iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** III **Month of publication:** March 2020

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue III Mar 2020- Available at www.ijraset.com

A Survey on an Approach to Reduce the Impact of Online Public Shaming using Machine Learning Framework

Vaishali Kor[1], Prof. D. M. Gohil[2]
[1]*Post Graduate Student,* [2]*Assistant Professor, Department of Computer Engineering, DY Pail College of Engineering, Akurdi, Pune*

Abstract: ***In the digital world, 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 irrelevant information gets spread across the world using online social networking sites. Twitter has become one of the most enormous platforms and it is the most popular micro blogging services to connect people with the same interests. 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 out of all the participating users who post comments in a particular occurrence, majority of them 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 non-shaming in Twitter. For reducing the impact of public shaming, the tweets can be categorized and those susceptible for shaming can be blocked.***
Keywords: Shamers, online user behaviour, public shaming, tweet classification.

I. 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 there are 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. For reducing the impact of public shaming, the tweets can be categorized and those susceptible for shaming can be blocked.

II. RELATED WORK

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 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 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'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 model reinforces experimental findings rather than trolling behaviour being mostly intrinsic, such behaviour can be 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 and clean. Hajime Watanabe [5] finds a pattern-based approach which is used to detect hate speech on Twitter. Patterns are extracted in pragmatic way from the training set and dene 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. 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. Gunjan [4] assesses the detection accuracy, true positive rate, false positive rate and the F-measure; the stability inspect show effectively that 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 classier 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 and SVM for classification. It shows Random forest performance is slightly outperforming. Twitter allows users to communicate freely, its instantaneous 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 FScore. 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 user generated 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. 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 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].

III.OPEN ISSUES

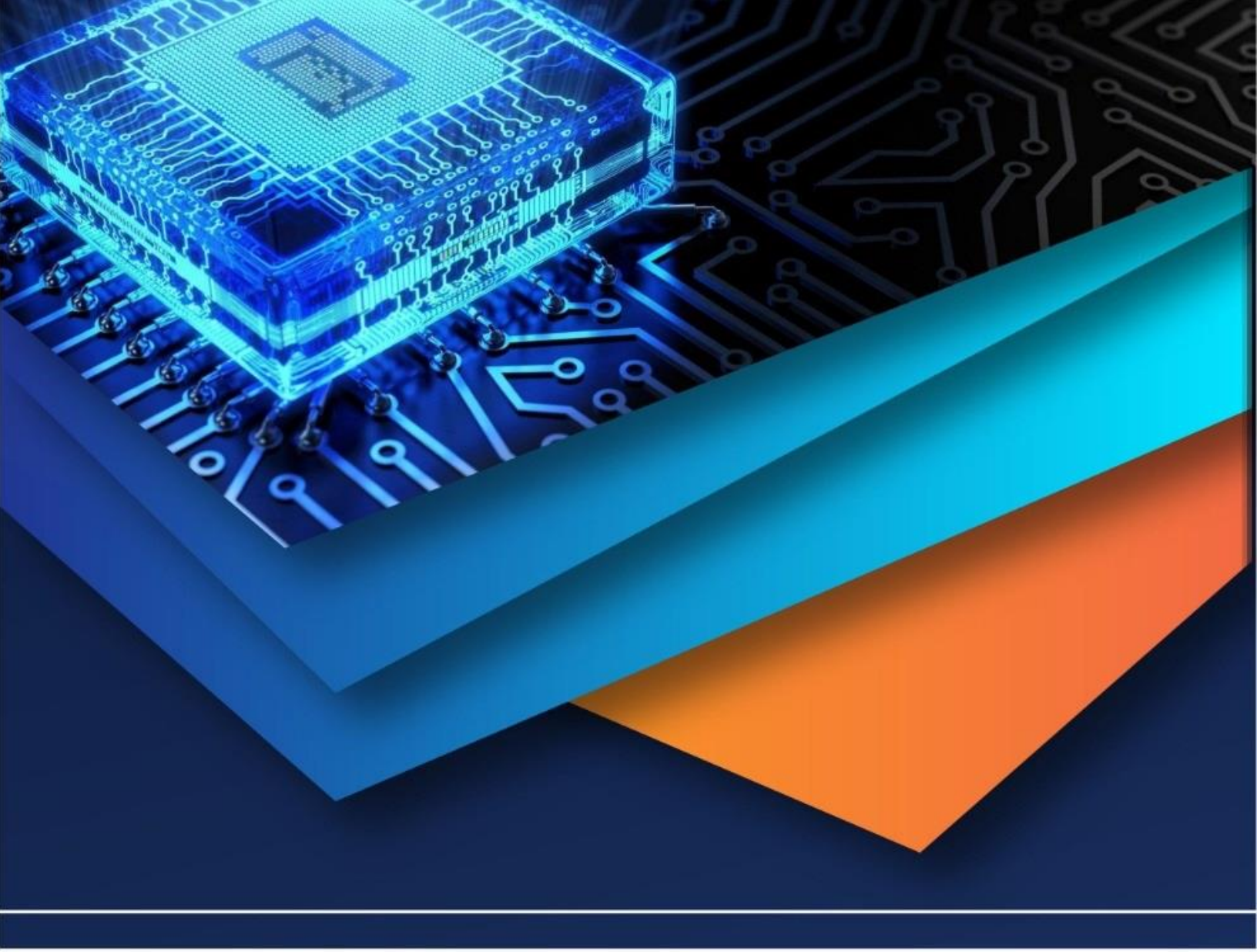
Lot of work has been done in this field because of its extensive usage and applications. In this section, some of the approaches which have been implemented to achieve the same purpose are mentioned. Due to limitations from Twitter on API calls from application, it is difficult to test algorithm on large date sets.

IV.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 survey allows users to find offensive word counts with the data and their overall polarity in percentage is calculated using classification by machine learning. 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.

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, APR2019.
[2] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, Jure Leskovec, "Anyone Can Become a Troll: Causes of Trolling Behaviour 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, "Dening 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] Mohd Fazil 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, 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.
[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: Proles 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. HuandB. Liu, "Mining and summarizing customer reviews," inProc. 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.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY