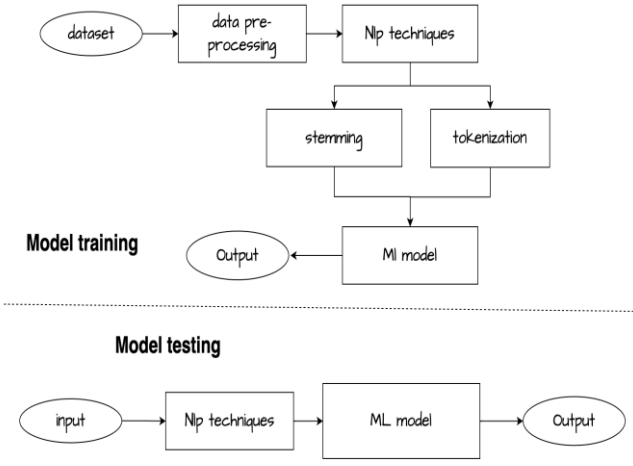
3) CYBER SIEVE : It is also a machine learning-based system that uses behavioral analytics to detect cyber bullying. This system analyzes online behavior patterns, including browsing history, social media activity and chat logs to identify users who may be engaged in cyber bullying.

4) ONLINE HATE SPEECH DETECTION : It uses machine learning techniques to detect hate speech and cyber bullying in online forums and social media platforms. This system uses a combination of NLP and deep learning algorithms to analyze text data and identify patterns of abusive language.

Proposed System:-

In this project, we will use all the different analyzes and techniques used to understand hate speech and improve the process so that it is accessible to different users to reduce the issue of hate speech. Tweets are collected from Twitter or datasets using natural language processing libraries and ML models, and we analyze their opinions. We'll classify harmful tweets as abusive content using these sentiments and additional abuse metrics. Current work uses natural language processing and machine learning techniques to understand hostile language, CNN algorithm is also used. We also used three machine learning models namely RFC (random forest classifier), SVM (support vector machine) and LRM (logistic regression model) for training and testing, among these three models logistic regression model has best accuracy, so we have saved that model and execution is done and then we have produced output of our project.

3. Implementation



Initially, data set is chosen and then it is pre processed using NLP techniques like stemming and tokenization. And then we split the data into training and testing sets. After splitting data, we use training set to train the machine learning models like random forest classifier, support vector machine and logistic regression models. Now we prepare the testing data set by cleaning it and we load those three trained machine learning models for testing and then we evaluate the performance of those models by comparing actual outputs with predicted outputs. After completion of testing these models, we have observed that logistic regression model has best accuracy, so we save this model for execution and then it produces the output based on the input given by the user.

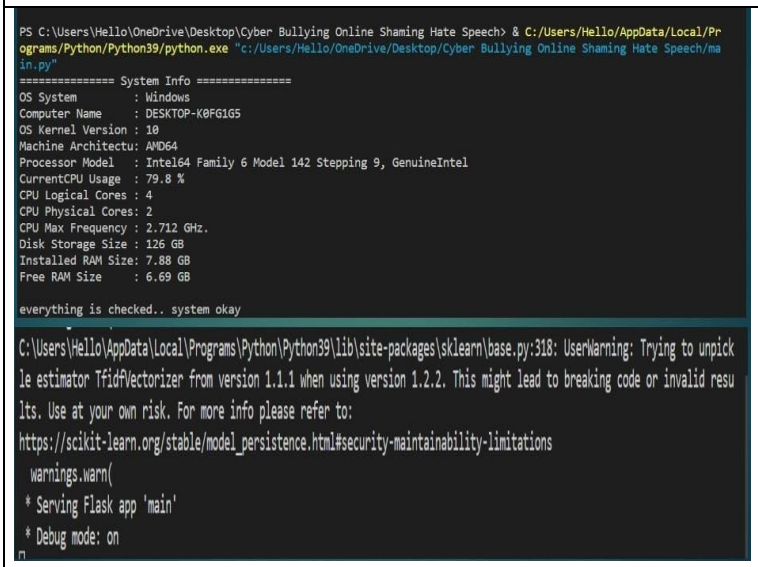
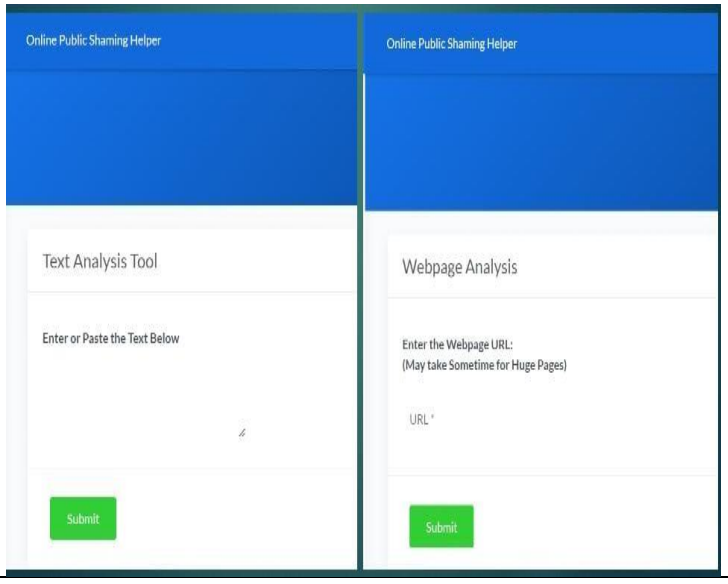
Working of LOGISTIC REGRESSION MODEL

Initially data is prepared by encoding categorical variables into numerical variables. Now logistic regression model is trained on the input data to estimate model parameters. Once the model parameters are estimated, sigmoid function is applied to linear combination of input features and parameter values to obtain predicted output.

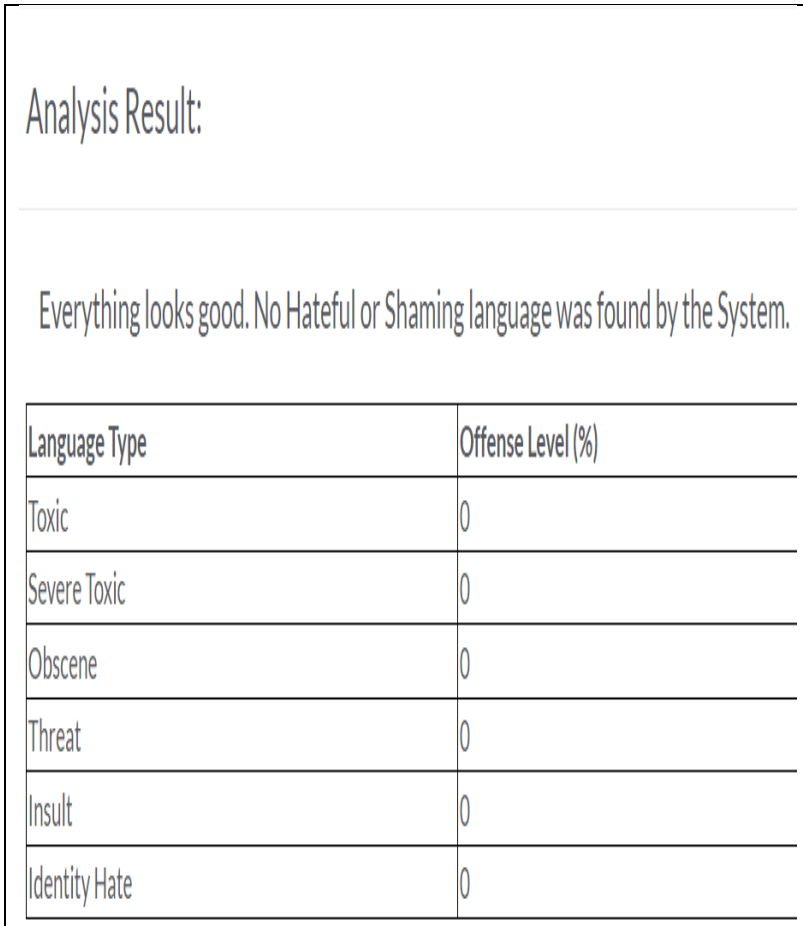
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

4.Results-

The proposed system is successfully implemented by detecting the toxicity and insulting contents of the input data set. We found that accuracy was great with no errors. Finally, with this we can detect the percentage of threat, insultive and toxic contents present in the data set and reduce the side effects of cyber bullying faced by cyber victims.

UI Design	*Design Description(functions,operations…)*
	This screen appears when we run the cyber bullying detection's main.py file and it gives assurance to the user that everything is checked and correct to proceed further. **1**
	These 2 pictures appear when we browse the following url present in main.py program code only after running the code. http://127.0.0.1:8080/ **2**

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362



This screen appears when your text input entered have positive meaning (OR) your entered web page url is secured.

3

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

Analysis Result:

Samples contained Hateful and Probelmatic language. The Language used is toxic. The Language used is little Obscene. The Language used is Insultive.

Language Type	Offense Level (%)
Toxic	99.57000000000001
Severe Toxic	0.22999999999999998
Obscene	21.85
Threat	3.52
Insult	55.97
Identity Hate	4.37

This screen appears when your text input entered is hateful or abusive or insultive (OR) your entered web page url have any threat.

4

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

6. Conclusion

In this paper, we proposed an approach to detect cyber- bullying using machine learning techniques. We evaluated our model on three classifiers namely RFC (Random Forest Classifier), SVM classifier (Support Vector Machine) and logistic regression classifier. Among these three classifiers, logistic regression has best accuracy, our work is definitely going to improve cyber-bullying detection to help people to use social media safely. However, detecting cyberbullying pattern is limited by the size of training data. Thus, a larger cyberbullying data is needed to improve the performance. Hence, deep learning techniques will be suitable in larger data as they are proven to outperform machine learning approaches over larger size data.

7. References

[1] N. A. Setyadi, M. Nasrun, and C. Setianingsih, ‘‘Text analysis for hate speech detection using backpropagation neural network,’’ in Proc. Int. Conf. Control, Electron., Renew. Energy Commun. (ICCEREC), Dec. 2018, pp. 159–165.

[2] J. Titcomb, ‘‘Facebook and Twitter promise to crack down on internet hatespeech,’’ The Telegraph. Accessed: Mar. 15, 2022. [Online]. Available: https://www.telegraph.co.U.K./technology/2016/05/31/facebook-and-twitter-promise-to-crack-down-on-internet-hate-speech/

[3] M. Rosemain. Exclusive: In a world first, Facebook to give data on hatespeech suspects to French courts. Reuters. Accessed: Mar. 15, 2022.

[4] Z. Waseem and D. Hovy, ‘‘Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter,’’ in Proc. NAACL Student

[5] A. A. Adebiyi, A. O. Adewumi, and C. K. Ayo, ‘‘Comparison of ARIMA and artificial neural networks models for stock price prediction,’’ J. Appl. Math., vol. 2014

[6] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti, ‘‘SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter,’’ in Proc. 13th Int. Workshop Semantic Eval., 2019, pp. 54–63.

[7] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung, ‘‘Multilingual and multi-aspect hate speech analysis,’’ 2019, arXiv:1908.11049.

[8] P. Burnap and M. L. Williams, ‘‘Cyber hate speech on Twitter: An appli- cation of machine classification and statistical modeling for policy and decision making,’’ Policy Internet, vol. 7, no. 2, pp. 223–242, 2015.

[9] Facebook. Facebook Community Standards—Hate Speech. Accessed: Mar. 15, 2022. [Online]. Available: https://www.facebook.com/communitystandards/hate_speech/

[10] YouTube. Hate Speech Policy. Accessed: Mar. 15, 2022. [Online]. Avail-able: https://support.google.com/youtube/answer/2801939?hl=en

8. Appendix

NLP - Natural Language Processing
CNN - Convolutional Neural Network
SVM - Support Vector Machine