Social Network Shaming Text: identification, inspection, reduction

Tailor Karan Rajesh[1], Navale Pradnya Pandurang [2], Bhatt Het Jaymin[3], Nehul Rutuja Vasant [4], Dhakulkar Bhagyashree [5]

[1,2,3,4]*UG. Student, Dept. of Computer Engineering, Dr. D. Y. Patil School of Engineering and Technology, Pune, India*

[5]*Professor, Dept. of Computer Engineering, Dr. D. Y. Patil School of Engineering and Technology, Pune, India*

***Abstract* - Twitter is a web social networking service that has quite 300 million users, generating an enormous amount of data daily. Twitter's most significant feature is its ability for users to tweet about events, situations, feelings, opinions, or maybe something new in real-time. there is no system of accuracy or reliability in place: Anyone can say just anything. It is a simple thanks to attacking your detractors for them to attack, the type of Twitter war. during this survey, various applications of machine learning and hate speech detection to ease the detection of shammers and shamming tweets were included. With the rise of online social networks, and the growth of publicly shaming events, voices against the callousness of the positioning owners are growing stronger, there is a necessity to investigate the shaming tweets, classify shamming tweets into different categories, and mitigate them by blocking them.**

***Index Terms* - Shaming, Social Network, OSN network, Performance metrics.**

I.INTRODUCTION

The limited knowledge of facts with the toxic nature of OSNs often translates into ignominy or loss or both for the victim. Unenthusiastic speech within the form of hate speech, bullying, profanity, flaming, trolling, etc in OSNs, is well studied within the literature. On the opposite hand, public shaming, which is that the condemnation of somebody who violates accepted social norms to arouse feelings of guilt in him or her, has not attracted much attention from a computational perspective. Nevertheless, these events are constantly being increasing for a few years. The immense volume of comments which is usually wont to shame an almost unknown victim speaks of the viral nature of such events. as an example, when Justine Sacco, a promotion person for an American Internet Company tweeted "Going to Africa. Hope I do not get AIDS. Just kidding. I'm white!" she had just 170 followers. Soon, a barrage of criticisms started pouring in, and thus the incident became one amongst the foremost talked-about topics on Twitter and also online, in general, within hours. She lost her job even before she landed in the state. Jon Ronson's "So You've Been Publicly Shamed" presents an account of the assorted online public shaming victims. The observation that we made from these diverse sets of events about the victims that are subjected to punishments disproportionate to the extent of crime they need committed. they need also formed a listing of victims, the year during which the event happened, the action that triggered public shaming together with the triggering medium, and its immediate consequences for every studied event The trigger is that the first action or word was spoken by the victim guilty for initiating public shaming. "Medium of triggering" is that the first communication media through which the final public became tuned in to the "Trigger." the implications for the victim, during or shortly after the event, are listed in "Immediate consequences." Henceforth, the two-letter abbreviations of the victim's name are accustomed to settling down with the respective shaming event. We proposed a system for automating the task of shaming tweet detection from the attitude of victims and exploring two major aspects which are events and shamers. Further, the shaming tweets are categorized into four types namely abusive, comparison, religious/ethnic, passing judgment, and every tweet is classed into one altogether these types or as non-shaming. It's discovered that out of all the participating users who post comments during a specific shaming event, the bulk of them are likely to shame the victim.

II.RELATED WORK

Rajesh [2] examine the shaming tweets which are classified into six types: abusive, comparison, religious, passing judgment, sarcasm/joke, whataboutery, and each tweet is classified into one of these types or as non-shaming. Support Vector Machine is used for classification. The web application called Block shame is used to block the shaming tweets. Categorization of shaming tweets, which helps in understanding the dynamics of the spread of online shaming events [12]. The probability of users trolling others generally depends on the bad mood and also noticing troll posts by others. Justin [3] introduces a trolling predictive model behavior that shows that mood and discussion together can show trolling behavior better than an individual's trolling history. A logistic regression model that precisely predicts whether an individual will troll in a mentioned post. This model also evaluates the corresponding importance of mood and discussion context. The experimental setup was a quiz followed by an online Discussion. Mind-set and talk setting together can clarify trolling conduct superior to a person's history of trolling. The multifaceted nature of the normal language development makes this undertaking testing and this framework perform broad examinations with different profound learning designs to learn semantic word embeddings to deal with this intricacy [15]. A deep neural network [8] is used for the classification of speech. Embedding learned from deep neural network models together with gradient boosted decision trees gave the best accuracy values. Hate speech refers to the use of attacking, harsh or insulting language. It mainly targets a specific group of people having a common property, whether this property is their gender, their community, race, or their believes and religion. Hajime Watanabe [7] finds a pattern-based approach that is used to detect hate speech on Twitter. The analysis also finds that the antisocial behavior of diverse groups of users of different levels can alter over time. Cyberbullying is broadly perceived as a genuine social issue, particularly for young people. Spammers sent spam emails in large volume and cybercriminals whose aim to get money from recipients that respond to email. Guanjun [5] assesses the detection accuracy, true positive rate, false-positive rate, and the F-measure; the stability inspects how effectively the algorithms perform when training samples are randomly selected and are of different sizes. Scalability aims to understand the effect of the parallel computing environment on the depletion of training and testing time of various machine learning algorithms. Random Forest would achieve better scalability and performance in a large-scale parallel environment. Vandebosch [15] gives a detailed survey of cyberbullies and their victims. There are a lot of reasons people troll others on online social media. Sometimes it is necessary to identify the posted whether the particular post is prone to troll or not. Panayiotis [8], shows the novel concept of troll vulnerability to characterize how susceptible a post is to trolls. for this, Built a classier that combines features related to the post and its history (i.e., the posts preceding it and their authors) to identify vulnerable posts. Twitter allows users to communicate freely, its instantaneous nature, and re-posting the tweet i.e. retweeting features can amplify hate speech. Twitter has a fairly large amount of tweets against some communities and is especially harmful in the Twitter community. Though this effect may not be obvious against a backdrop of half a billion tweets a day. Kwok [9] uses a supervised machine learning approach to detect hate speech on different Twitter accounts to pursue a binary classifier for the labels "racist" and "neutral". A hybrid approach for identifying automated spammers by grouping community-based features with other feature categories, namely metadata, content, and interaction-based features. K. Dinakar [11] contemplates three occasions that help to get an understanding of different parts of disgracing done through Twitter. A significant commitment of the work is a classification of disgracing tweets, which helps in understanding the elements of the spread of web-based disgracing occasions. It likewise encourages robotized isolation of disgracing tweets from non-disgracing ones. As online communities get large and the amount of user-generated data become greater in size, then the necessity of community management also rises. Sood [12] used a machine learning technique for automatic detection of bad user contributions. Every comment is labeled whether there exists the presence of insults, profanity, and the motive of the insults. These data are used for training Support vector machines and are combined with appropriate analysis systems in a multistep approach for the detection of bad user contributions. M. Hu and B. Liu [17] aimed to mine and to summarize customer reviews of a product from various merchant sites using features of the product on which the customer expressed opinions as positive or negative. Sarcasm or joking is nothing but using the words in such a way that meaning is opposite to tease others. For the mining of sarcasm tweets, communicative context improves the accuracy because Sarcasm requires some shared knowledge between speaker and audience. It helps to achieve the best precision values compared to purely linguistic characteristics in the detection of this sarcasm phenomenon[13].

III.METHODOLOGY

A) Dataset

We have used the Twitter dataset for classification purposes. On this, Social media Platform we can right our reviews, suggestions, messages, and upload pictures also which is considered as "tweets". Before, 2017 users could only tweet in 140 Characters. After 2017, Twitter has increased its limit up to 280 characters for all languages except Korean, Chinese, and Japanese. Users who are registered on Twitter can post tweets, like other's tweets, comment on other's tweets, retweet other's tweets, share tweets, and also communicate with others whereas Users who are not registered on Twitter are only able to see the message. So we have collected the dataset of these tweets from the Twitter developer account.

1)We use Twitter API for real-time data and 2) API website: apps.twitter.com

B) System Architecture:

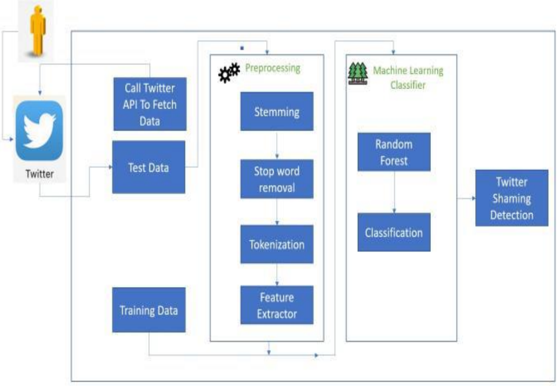


Fig 1: System Architecture

C) Methodology:

A. Preprocessing of Tweet

We perform a series of pre-processing steps before classification takes place. Named entity (NE) recognition, coreference resolution, and dependence parsing carried out using the Stanford CoreNLP library. All references to victims include names or surnames preceded by salutations, mentions, and so on, and are replaced with a uniform victim marker after the dependency parsing step. We also remove user mentions, retweet markers, hashtags, and links from the tweet text after dependency parsing and before parts of speech (POS) tagging with Stanford CoreNLP. After the elimination of useless entities, we convert all the letters to lowercase, and these are saved in a dataset named 'Clean Tweets'.

B. Shaming Tweet Detection

In the detection of Shaming Tweets, we have done feature extraction on the text of tweets. In this feature extraction, We have extracted positive and negative words from the tweets using feature vectorization. If a tweet contains any negative word then it will get labeled as 1 and if a tweet has no negative words then it will get labeled as 0 in a dataset.

After this, we trained our model and for that, we applied a total of three algorithms to check the accuracy. We applied to KNN, Random Forest, and Naïve Bayes. Among these algorithms, Random Forest gave us good accuracy that's Why proceed further with the Random Forest.

Algorithm: Random Forest

Random forest is a supervised learning algorithm that is used for both shaming detection and classification. However, it is mainly used for classification problems in which first it detects whether the tweet is shaming or not, and if it is shaming then it will classify it in a specific category. As I know that a forest is made up of trees and more trees means a more robust forest. Similarly, a random forest algorithm creates decision trees on data samples and then gets the prediction from each of them, and finally selects the best solution using Twitter. It is an ensemble method that is better than a single decision tree because it reduces overfitting by averaging the result. Working of Random Forest Algorithm

Step 1: First, start with the selection of random samples from a given Twitter dataset

Step 2: Next, What random forest does is it will construct a decision tree for every sample and after that, it will get prediction results from every decision tree.

Step 3: In this step, Random Forest will be performed for every predicted result.

Step 4: At last, select the emotion prediction result as the final prediction result.

So, That's how we trained our model for shaming detection of tweets using the Random Forest algorithm.

C. Tweet Classification

We are classifying the tweets automatically in the following four categories:

1. Abusive: When the shamer abuses victim then that comment will fall in this category. It is also observed that this list of abusive words is not enough to detect this kind of shaming, because some comments may contain abusive words which can still be in support of the victim. However, some abusive words associated with the victim as found from dependency parsing of the comment are a strong marker of abusive kind of shaming.

2. Religious: When there is maligning of religious group identities concerning the victim then it will fall under this category. We have given keywords of some religious identities and we also assume that we know the religious identity with which the victim associates. We have given many possible keywords for this shaming. Religious shaming is such a kind of shaming that can cause very serious issues or problems not only in the individual but also between communities. Whenever there is shaming with the religious identity which the victim associates with then that comment will get categorized in religious shaming, So then we can mitigate it.

3. Comparison: When the victim's action or behavior is compared or contrasted with another entity then that comment will fall in this category. Here, automatic detection of perception of the entity which is mentioned in the comment so that we can determine whether the comparison is shaming or not this is the main challenge. Sometimes text itself may not contain enough hints, e.g., adjectives with polarity related to the entity. In such cases, for the necessary context, the author of the comment depends on the collective memory of the social network users. Most of the time it is true when said entity appeared recently in other events.

4. Passing Judgement: When shamers pass the judgments vilifying the victim then this shaming falls under this category. Passing judgment often overlaps with other categories. But only those comments are passing judgments that do not fall under other categories. When someone judges with self-righteousness and does not apply the same standards to the actions and motivations so this falls under the passing judgment category. Passing judgment often starts with a verb and contains modal auxiliary verbs.

While training the classifier, shaming tweets from all categories and no-shame comments are assumed as negative examples. After that, based on the precision of the test set, classifiers are arranged by placing one with higher precision above one with lower precision. For classification, we have used Random Forest. We have used this because it reduces overfitting in decision trees which helps to improve the accuracy of our project and also it works well with categorical values. An equal number of tweets are sampled from all categories to classify them in the three categories mentioned above.

D. Mitigation

The control of irrelevant behavior on Twitter is being managed by the admin. Whenever a single account tweets 3 shamming tweets the account is automatically blocked from the admin side. The shammers are blocked from the admin side and only the admin has the rights to unblock the tweets.

The workflow of Blocking shammer includes the following steps:

1. Admin has the authorization of Blocking accounts with the occurrences of shaming tweets.
2. Admin can set a choice of actions according to the occurrences of shaming tweets.
3. If a user tweets shaming tweets 3 times the user is automatically logged out and the screen is flash with a Shaming Detected message.
4. If the tweets are non-shaming the screen is flash with no-shaming detected and the user can tweet the next tweet.
5. User side tweets are read and given as input to the classifier
6. The obtained tweets are classified using the Random Forest model.

7. Admin is responsible for making choices according to step 2
8. The blocked account is unblocked from the admin side.

D) Mathematical Model

Let S is the whole system consists:

S={ I, P, O}

WHERE

S = whole system

I = Input

P = process

O = output

I = {I0, I1, I2, I3, I4, I5}

I0 = Feedback post

I1 = bag-of-words

I2 = support of feed

I3 = confidence of feed

I4 = feeds of user

I5 = MODEL

P ={ P0, P1, P2, P3, P4, P5}

P0 = read posts

P1 = stop word removal

P2 = tokenization

P3 = train the model

P4 = classification of tweets

P5 = update the MODEL

O ={O0,O1,O2,O3}

O0 = token array

O1 = bag of words array

O2 = trained model

O3 = classification of MODEL

COMPARISION WITH EXISTING SYSTEM

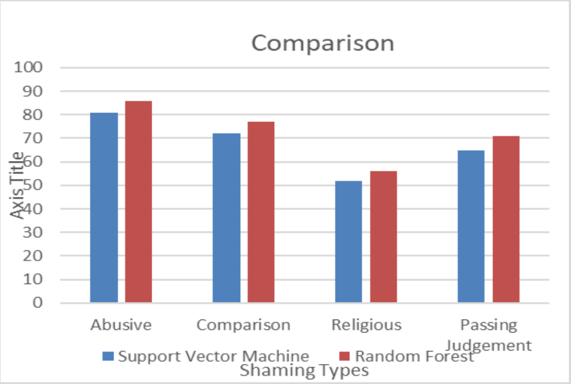


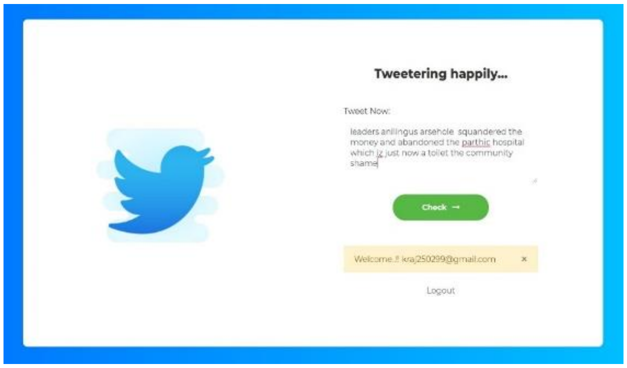
Chart: Comparison with Existing system

Shaming Type	Support Vector Machine	Random Forest
Abusive	81%	86%
Comparison	72%	77%
Religious	52%	56%
Passing Judgement	65%	71%

Table: Comparison with Existing system

V.APPLICATION RESULTS

A) Input Tweet

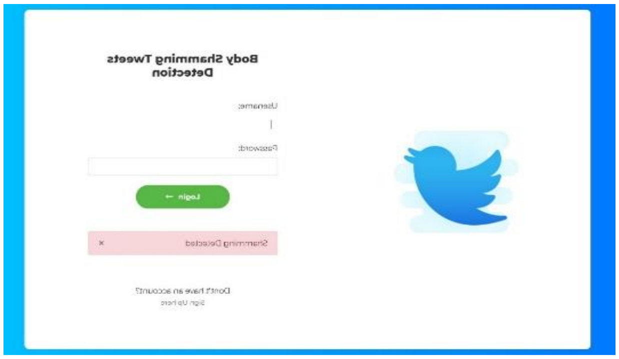

Tweetering happily...

Tweet Now:

leaders anilingus arsehole squandered the money and abandoned the parthic hospital which iz just now a toilet the community shame

Fig: Input Tweet Screen

B) Shamming Tweet Detection

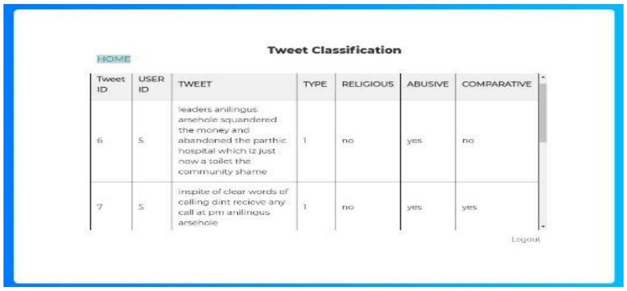

Body Shamming Tweets Detection

Username:

Password:

Fig: Shamming Detected

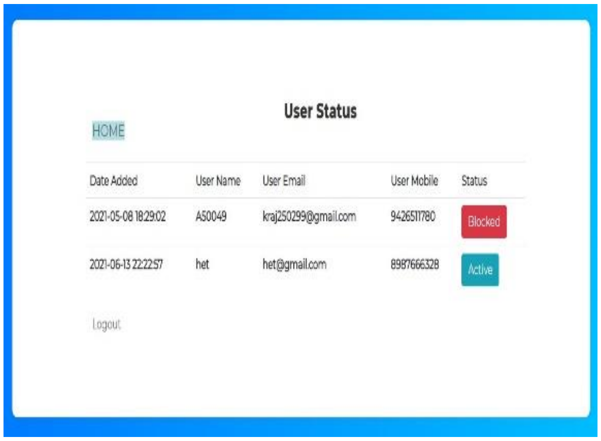
C) Tweet Classification


Tweet Classification

Tweet ID	USER ID	TWEET	TYPE	RELIGIOUS	ABUSIVE	COMPARATIVE
6	5	leaders anilingus arsehole squandered the money and abandoned the parthic hospital which iz just now a toilet the community shame	1	no	yes	no
7	5	inspite of clear words of calling dint recieve any call at pm anilingus arsehole	1	no	yes	yes

Fig.Classified Tweets

D) User Blocked



User Status

Date Added	User Name	User Email	User Mobile	Status
2021-05-08 18:29:02	A50049	kraj250299@gmail.com	9426511780	Blocked
2021-06-13 22:22:57	het	het@gmail.com	8987666328	Active

Fig: User Blocked

VI.CONCLUSION

Overall, the application is very important because as online social networks are increasing, public shaming events on social media platforms are also growing. So to prevent this kind of shaming events we need applications and software. This paper represents an analysis of the various applications of machine learning and hates speech detection to make the easy detection of shammers and shaming tweets. After going through the literature survey, a new system proposed a potential solution for detecting and categorizing shaming tweets into different categories using the Random Forest algorithm which is better than the Support vector machine which is being used in the existing system and if a tweet has been detected as shaming for three times from the same user then the user will get blocked.

REFERENCES

[1] K.Tailor,H.Bhatt,P.Navale,R.Nehul , “Survey On Shaming Sentence Detection On Social Network” in Int. Journal IJIRSET,2020.

[2] 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 transactionson computational social systems, vol. 6, NO. 2, APR 2019.

[3] Justin Cheng, Michael Bernstein, Cristian Danescu-NiculescuMizil, Jure Leskovec, “Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions”, ACM-2017

[4] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, Vasudeva Varma, “Deep Learning for Hate Speech Detection in Tweets”, International World Wide Web Conference Committee-2017

[5] Guanjun Lin, Sun, Surya Nepal, Jun Zhang, Yang Xiang, Senior Member, Houcine Hassan, “Statistical Twitter Spam Detection Demystified:

[6] Performance, Stability and Scalability”, IEEE transactions – 2017.

[7] Hajime watanabe, mondher bouazizi, and tomoakiOhtsuki, “Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection”, Digital Object Identifier – 2017

[8] Panayiotis Tsapara, “Dening and predicting troll vulnerability in online social media”, Springer - 2017.

[9] I. Kwok and Y.Wang, “Locate the hate: Detecting tweets against blacks,” in Proc. AAAI, 2013, pp. 1621–1622.

[10] Mohd Fazil and Muhammad Abulaish, “A Hybrid Approach for Detecting Automated Spammers in Twitter,” IEEE Transactions, 2019.

[11] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, “Common sense reasoning for detection, prevention, and mitigation of cyberbullying,” ACM Trans. Interact. Intell. Syst., vol. 2, no. 3, p. 18, 2012.

[12] S. O. Sood, E. F. Churchill, and J. Antin, “Automatic identification of personal insults on social news sites,” J. Assoc. Inf. Sci. Technol., vol. 63, no. 2, pp. 270–285, 2012.

[13] Rajesh Basak, Niloy Ganguly, Shamik Sural, Soumya K Ghosh,” Look Before You Shame: A Study on Shaming Activities on Twitter”, ACM 978-1-4503-4144-8/16/04.

[14] Rajadesingan, R. Zafarani, and H. Liu, “Sarcasm detection on Twitter: A behavioral modeling approach,” in Proc. 8th ACM Int. Conf. Web Search Data Mining, 2015, pp. 97–106.

[15] H. Vandebosch and K. Van Cleemput, “Cyberbullying among youngsters: Proles of bullies and victims,” New Media Soc., vol. 11, no. 8, pp. 1349–1371, 2009.

[16] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, “The Stanford CoreNLP natural language processing toolkit,” in

Proc. Assoc. Comput. Linguistics (ACL) Syst. Demonstrations, 2014, pp. 55–60. [Online].

[17] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining,2004, pp. 168–177.