IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

Online Shaming on Social Media: Analyze and Mitigation

Shraddha K Srinivas[1], E. Aparnaa[2], R. Kalpana[3], M. Preetha[4]
Students, Department of Information Technology[1,2]
Professor, Department of Information Technology[3]
Head of Department, Department of Information Technology[4]
Prince Shri Venkateshwara Padmavathy engineering college, Chennai, India

Abstract: *In this paper we would be discussing about online shaming and analyse different type of shaming and how to deal with it. The task of public shaming detection in social media is automated from the perspective of victims. It explores primarily about two aspects, namely, events and shamers. Based on classification of shaming tweets, a web application has been designed and deployed especially for one type of shaming tweet that of sarcasm, and the website also provide information about shamer who has used abusive comments under the user profile more than three time and sent alert message to user about the informing about the shamer.*

Keywords: Online Public Shaming, Machine Learning, Sentimental Analysis, Naïve Bayes, SVM, etc.

I. INTRODUCTION

Online shaming is a form of public shaming in which targets are publicly humiliated on the internet, via social media platforms (e.g., Twitter or Facebook), or more localized media (e.g., email groups). As online shaming frequently involves exposing private information on the Internet, the ethics of public humiliation has been a source of debate over internet privacy and media ethics [1]. The fundamental aspect of shaming is the societal processes of expressing social disapproval with the result of regret in the offender and/or disapproval from their peers.Public shaming in online social networks and related online public forums like Twitter has been increasing in recent years.The modernity of the phenomena and its subjectivity has led to cyber harassment ,cyber bullying and trolling.A suitable methodology is proposed for the detection and mitigation of the ill effects of online public shaming. In the past, work on this topic has been done from the perspective of administrators who want to filter out any content perceived as malicious according to their website policy. However, none of these considers any specific victim. On the contrary, we look at the problem from the victim's perspective.

II. LITERATURE SURVEY

A. SMOKEY

Abusive are one of the current hazards of on-line communication. While some people enjoy exchanging of abusive words, most users consider these abusive and insulting messages to be a nuisance or even upsetting. Smokey, a prototype system to automatically recognize email flames or abusive words was proposed which, combines natural-language processing and sociolinguistic observations to identify messages that not only contain insulting words but use them in an insulting manner. Smokey [2] bring about change in software field dealing about the online shaming process.

B. ONLINE SHAMING

This study focuses on the factors that lay behind an individual carrying out shaming on the internet as well as the prevalence of online shaming activity. One of the defining features of the internet landscape is the apparent anonymity it offers to the user. Anonymity on the internet can be a tricky concept to nail down. By using this anonymity many people leave comments or direct message them with some bad words with some bad meanings thinking that they would not be caught in the feature and can change their identity to some other if they caught by changes. The study on online shaming [3] also should how it may mentally affect a person and cause lot of mental problems for them.

C. BLOCKSHAME

Blockshame [4] is a website created as a project where the comments are filtered for good and bad comments. The abusive comments are categorized into many types of comments such as abusive, comparison, passing judgment, religious/ethnic, sarcasm/joke, and each tweet is classified into one of these types or as non-shaming.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

In this they have used support vector machine algorithm for categorizing the comments into shaming or non-shaming comments.

III. Existing System

Academic research was done earlier, it used different nomenclatures including abusive, flame, personal attack, bullying, hate speech, etc., often grouping more than a single category under a single name Efforts to moderate user generated content on the internet started very early, smoky is one of earliest works on classifying insulting post on labelled comments from web forms, In the existing system, large number of datasets cannot be compiled. Detection of shaming comments are not accurate. One major concern about it would be the finding of the comments which are about sarcasm which may not be detected by the normal algorithm as it may not look like an abusive comment but this would still bring down the mental health of the user.

IV. Proposed System

An application programming interface is proposed to satisfy all type of online shaming such as abusive words, especially sarcasm, jokes etc. Sentiment analysis and natural language processing is used analyse the whole comment under the user's social media account to identify the emotional tone behind a body of text The polarity prediction score is used to classify it as a negative, positive or neutral comment. naive Bayes, logistic regression and support vector machine (SVM) with a linear kernel is used here for improving the high accuracy rate when compare with previous method Here the polarity of reviews was identified correctly as we use more datasets and additionally feature is also added where if same person leave abusive comments more than 3 times then an alert message is to user's mail id.

V. System Architecture

The goal is classification of tweets automatically in given categories which classify if they are a sarcastic comment or not. The main functional units are shown in fig 1. The labelled training set and test set for each category go through the pre-processing and feature extraction steps. The training set is used to train the Random Forest (RM). A tweet is labelled non shame if all the classifier label it as negative.

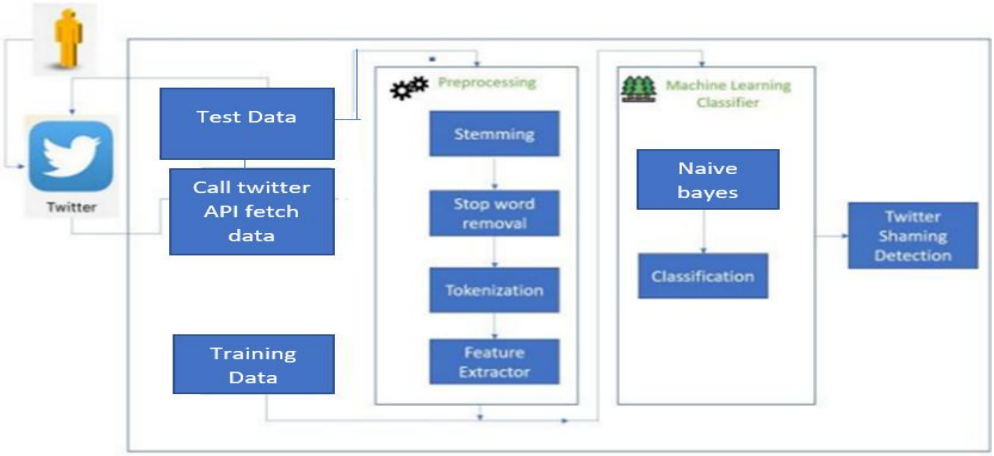

Figure 1: System Architecture

VI. Algorithm Used

A. SUPPORT VECTOR MACHINE WITH LINEAR KERNAL

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Linear Kernel [5] is used when the data is Linearly separable, that is, it can be separated using a single Line. It is one of the most common kernels to be used. It is mostly used when there are a large number of Features in a particular Data Set, Training the data set with linear kernel much faster than any other kernel which why we would be using the linear kernel. Previously this algorithm was used for classification of the good and bad comments in blockshame [4] web interface, but without the linear kernel was used.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

B. NAÏVE BAYES ALGORITHM

Naive Bayes is a classification algorithm that works based on the Bayes theorem. Bayes theorem is used to find the probability of a hypothesis with given evidence.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

By using Bayes theorem, [6] we can find the probability of A, given that B occurred. A is the hypothesis and B is the evidence. P(B|A) is the probability of B given that A is True. P(A) and P(B) is the independent probabilities of A and B. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e., every pair of features being classified is independent of each other. The dataset is divided into two parts, namely, **feature matrix** and the **response vector**. We generally use Gaussian naïve bayes continuous values associated with each feature are assumed to be distributed according to a **Gaussian distribution**

VII. CONCLUSION

Using Twitter API, all the sample dataset for training set has been collected and these datasets will be trained by SVM algorithm, Naïve Bayes algorithm. The main objective of the paper is to detect a comment which is a sarcastic comment as sarcasm comes under abusive words that is where many found it difficult to on how to detect the comment to be sarcastic or not. A web application is formed specially to find whether a comment is sarcastic or not and other abusive comment, this interface has register and login page for user who can give their social media account details, after which the system runs and finds about all the abusive comment left under user comment section. If a person leaves shaming comment more than 3 times, then user gets an alert message about the user warning us that the person left many abusive comments.

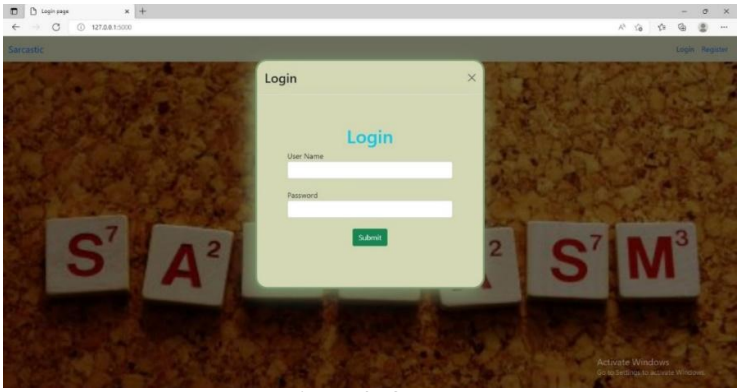

Figure 2: Output

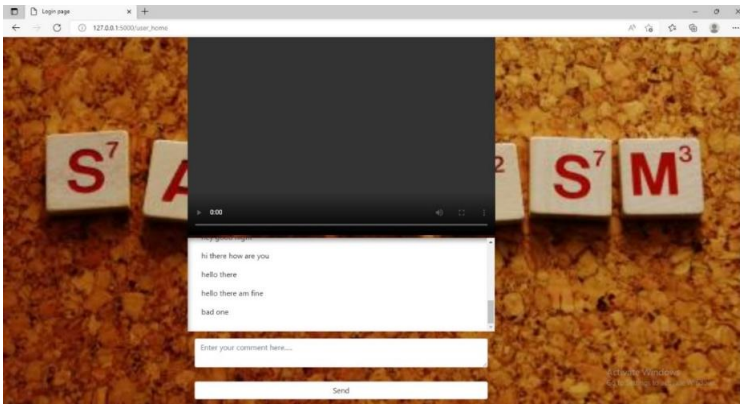

Figure 3: Output

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

Figure 4: Output

ACKNOWLEDGEMENT

We Would like to thank all our faculties, friends and family members who have directly and indirectly helped in easy completion of our project. Special thanks to Mrs S. Kalpana, Associate Professor, Department of Information Technology, Prince Shri Venkateshwara Padmavathy Engineering College, Chennai, Tamil Nadu, India for guiding us in completing this project successful.

REFERENCES

[1] https://en.wikipedia.org/wiki/Online_shaming
[2] E. Spertus, "Smokey: Automatic recognition of hostile messages," in Proc. AAAI/IAAI, 1997, pp. 1058–1065
[3] http://arno.uvt.nl/show.cgi?fid=143336
[4] https://www.researchgate.net/publication/331236549_Online_Public_Shaming_on_Twitter_Detection_Analysis_and_Mitigation.
[5] https://www.geeksforgeeks.org/creating-linear-kernel-svm-in-python/#:~:text=Linear%20Kernel%20is%20used%20when,in%20a%20particular%20Data%20Set.
[6] https://inpressco.com/wp-content/uploads/2021/02/Paper208997-1000.pdf