iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VII **Month of publication:** July 2021

DOI: https://doi.org/10.22214/ijraset.2021.37020

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 9 Issue VII July 2021- Available at www.ijraset.com

Public Trolling Detection on Twitter Using Machine Learning

Miss. Pooja Dilip Dhotre[1], Dr. N. A. Deshpande[2]
[1]*Student,* [2]*Professor, Gokhale Education Society's R. H. Sapat College of Engineering, Nashik*

Abstract: ***Social media websites are among the internet's most far-reaching digital sites. Billions of social network users exist Users' frequent interactions with social networking sites, like Twitter, have a widespread and sometimes unfortunate effect on day-to-day life. Social networking sites make it easy for large amounts of unwanted and unrelated information to spread around the world. Twitter is a popular micro blogging service where users connect with others with similar interests. Because of the current popularity of Twitter, it is vulnerable to public shaming. Recently, Twitter has emerged as a rich source of human-generated information, with the added benefit of connecting you with customers and enabling two-way communication. It is generally accepted that when someone posts a comment in an occurrence, it is likely to humiliate the victim. The fact that shaming users' follower counts increase faster than that of the people who don't use shame is interesting. Using machine learning algorithms, users will be able to identify disrespectful words, as well as the overall negativity of those words, which is displayed in a percentage.***

Index Terms: Trolls, online comments, public shaming, categorizing tweets.

I. INTRODUCTION

Using dedicated websites applications, people use the online social network (OSN) to interact with each other or to find those with similar interests. People of all ages can stay in touch regardless of where they are on the globe with social network sites like Facebook. For some children, exposure to unpleasant experiences and mistreatment becomes part of their experience in the world. A lot of the attacks on social network sites remain unnoticed by those who use the sites. Today, the Internet has become an important part of daily life for everyone. Social networks enable users to connect to many other web pages, such as education, marketing, online shopping, business, e-commerce, and. More and more people use social networks like Facebook, LinkedIn, Myspace, and Twitter these days. In order to find out if there are offensive (e.g. related to religion, racism, defecation, etc.) language present in a given document, a programmer analyses the content and classifies the file accordingly. A large collection of texts—tweets, blog posts, social media comments, etc.—in English will be used to classifier the document that will be classified in abusive word detection. The usage of Twitter is classified into one of the aforementioned categories or it is non-degrading. Recent years have seen an increase in public shaming in online social networks. This kind of attack has an especially destructive effect on the victims' social, political, and financial well-being. In diverse shaming scenarios, victims are disproportionally subjected to punishments that go far beyond the level of criminal culpability they appear to have demonstrated.

II. LITERATURE SURVEY

The “shaming tweets” that were categorized into six types are investigated and classified according to abusive, comparison, religious, passing judgment, sarcasm/joke, and whataboutery. Classification is made possible by SVMs. Block shame is a web application that is used to stop the bullying tweets. It aids in our understanding of how the spread of online shaming events progresses when we categorise the shaming tweets. Most users will troll when they are in a bad mood, and they will notice troll posts if they are looking [1].

An advanced trolling predictive model is introduced when put together, discussions and moods provide more information about a troll's identity than any individual characteristic. A logistic regression model that is perfectly capable of accurately predicting whether or not a particular individual will troll in a discussion thread mentioned. Additionally, the model will also consider mood and discussion context to be of equal importance. The model aims to confirm experimental findings, and does not promote trolling behavior as being mostly intrinsic. The discussion's context, as well as the users recent posting history, are important factors in this regard. After an experiment, people were asked to fill out a questionnaire followed by an online discussion. As well as the mind-set and the talk setting, understanding trolling behavior solely through a person's history of trolling falls short. It is critical for programmers like “controversial incident extraction,” “AI chatterbots,” “opinion mining,” and “content recommendation” to have hate speech identification in place on Twitter.

©IJRASET: All Rights are Reserved

She sees this task as allowing her to categories a tweet as sexist, chauvinistic, or bigoted. Normal language develops in such a way that the project's tests must meet numerous challenges. The project framework must, as a result, encompass various sophisticated learning designs to learn semantic word embedding to meet this complexity [2].

Speech classification is being accomplished with deep neural networks. By blending gradient boosted decision trees with deep neural network models, the highest accuracy values were attained.

The term hate speech refers to insults, profanity, or hostile language. It is directed at a specific demographic group, whether they are people of a certain gender, people who come from a certain community, or a group of people who all have the same race or religion. Clean, offensive, and hateful tweets are organized into the ternary classification of primary, secondary, and tertiary tweets. Using a pattern-based approach, Hajime Watanabe has found that Twitter is used to express hate speech. Instead of "cherry-picking" patterns from the training set, we extract them based on practical needs and define a list of parameters to optimize the collection of patterns. When network input is unforgiving, reserve conduct becomes even more imperative. In addition, analysis reveals that diverse groups of users with varying levels of antisocial behavior can exist at any time. Young people consider cyber bullying to be a serious social problem. Because of the large volume of spam emails sent, spammers and cybercriminals whose goal is to obtain money from responders all utilized this strategy [3].

Uses a true positive rate, false positive rate, and F-measure to assess how stable the detection process is runs algorithms through simulated data where randomly-chosen samples are of varying sizes to see how well they work. The purpose of scalability is to discover the relationships between parallel processing and machine learning algorithm training and testing time. In a parallel environment, Random Forest can perform better scalability and performance. As a survey of cyber bullies and their victims, Vandebosch offers a detailed assessment. Many people like to torment others on social media for a variety of reasons. It is important to identify when the post is likely to be trolled due to inclement weather [4].

This show demonstrates a novel concept in trolling called "troll vulnerability," showing how prone a post is to trolls. By building a classifier that includes features such as previous posts and authors, these article identifies vulnerable posts. New algorithms, such as random forest and SVM, have been applied to classification. Random forest performance appears to be slightly better. While Twitter gives users the ability to communicate freely, it also facilitates an amplification of hate speech by the practice of re-tweeting tweets, which is also referred to as retweeting. Twitter contains many harmful tweets about a particular community, and those tweets are especially problematic for the community on Twitter. Despite the enormous number of tweets generated every day, this can go unnoticed unless you're looking for it [5].

For the purposes of our binary classifier, we will utilize a supervised machine learning approach to determine if various twitter accounts contain hate speech by looking for “racist” and “neutral” phrases. A hybrid method of classifying automated spammers takes into account features that are provided to the community, like metadata, content, and interactivity, as well as other features, such as metadata, content, and interaction. Random forest [8] has the best detection rate, false positive rate, and score on all three metrics. DR and F-Score can be optimized using the decision tree algorithm. In comparison to other supervised learning algorithms, the Bayesian network is far better at reducing the FPR (False Positive Rate) and F-Score (False Alarm Rate), but it doesn't quite cut it when it comes to the overall detection rate (DR). Online organizations run the risk of being inundated with scalding comments about poor behavior [6].

Studies three times that aid in explaining disgracing that is caused through Twitter. Dedicating an inordinate amount of time to classifying disgracing tweets serves an invaluable purpose in explaining how web-based disgracing events are transmitted. This does the same thing, encouraging robotized isolation of tweets of shame from tweets of not shame. Because of the increase in the number of online communities and the amount of user-generated data, the need for effective community management increases [7].

Author [10]used machine learning to automatically identify poor user contributions using an algorithm. Comments are labeled based on whether or not there is profanity, insults, and the purpose of the insults. The use of these data is for training Support Vector Machines (SVM) and is part of a multistep process for detecting bad user contributions.

III. PROPOSED METHODOLOGY

To perform the proposed classification of problems, we formulate the task as detection and mitigation of online public disgracing side effects. The two major contributions of this study are: This project classifies and automatically classifies tweets that embarrass others. Creating a web application for Twitter users to identify Shamers 2) also Develop a web application for Twitter users to identify Shamers.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

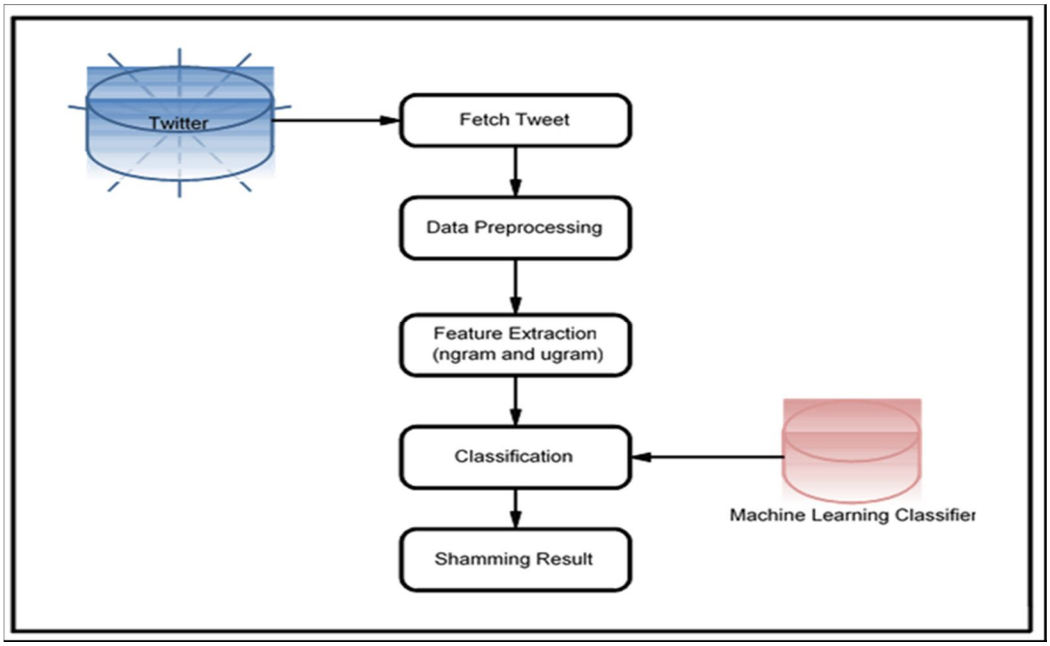


Fig. 1. System Architecture

A. Algorithm

1) Decision Tree

a) Step 1: Dataset upload

b) Step 2: The input attributes set is the text set

c) Step 3: The output attributes are set as Shaming

d) Step 4: Sample is an exercise set of data

Iterative function Dichotomiser gives back a tree of decision. Build the tree root node.If (all inputs are good, return the positive leaf node)\ If Else (when all inputs are negative, return the negative leaf node) Else (Some inputs are positive, some inputs are negative, check condition, then return result)

2) Calculate H(S) current entropy

3) Compute the entropy with regard to 'X,' which is referred to as H(S,X) for each attribute.

4) Choose a maximum IG(S,attribute. X)

5) Remove from the set of attributes the attribute with the highest value

6) Repeat until all attributes are running out, or all leaf nodes have been in the decision tree.

B. Output

The value of the data set is recovered.

IV. RESULTS AND DISCUSSION

Large numbers of real-time tweets are mined from Twitter's API. Once the nature of tweets has been defined, sentiment analysis is done. Shamming classification is finally done, and semantic analysis comes next. A summary of the evaluation metrics for each run is given in table.

	Naive Bayes	Decision Tree
Precision	66.09%	60.78%
Recall	77.15%	70.92%
F-measure	61.99%	65.03%
Accuracy	81.20%	82.21%

Table 1: Comparison with Existing system

©IJRASET: All Rights are Reserved

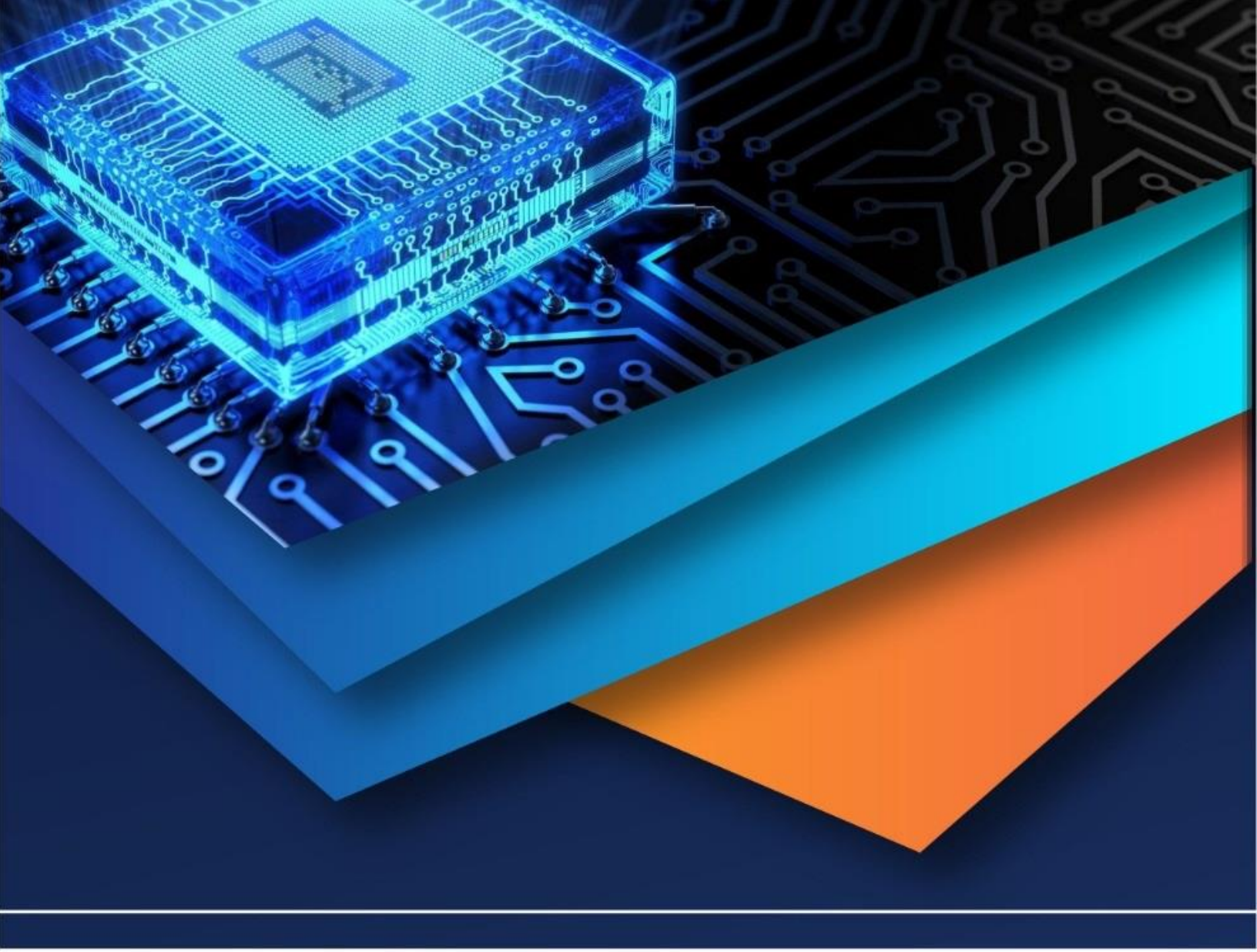
V. CONCLUSION

Bullying detection has made it possible to identify what causes shame. Words that are considered shameful, like 'failure' and 'dumb,' can be found on social media. In the past few years, shaming detection has seen an uptick in usage. This proposed system enables users to discover the number of offensive words in their data with the aid of the data and their polarity is estimated using Random Forest. What Twitter could do to minimise online public shaming is to develop a set of classifiers to detect and assign shaming comments into different categories. Moreover, we will continue to investigate new problems with respect to a social network service provider, such as Facebook or Instagram, for OSN users to enhance their well-being.

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 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] 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: 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.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)