Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 7, Issue 07, July -2020

Mitigation of Online Public Shaming Using Machine Learning Framework

[1]Mrs.Vaishali Kor, [2]Prof. Mrs. D. M. Gohil

[1]*Department of Computer Engineering, DY Patil college of Engineering, Akurdi, Pune*
Savitribai Phule Pune University Pune, India

Abstract— *In the digital world, currently some of the most far-reaching sites are social media sites on the internet. Billions of users are associated with social network sites. User interactions with these social sites, like twitter has an enormous and occasionally undesirable impact implications for daily life. Large amount of unwanted and unrelated information gets spread across the world using online social networking sites. Twitter is one of the most extensive platforms and it is the most popular micro blogging services to connect people with the same interests. Due to the popularity of twitter, it becomes a main target for shaming activities. Nowadays, Twitter is a rich source of human generated information which includes potential customers which allows direct two-way communication with customers. It is noticed that most of the participating users post comments in a particular occurrence are likely to embarrass the victim. Interestingly, it is also the case that shaming whose follower counts increase at higher speed than that of the nonshaming in Twitter. The proposed system allows users to find disrespectful words and their overall polarity in percentage is calculated using machine learning algorithm. Shaming tweets are grouped into nine types: abusive, comparison, religious, passing judgment, jokes on personal issues, vulgar, spam, non-spam and whataboutery by choosing appropriate features and designing a set of classifiers to detect it.*

Index Terms—*Shamers, online user behaviour, public shaming, tweet classification.*

INTRODUCTION

Online social network (OSN) is the use of dedicated websites applications that allow users to interact with other users or to find people with similar own interest Social networks sites allow people around the world to keep in touch with each other regardless of age [1] [7]. Sometimes children are introduced to a bad world of worst experiences and harassment. Users of social network sites may not be aware of numerous vulnerable attacks hosted by attackers on these sites. Today the Internet has become part of the people daily life. People use social networks to share images, music, videos, etc., social networks allows the user to connect to several other pages in the web, including some useful sites like education, marketing, online shopping, business, e-commerce and Social networks like Facebook, LinkedIn, Myspace, Twitter are more popular lately [8][9]. The offensive language detection is a processing activity of natural language that deals with find out if thereare shaming (e.g. related to religion, racism, defecation, etc.) present in a given document and classify the file document accordingly [1]. The document that will be classified in abusive word detection is in English text format that can be extracted from tweets, comments on social networks, movie reviews, political reviews. The work is divided into two parts: Shaming tweets are grouped into nine types. 1) Abusive 2) Comparison 3) Passing judgement 4) Religious 5) Sarcasm 6) What aboutery 7) Vulgar 8) Spam 9) Non spam. Tweet is classified into one of the mentioned types or as non-shaming. Public shaming in online social networks has been increasing in recent years [1]. These events have devasting impact on victim's social, political and financial life. In a diverse set of shaming events victims are subjected to punishments disproportionate to the level of crime they have apparently committed. Web application for twitter to help for blocking shamers attacking a victim.

II. LITERATURE SURVEY

Rajesh [1] examine the shaming tweets which are classified into six types: abusive, comparison, religious, passing judgment, sarcasm/joke, and whataboutery, and each tweet is classified into one of these types or as nonshaming. 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 spread of online shaming events [11]. The probability of users to troll others generally depends on bad mood and also noticing troll posts by others.

Justin [2] introduces a trolling predictive model behaviour shows that mood and discussion together can shows trolling behaviour better than an individual'strolling history. Alogistic regression model that precisely predicts whether an individual will troll ina mentioned post. This model also evaluates the corresponding importance of mood and discussion context. The model reinforces experimental findings rather than trolling behaviour being mostly intrinsic, such behaviour canbe mainly explained by the discussion's context, as well as the user's mood as revealed through their recent posting history. The experimental setup was quiz followed by online Discussion. Mind-set and talk setting together can clarify trolling conduct

superior to a person's history of trolling. Hate speech identification on Twitter is crucial for applications like controversial incident extraction, constructing AI chatterbots, opinion mining and recommendation of content. Creator characterize this errand as having the option to group a tweet as bigot, chauvinist or not one or the other. The multifaceted nature of the normal language develops makes this undertaking testing and this framework perform broad examinations with different profound learning designs to learn semantic word embedding to deal with this intricacy [14].

Deep neural network [3] is used for the classification of speech. Embedding learned from deep neural network models together with gradient boosted decision trees gave 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. Ternary classification of tweets into, hateful, offensive andclean. Hajime Watanabe [5] find sapattern-based approach which is used to detect hate speech on Twitter. Patterns are extracted in pragmatic way from the training set and define a set of parameters to optimize the collection of patterns. Reserved conduct is exacerbated when network input is excessively unforgiving. Analysis also finds that the antisocial behaviour of diverse groups of users of different levels that can alter over the time.Cyberbullyingisbroadly perceived asagenuinesocialissue, particularly for young people. Spammers sent spam emails in large volume and cybercriminals whose aim to get money from recipients that respond to email.

Gunjan [4] assesses the detection accuracy, true positive rate, false positive rate and the F-measure the stability inspect show effectively the algorithms perform when training samples are randomly selected and are of different sizes. The aim of scalability is 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 of parallel environment. Vandebosch [13] gives a detailed survey of cyberbullies and their victims. There are lot of reasons people troll others in online social media. Sometimes it is necessary to identify the post weather the particular post is prone to troll or not.

Panayiotis [6], shows novel concept of troll vulnerability to characterize how susceptible a post is to trolls. for this, built a classifier that combines features related to the post and its history (i.e., the posts preceding it and their authors) to identify vulnerable posts. Additional efforts have been done with random forest andSVM for classification. It shows Random forest performance is slightly outperforming. Twitter all owsusers to communicate freely, its instant aneous nature and re-posting the tweet i.e. retweeting features can amplify hate speech. As Twitter has a fairly large number of tweets against some community and are especially harmful in the Twitter community. Though this effect may not be obvious against a backdrop of half a billion tweets a day.

Kwok [7] use a supervised machine learning approach to detect hate speech on different twitter accounts to pursue a binary classifier for the labels "racist" and "neutral". Hybrid approach for identifying automated spammers by grouping community-based features with other feature categories, namely metadata, content, and interaction-based features. Random forest [8] gives best result in terms of all three metrics Detection Rate (DR), False positive Rate (FPR), and Score. Decision tree algorithm is good with regard to DR and F-Score. Bayesian network performs notably good with regard to False positive Rate (FPR) and F-Score, but it does not perform good enough with regard to Detection Rate (DR). Online informal organizations are frequently overwhelmed with scorching comments against people or organizations on their apparent bad behavior.

K. Dinakar [9] contemplates three occasions that help to get understanding into different parts of disgracing done through twitter. A significant commitment of the work is classification of disgracing tweets, which helps in understanding the elements of 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 usergenerated data become greater in size, then necessity of community management also rises. Sood [10] used a machine learning technique for automatic detection of bad user contributions. Every comment is labeled whether there exists 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 [15] aimed to mine and to summarize customer reviewsofaproductfromvariousmerchantsitesusingfeatures of the product on which the customer expressed opinions as positive or negative. Sarcasm or joking is nothing but use the words in such a way that meaning is opposite to tease others. For the mining of sarcasm tweet, 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 [12][16]. When any event happens, people discuss it on social media. The rapid growth of social media data on web has encouraged the big data mining area welcomed by researchers from academia as well as industry. It is challenging to figure out people's emotion on that. An innovative method of construction of Word Emotion Association Network (WEAN) [17] is used for emotion detection. Due to the popularityof Twitter, it became target for spamming activities. So, researchers started to apply different machine learning algorithms to detect spam in twitter. For the thorough evaluation a large data set of over 600 million public tweets is collected. it labelled around 6.5 million spam tweets and extracted 12 light-weight features, which can be used for online detection [18].

III. PROPOSED METHODOLOGY

In the proposed systemic approach, we formulate the task as classification of problem for the detection and mitigation of side effects of online public disgracing. Two main contributions are: 1) Categorization and automatic classification of disgracing tweets. 2) Develop a web application for Twitter user to identify Shamers.

A. Architecture

The goal is classification of tweets automatically in nine categories. 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 nonshame if all the classifiers label it as negative.

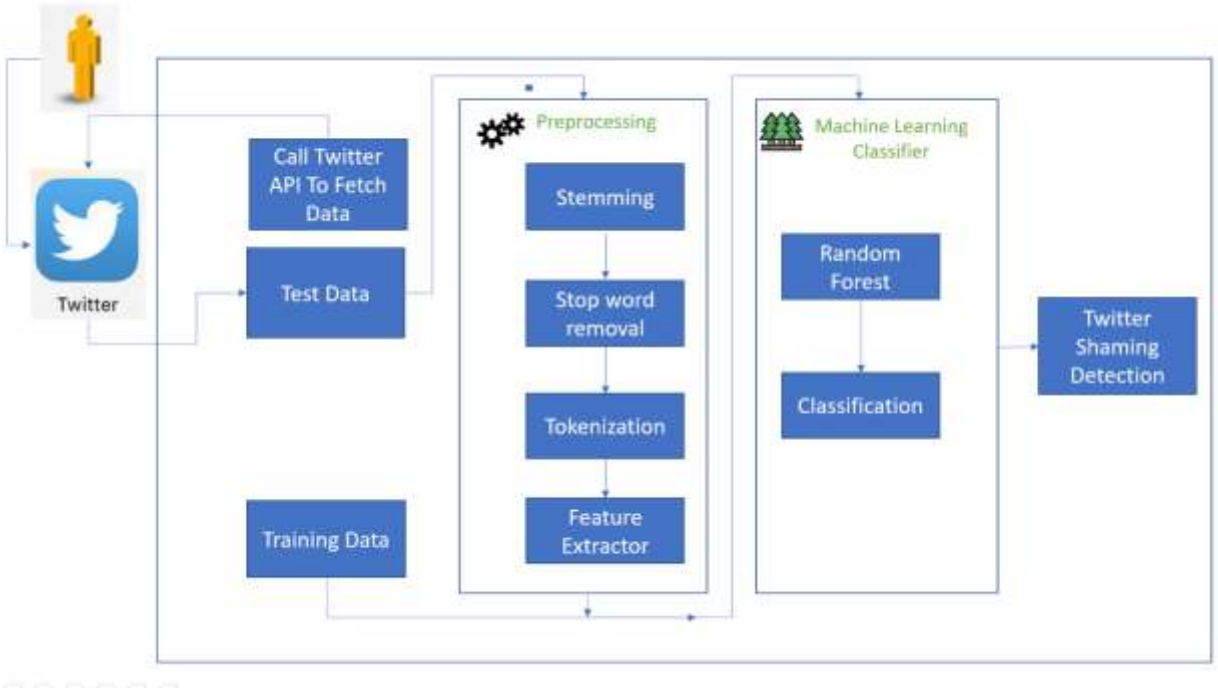

Fig. 1. System Architecture

B. 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 not pruned (as may be done in constructing a normal tree classifier). For forecasting, a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in. This procedure is repeated over all trees in the ensemble, and the average vote of all trees is reported as random forest prediction. i.e. classifier having most votes.

C. Mathematical Model The mathematical model for Shamming Detection System is as

S = {I, F, O}

Where, I = Set of inputs

The input consists of set of Words. It uses Twitter dataset. F = Set of functions

F = {F1, F2, F3....FN}

F1: Tweets Extraction
F2: Tweets Preprocessing
F3: Feature Extraction
F4: Shamming Classification
O: Shamming Detection and Block Shamers

D. Dataset Twitter

Twitter dataset is used for the classification purpose.In this social networking service users can freely communicate. They post and communicate with messages known as" tweets". Originallytherewasrestrictionoftweetscharactersthatis140, but from November 7, 2017, this limit was increased to 280 for all languages except Chinese, Japanese, and Korean. Registered users can post, like, and retweet tweets, but unregistered users can only read the messages. Users access Twitter through its website interface, through Short Message Service (SMS) or its mobile-device application software ("app"). Twitter, Inc. is based in San Francisco, California, and has more than 25 offices around the world.

1) We use twitter real-time data using Twitter API.
2) apps.twitter.com API Website.

IV. RESULTS AND DISCUSSION

Using Twitter application programming interface (API), a large number of real time tweets are collected. Then to understand the overall nature of tweets sentiment analysis is performed. Finally, after semantic analysis shamming classification is done. Evaluation metrics for each run are shown in fig.

Evaluation Metrics	Support Vector Machine	Random Forest
Precision	59%	60.78%
Recall	63.15%	69.92%
F-measure	61%	65.03%
Accuracy	75%	79.27%

Table 1: Comparison with Existing system

Fig. 2: Classification Performance

V. CONCLUSION

Shaming detection has led to identify Shaming contents. Shaming words can be mined from social media. Shaming detection has become quite popular with its application. In this proposed system allows users to find offensive word counts with the data and their overall polarity in percentage is calculated using Random Forest. Potential solution for countering the menace of online public shaming in Twitter by categorizing shaming comments in nine types, choosing appropriate features, and designing a set of classifiers to detect it.Furthermore, we intend to continue to explore new problems from the point of view of a social network service provider, such as Facebook or Instagram, to improve the well-being of OS Nusers and We would like to improve our experiments with an even larger annotated data set to improve the performance further.

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 ONCOMPUTATIONALSOCIALSYSTEMS,VOL.6,NO.2,APR2019.

[2] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, Jure Leskovec, "Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions", ACM-2017.

[3] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, Vasudeva Varma, "Deep Learning for Hate Speech Detection in Tweets", International World Wide Web Conference Committee-2017.
[4] 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.
[5] 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.
[6] Panayiotis Tsapara, "Defining and predicting troll vulnerability in online social media", Springer - 2017.
[7] I. Kwok and Y.Wang, "Locate the hate: Detecting tweets against blacks", in Proc. AAAI, 2013, pp. 1621–1622.
[8] MohdFazil and Muhammad Abulaish, "A Hybrid Approach for Detecting Automated Spammers in Twitter", IEEE Transactions, 2019.
[9] 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.
[10] 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.
[11] Rajesh Basak, NiloyGanguly, Shamik Sural, Soumya K Ghosh, "Look Before You Shame: A Study on Shaming Activities on Twitter", ACM 978-1-4503-4144-8/16/04.
[12] 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.
[13] H. Vandebosch and K. Van Cleemput, "Cyberbullying among youngsters: Profiles of bullies and victims", New Media Soc., vol. 11, no. 8, pp. 1349–1371, 2009.
[14] 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].
[15] 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.
[16] D. Bamman and N. A. Smith, "Contextualized sarcasm detection on Twitter," in Proc. ICWSM, 2015, pp. 574–577.
[17] DANDAN JIANG1, XIANGFENG LUO1,2, JUNYU XUAN3, AND ZHENG XU4, "Sentiment Computing for the News Event Based on the Social Media Big Data", IEEE Acess-2016.
[18] Hao Chen; Jun Zhang; Xiao Chen; Yang Xiang; Wanlei Zhou, "6 million spam tweets: A large ground truth for timely Twitter spam detection", IEEE Int. conference- 2015.

