IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Mitigation Of Online Public Shaming Using Machine Learning Framework

[1]Dige Vaishnavi A., [2]Gujar Shubhangi S.,

[1,2] *Computer Engineering,*

[1,2]*Dr. Babasaheb Ambedkar Technological University, Lonere.*

Abstract: Social networking sites involve billions of users worldwide. User interaction with these social networking sites, such as twitter, has major and unpleasant effects from time to time in everyday life. Major social networking sites have become a platform for users to spread a lot of irrelevant and unwanted information. Twitter has become one of the best platforms of all time and the most popular small blogging service that is widely used to share mindless ideas. In this proposed project change the task of detecting public embarrassment on Twitter. Embarrassing tweets are divided into nine categories: harassment, comparisons, judgment, religious, jokes on personal matters, profanity, spam, non-spam, whataboutery and each tweet is classified into one of these types or as non-embarrassing. It is evident that of all the participating users who post comments on a particular event, most of them are likely to embarrass the victim. Interestingly, it is also a disgrace that his followers say that he climbed faster than the shameless one on Twitter.

***Index Terms*- Remove shamers, Online user behaviour, Tweet analysis, Public shaming, Tweet classification.**

I.INTRODUCTION

Public stigma on social networking sites has increased over the years. These events have a devastating effect on the victim's social, political, and financial well-being. In various embarrassing situations, victims are given punishments that are not commensurate with the level of crime they have apparently committed. Twitter app help to prevent shamers from attacking the victim.

OSN is a website with apps designed to allow users to interact with other users or find users with similar interests. People from all over the world can keep in touch through social networks, regardless of their age. Those most in danger would be brought into a dark, violent world full of terrible things. Attackers host multiple attacks on social networking sites without the users knowing. People today have access to the Internet as an integral part of their daily lives. Many users use social media to share photos, music, videos, and more, which can link the user to other websites such as education, marketing, e-commerce, and business. Over time, social networking sites Facebook, LinkedIn, MySpace, and Twitter have gained popularity. Indigenous language analysis is about shame (e.g. English text found in tweets, social media comments, movies review, and political reviews are considered as a toolkit to detect fraud.

Twitter separates tweets into one of three different categories: insensitivity, harassment, or both. Social networks such as Twitter have seen an increase in social stigma in recent years. These incidents lead to serious personal, political, and financial consequences for victims. Victims are often equally punished for various forms of shame. Twitter web applications help people to avoid bullies.

A. Motivation

1. These days, social networks involve billions of users worldwide.
2. User interaction with these social networking sites, such as twitter, has major and sometimes unpleasant effects on everyday life.
3. Trolls disrupt meaningful conversations in online communities by posting unimportant comments.
4. Victims are given punishments that are not commensurate with the crime they have committed

This paper is organized as follows. Phase 2 is a related activity that outlines different approaches. Section 3 outlines open issues that highlight the challenges of the Program. The method of identifying and preventing such incidents is proposed in Section 4. The results and discussion are directly related to Section 5.

II. RELATED WORK

Research discovery of hate speech given the rapidly growing body of social media content, the number of hate speech online is also growing. Due to the large size of the web, methods are needed to detect hate speech. This study describes the key areas that have been tested for the automatic detection of these forms of speech using natural language processing and in this paper we also discuss the limitations of those methods [1].

Guntur et al. [2] has developed a hate speech modelling model that uses word representation with a continuous word processor (CBOW) and a fast text algorithm. This algorithm was chosen, as it is capable of achieving optimal performance, especially in the case of unfamiliar words using character level details. Based on this result, we see that there is nothing different, more diverse than anything else. Chaya Libeskind [3] the purpose of commenting is offensive or non-abusive. This paper forms the Hebrew corpus of user comments defined for abusive language. After that, we investigate the smallest n-gram representation and the denser n-gram character presentations for the subtle abuse of comments. As noted, social media is usually short, and we are investigating four ways to reduce the size, which produces word processors that wrap the same words in groups. Mukul et al. [4] Kaggle's toxic comment database is used to train in-depth reading models and classify comments into the following categories: toxic, highly toxic, obscene, threatening, insulting, and hateful. The database is trained in a variety of in-depth reading strategies and analyses which in-depth reading model is best for commentary sharing. In-depth learning strategies such as short-term memory (LSTM) with chemicals used with and without the name Glove, using the Convolution neural network (CNN) with or without Glove, and the Glove model associated with it is used for isolation.

Indonesian marginal language acquisition program by solving this problem as a task of isolating and resolving it using the following dividers: Naive Bayes, SVM and KNN. This paper also conducts the process of selecting a feature based on the value of the Partnership Information between words [5]. Aneta and Gareth [6] provide a complete set of features based on user symbols, as well as public graph metadata. The first includes metadata about the account itself, and the last is calculated from the public graph between each sender and recipient of each message. Person-based features are useful for defining user accounts on OSN, while graph-based features may reflect the power of information distribution across the network. In particular, in this paper you find the Jaccard index as a key element to reveal the dangerous or dangerous nature of messages directed at Twitter. To the best of our knowledge, we are the first to suggest the same metric to illustrate the abuse on Twitter. Justin Cheng [7] Both the negative attitude and seeing the individual trolley posts could mean that it greatly increases the chances of being trampled by the user, and together doubles this opportunity. The predictable model of trolling behaviour shows that mood and chatter together can define trolling behaviour better than human treadmill history.

Pinkesh et al. [8] describe the discovery of hate speech on Twitter as critical of applications such as the release of a controversial event, the creation of AI discussion bots, content recommendations, and emotional analysis. In this paper describe this function as being able to classify a tweet as racist, sexual or non-existent. The complexity of natural language construction makes this task a great challenge and this program conducts extensive experiments with many in-depth learning structures to learn the embedding of comprehension words to handle these difficulties.

Guanjun and Surya [9] mention that cyber bullying (harassment on social media) is known as a major social problem, especially among young people. It threatens the existence of social networking sites such as spam and what should have been emailed in the early days of the Internet. Current work to address this problem has included social and psychological studies on its prevalence and adverse effects on young people. While real solutions focus on teaching young people how to have positive relationships with people, few consider new social media software as a tool to reduce the problem. Reducing cyber bullying involves two key factors: effective discovery strategies and easy-to-use indicators that encourage users to think about their behaviour and decisions.

Hajime et al. [10] demonstrates social misconduct in the three online chat communities by analyzing users who were blocked in these communities. In this paper, it is found that such users tend to focus their efforts on a small number of threads, are more likely to post carelessly, and are more effective at collecting feedback from other users. Reading about the emergence of these users from when they joined the community until they were banned, I find that they not only write worse than other users over time, but also become increasingly intolerant by the community. In addition, this paper finds that human misconduct increases when the public response is extremely harsh. Our analysis also reveals that different groups of users have different levels of different behaviours that may change over time.

III. OPEN ISSUES

Much work has been done in this field due to its widespread use and application. In this section, some of the methods used to achieve the same purpose are mentioned. Due to restrictions from Twitter to API calls from applications, it is difficult to test algorithms in big data sets.

The biggest obstacles are,

- Creating and using web applications help to prevent shamers attacking victims on social media platforms such as Twitter.
- Category and automatic segmentation of derogatory tweets
- Creating and developing web-based applications helps prevent spammers from using machine learning.

IV. PROPOSED METHODOLOGY

A. Proposed Methodology

With the existing system, we begin the process of integrating adoption problems and minimizing the negative effects of online public embarrassment. Two Methods Segmentation and automatic editing of scandalous tweet & develop a web application for Twitter users to target Shamers.

Limitations Are,

- Twitter limits users to 150 applications per hour.
- Search is limited to 1500 post returns for a given request.
- The use of distance-based methods is very limited.

B. Proposed Solution

The goal of the proposed systemic approach is to use classification to help with the detection and mitigation of side effects on the online public standing of those that have been disgraced. The first of these contributions are: Classification and classification by automatics are used for disgracing tweets. Identify Shamers with a web application that is specific to twitter users.

To accomplish this, the system has to classify tweets into nine different categories. In figure 1 you can see the main functional units. The pre-processing and feature extraction steps are performed on the labelled training set and test set for each category. The tweet was labelled as embarrassing when all the team managers called it negative. Random Forest is trained on the training set (RF). Classifiers will not call something non-shame if the tweet says that it is so.

Objectives Are,

- To Mitigation and automatic classification of shaming and non-Shaming tweets.
- Help to block shamers attacking a victim on social media.
- To provide insights into shaming events and shamers.
- Try to improve classification accuracy using machine learning and real-time twitter data.

The main Purpose of Project is to find the tweets automatically in the nine types with the use of random forest algorithm

C. Architecture

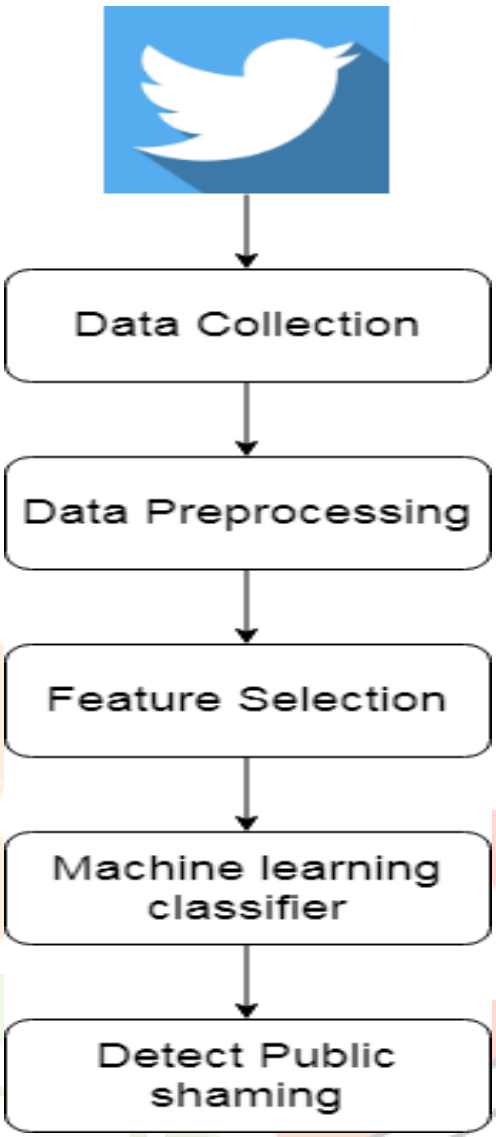

Fig. System Architecture

D. Algorithm

The algorithm used here is Random Forest. Random Forest is the most popular and powerful algorithm of machine learning

Step 1: Assume N as number of training samples and M as number of variables within the classifier.
Step 2: The number m as input variables to decide the decision at each node of the tree; m should be much less than M.
Step 3: Consider training set by picking n times with replacement from all N available training samples. Use the remaining of the cases to estimate the error of the tree, by forecasting their classes.
Step 4: Randomly select m variables for each node on which to base the choice at that node. Evaluate the best split based on these m variables in the training set.
Step 5: Each tree is fully grown and pruned (as can be done in the construction of a conventional tree trunk). To predict, a new sample is drawn under a tree. Training sample labelled at the final location. This process is repeated over all the trees in this union, and the average vote for all the trees is reported as a random guessing of the forest. i.e. dividers with the most votes.

E. Mathematical Modelling

The mathematical model for Mitigation of online public shaming is as-

$$S= \{I, F, O\}$$

Where,
I = Set of inputs
The input consists of a set of Words. It uses Twitter dataset.
F = Set of functions

$$F= \{F1, F2, F3\}$$

F1: Tweets Extraction
F2: Tweets Pre-processing
F3: Feature Extraction
F4: Shamming Classification
O: Shaming Detection and Block Shamers

F. Base Algorithm Technique

1. Sentiment Analysis using Sentiwordnet Dictionary
2. Probabilistic Latent Semantic Indexing (PLSI)
3. Latent Dirichlet Allocation (Keyword Extraction)
4. Part-of-speech Tagging Method
5. Random forest classification

G. Algorithmic Description:

1. Sentiment Analysis using Sentiwordnet Dictionary

```
polarizedTokensList ← newList()
whiletokenizedTicket.hasNext() do
token←tokenizedTicket.next()
lemma←token.lemma
polarityScore←null
ifDomainDictionary.contains(lemma,pos) then
ifSentiWordNet.contains(lemma,pos) and
SentiWordNet.getPolarity(lemma,pos) != 0) then
polarity Score ← SentiWordNet.getPolarity(lemma, pos)
else
domainDicToken←DomainDictionary.getToken(lemma, pos)
ifdomainDicToken.PolarityOrientation == ”POSITIVE” then
polarityScore ← DefaultPolarity.positive
else
polarityScore ← DefaultPolarity.negative
end if
end if
polarizedTokensList.add(token, polarityScore)
end if
end while
returnpolarizedTokensList
```

2. LDA Algorithm

First and foremost, LDA produces a procreative model that expresses how the documents in a dataset were created. In this context, a dataset is a collection of documents. Document is a collection of words. So our production model explains how each document gets its names. Initially, let's assume we know topic distributions for our dataset, meaning multinomial containing elements each, where is the number of terms in our corpus. Let βi represent the multinomial for the ith topic, where the sizeofβiisV:|βi|=V. Given this distribution, the LDA production process is as follows:
Steps:

1. for each document:

(a)Randomly select the distribution in the headings (variety of length K)
(b) For each word in the document:
(i) Probabilistically draw one of the topics from the distribution over topics obtained in (a), say topicβj
(ii) Probabilistically draw one of the words fromβj

3. Random Forest

The algorithm used here is Random Forest. Random Forest is the most popular and powerful algorithm of machine learning.

Step 1: Assume N as number of training samples and M as number of variables within the classifier.
Step 2: The number m as input variables to decide the decision at each node of the tree; m should be much less than M.
Step 3: Consider training set by picking n times with replacement from all N available training samples. Use the remaining of the cases to estimate the error of the tree, by forecasting their classes.
Step 4: Randomly select m variables for each node on which to base the choice at that node. Evaluate the best split based on these m variables in the training set.
Step 5: Each tree is fully grown and pruned (as can be done in the construction of a regular tree). To predict, a new sample is drawn under a tree. Training sample labelled at the final location. This process is repeated over all the trees in this union, and the average vote for all the trees is reported as a random guessing of the forest. i.e. dividers with the most votes.

4. Simulation Parameters

- Positive/Negative Tweets
- Similar Tweets using PLSA & Text Clustering Algorithm
- Topics extraction using LDA Algorithm
- Keyword extraction
- Shaming detection

V. RESULTS AND DISCUSSION

Using Twitter application programming interface (API), a large number of tweets are collected. Performance scores for the nine classifiers are shown in Fig 2. Twitter classifier analysis of existing and proposed systems are given in Table 1.

Table. 1 Twitter classifier analysis of existing and proposed system

Shaming Types	Support Vector Machine	Random Forest
Abusive	81%	86%
Comparison	71%	76%
Passing Judgment	68%	72%
Religious/Ethnic	52%	55%
Sarcasm/Joke	69%	73%
Whataboutery	47%	56%
Spam	65%	69%
Non-Spam	60%	65%
Vulgar	44%	50%

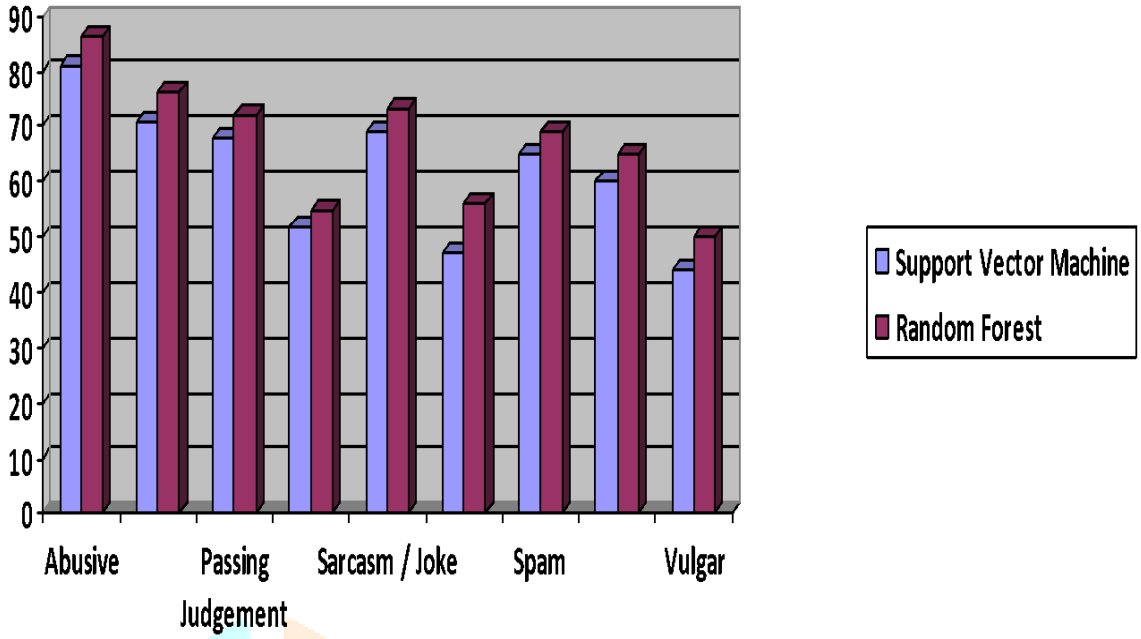

Fig. 2 Performance scores of Twitter classifier

CONCLUSION

Shame detection has led to the identification of Shaming content Shame words can be removed from social media. Shame detection has become very popular with the app. This program allows users to obtain aggressive word counts and data and their total percentage is calculated using machine learning classification. A possible solution to combat the threat of public embarrassment online on Twitter by dividing the shameful ideas into nine categories, choosing the right features, and designing a classier set to find you.

ACKNOWLEDGMENT

Express grateful and sincere gratitude to my project guide Prof. Gujar S. S. Mam for her precious collaboration and guidance that gave me during my research. Inspire me and provide me with all the laboratory facilities and allow me to carry out this research work in a very simple and practical way. I would also like to express my thanks to the (H.O.D.) S. R. Tandale sir, all faculty members and friends who support me during my hard work.

REFERENCES

[1] Rajesh Basak, Shamik Sural , Senior Member, IEEE, Niloy Ganguly, and Soumya K. Ghosh, Member, IEEE , " Online Public Shaming on Twitter: Detection, Analysis, and Mitigation", IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, VOL. 6, NO. 2, APR 2019.
[2] Guntur Budi Herwanto, Annisa Maulida Ningtyas, Kurniawan Eka Nugrahaz, Nyoman Prayana Trisna" Hate Speech and Abusive Language Classification using fast Text" ISRITI 2019.
[3] Chaya Liebeskind, Shmuel Liebeskind" Identifying Abusive Comments in Hebrew Facebook" 2018 ICSEE.
[4] Mukul Anand, Dr. R. Eswari" Classification of Abusive Comments in Social Media using Deep Learning" ICCMC 2019.
[5] Dhamir Raniah Kiasati Desrul, Ade Romadhony" Abusive Language Detection on Indonesian Online.News Comments" ISRITI 2019.
[6] Alvaro Garcia-Recuero, Aneta Morawin and Gareth Tyson Trolls layer: crowd sourcing and Characterization of Abusive Birds in Twitter" SNAMS 2018
[7] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, Jure Leskovec , "Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions", ACM-2017
[8] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, Vasudeva Varma, "Deep Learning for Hate Speech Detection in Tweets", International World Wide Web Conference Committee-2017
[9] Guanjun Lin, Sun, Surya Nepal, Jun Zhang, Yang Xiang, Senior Member, Houcine Hassan, "Statistical Twitter Spam Detection Demystified: Performance, Stability and Scalability", IEEE TRANSACTIONS – 2017.
[10] Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection", Digital Object Identifier – 2017.