ONLINE SHAMING ON SOCIAL MEDIA: ANALYZE AND MITIGATION

APARNAA.E [1], SHRADDHA K SRINIVAS [2], R. KALPANA[3], Dr. M. PREETHA[4]
[1,2] STUDENT, DEPARTMENT OF INFORMATION TECHNOLOGY, PSVPEC, CHENNAI, INDIA
[3] ASSISTANT PROFFESOR, DEPARTMENT OF INFORMATION TECHNOLOGY, PSVPEC, CHENNAI, INDIA
[4] PROFFESOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, PSVPEC, CHENNAI, INDIA

ABSTRACT

- *In this paper we would be discussing about online shaming and analyze 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.*

.

Keyword: - *Online public shaming, machine learning, sentimental analysis, naïve Bayes*

1. 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

2. LITERATURE SURVEY

2.1 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.

2.2 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.

2.3 BLOCKSHAM

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

2.4 DETECTING HATE SPEECH ON WORLD WIDE WEB

We present an approach to detecting hate speech in online text, where hate speech is defined as abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation. While hate speech against any group may exhibit some common characteristics, we have observed that hatred against each different group is typically characterized using a small set of high frequency stereotypical words; however, such words may be used in either a positive or a negative sense, making our task similar to that of words sense disambiguation. In this paper we describe our definition of hate speech, the collection and annotation of our hate speech corpus, and a mechanism for detecting some commonly used methods of evading common "dirty word" filters. We describe pilot classification experiments in which we classify antisemitic speech reaching an accuracy 94%, precision of 68% and recall at 60%, for an F1 measure of. 6375.

2.5 A STUDY OF TEXT REPSENTATION OF HATE SPEECH DETECTION

The pervasiveness of the Internet and social media have enabled the rapid and anonymous spread of Hate Speech content on microblogging platforms such as Twitter. Current EU and US legislation against hateful language, in conjunction with the large amount of data produced in these platforms has led to automatic tools being a necessary component of the Hate Speech detection task and pipeline. In this study, we examine the performance of several, diverse text representation techniques paired with multiple classification algorithms, on the automatic Hate Speech detection and abusive language discrimination task.

3. EXISITING 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, smokey 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 look as not abusive comments but this would still bring down the mental health of the user.

4. 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 were 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 a alert message is to users mail id.

5. SOFTWARE REQUIREMENTS SPECIFICATION

5.1 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 labeled training set and test set for each category go through the preprocessing and feature extraction steps. The training set is used to train the Random Forest (RM). A tweet is labeled non shame if all the classifier label it as negative.

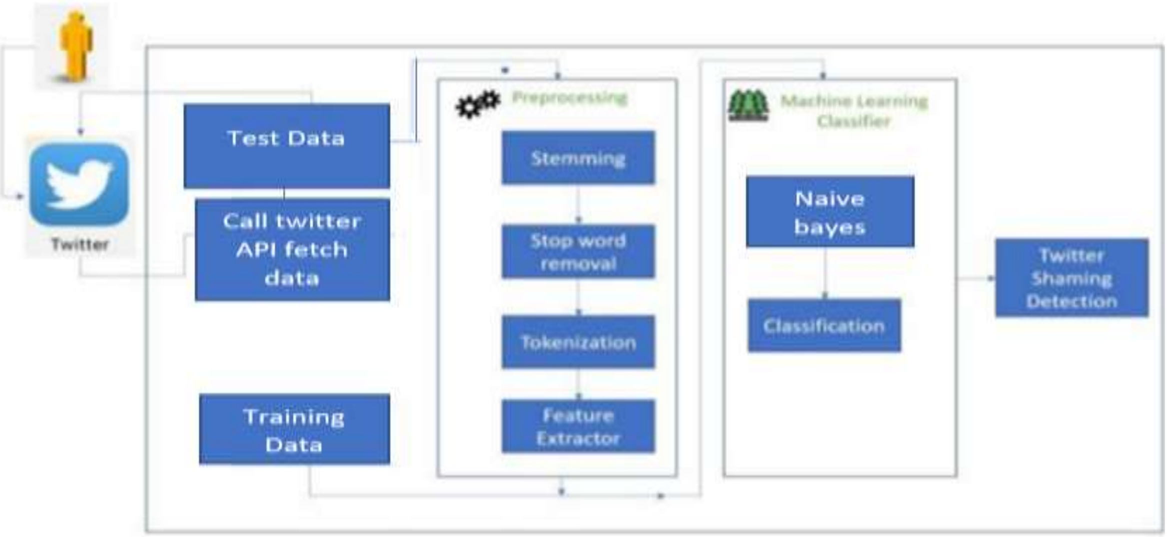
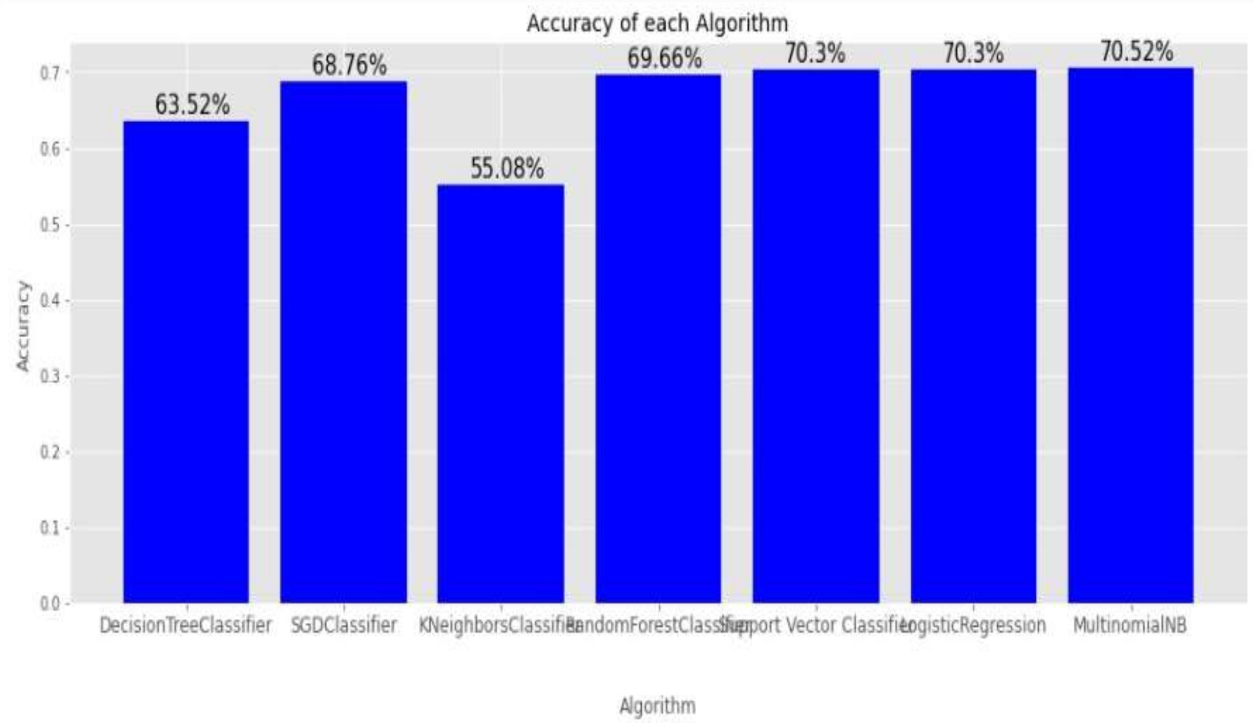


Fig -1: SYSTEM ARCHITECTURE

5.2 COMPARISON OF DIFFERENT ALGORITHM

Comparison of different algorithm has been done using jupyter model creation has taken place and accuracy of each of the model is noted with and decided on which of the algorithm is used for creation of the sarcasm detecting software

ALGORITHM	ACCURACY	AVAREAGE TIME
SGD CLASSIFIER	0.6876	0.331114054
KNNEIGHBOUR CLASSIFIER	0.5580	3.526567936
SVM CLASSIFIER	0.703	83.90283251
LOGISTIC REGRESSION	0.703	2.318800688
DECISION TREE CLASSIFIER	0.636	8.748605721
RANDOM FOREST CLASSIFIER	0.696	107.9681305
NAÏVE BAYES CLASSIFIER	0.705	0.008976221



6. ALGORITHM USED

6.1 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**

7. 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

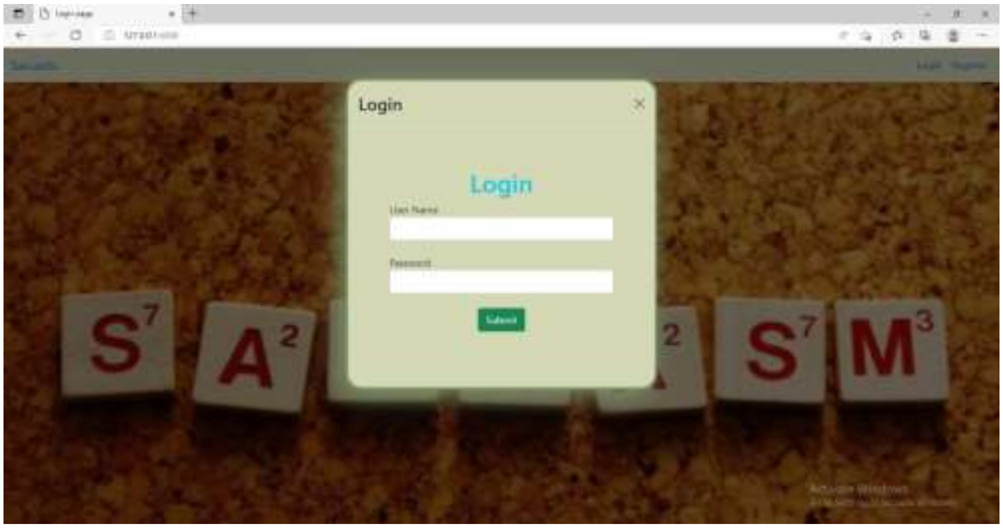
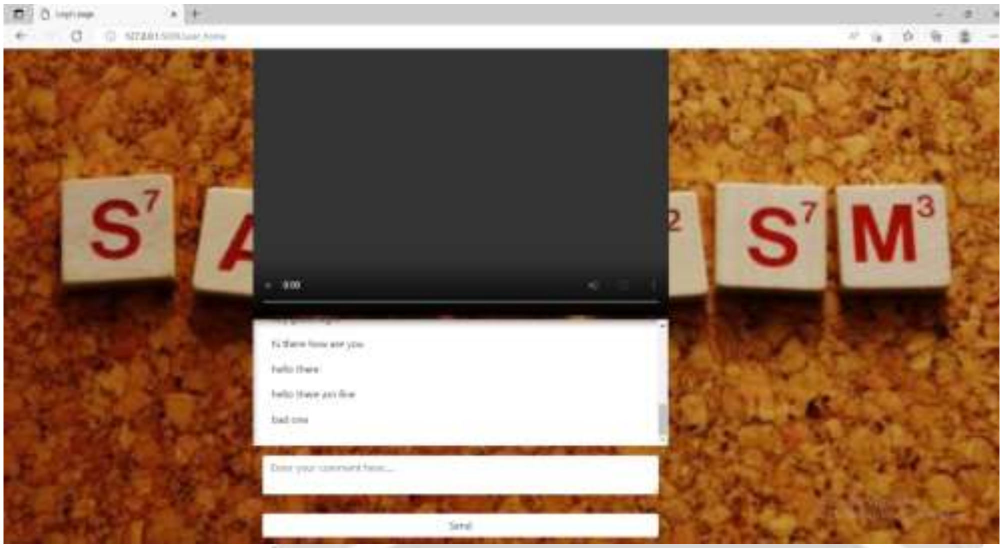


Fig -2: OUTPUT 1



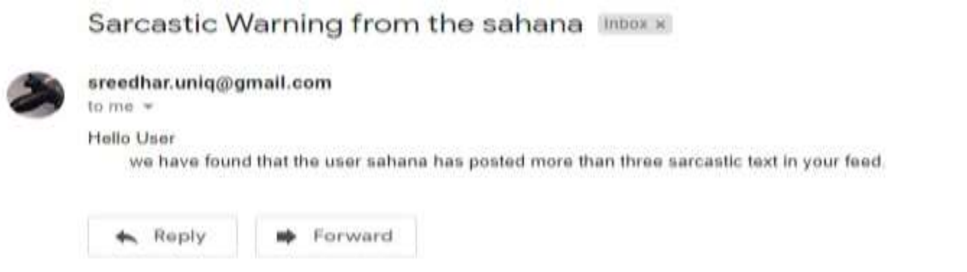
hi there how are you

hello there

hello there yes fine

bad one

Fig -3: OUTPUT



Sarcastic Warning from the sahana

sreedhar.uniq@gmail.com

to me

Hello User

we have found that the user sahana has posted more than three sarcastic text in your feed

Fig -4: OUTPUT

8. REFERENCES

1. E. Spertus, “Smokey: Automatic recognition of hostile messages,” in Proc. AAAI/IAAI, 1997, pp. 1058–1065
2. http://arno.uvt.nl/show.cgi?fid=143336
3. https://www.researchgate.net/publication/331236549_Online_Public_Shaming_on_Twitter_Detection_Analysis_and_Mitigation
4. W. Warner and J. Hirschberg, “Detecting hate speech on the world wide web,” in Proceedings of the Second Workshop on Language in social media. Association for Computational Linguistics, 2012, pp. 19–26.
5. A. Schmidt and M. Wiegand, “A survey on hate speech detection using natural language processing,” in Proceedings of the Fifth International Workshop on Natural Language Processing for social media. Association for Computational Linguistics, Valencia, Spain, 2017, pp. 1–10