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

Analysis and Detection of Online Public Shaming on Twitter using Bayes Classifier

Jayanthi R[1], Kousalya O[2], Priyanka S[3], Vinitha V[4], Stephie Rachel I[5]
[1, 2, 3, 4]*UG Students,* [5]*Teaching Fellow, Department of IT, University College Of Engineering, Nagercoil, Tamil Nadu, India.*

Abstract: ***Public shaming in online social networks has been increasing in recent years mainly on Twitter . These events are known to have a devastating impact on the victim's social, political, and financial life. The task of public shaming detection is automated from the perspective of victims on Twitter and explore primarily two aspects, namely, events and shamers. Shaming tweets are categorized into six types: abusive, comparison, passing judgment, religious/ethnic, sarcasm /joke, what aboutery and each tweet is classified into one of these types or as non-shaming. Finally, based on the categorization and classification of shaming tweets using Bayes classifiers, a web application called Block Shame has been designed and deployed for on-the-fly muting/blocking of shamers attacking a victim using automatic blocking algorithm.***
Keywords: Tweets, Shamers, Blocking, Classification, Analysis

I. INTRODUCTION

Online social networks (OSNs) are frequently flooded with scathing remarks against individuals or organizations on their perceived wrongdoing. When some of these remarks pertain to objective fact about the event, a sizable proportion attempts to malign the subject by passing quick judgments based on false or partially true facts. Limited scope of fact check ability coupled with the virulent nature of OSN often translates into ignominy or financial loss or both for the victim. Negative discourse in the form of hate speech, bullying, profanity, flaming, trolling, etc., in OSNs. Public shaming is condemnation of someone who is in violation of accepted social norms to arouse feeling of guilt in him or her, has not attracted much attention from a computational perspective. Nevertheless, these events are constantly being on the rise for some years.

Public shaming events have far-reaching impact on virtually every aspect of the victim's life. In public shaming, a shame is seldom repetitive as opposed to bullying. Public shaming events have far-reaching impact on virtually every aspect of the victim's life. Such events have certain distinctive characteristics that set them apart from other similar phenomena: 1) A definite single target or victim; 2) An action committed by the victim perceived to be wrong; and 3) A cascade of condemnation from the society. In public shaming, a shame is seldom repetitive as opposed to bullying. This paper looks at the problem from the victim's perspective. We consider a comment to be shaming only when it criticizes the target of the shaming event.

II. LITERATURE REVIEW

Karthik Dhinakar, Birago Jones, Catherine Havasi, Henry Lieberman and Rosalind Picard [1] studied that cyber bullying (harassment on social networks) is widely recognized as a serious social problem, especially for adolescents. It is as much a threat to the feasibility of online social networks for youth today as spam once was to email in the early days of the Internet. To promote empathy among social network participants, an air traffic control-like dashboard is proposed, which alerts to be escalated and helps prioritize the current deluge of user complaints.

Sara Owsley Sood, Judd Antin, Elizabeth F. Churchill [2] concluded that research in computer vision focuses on detection of inappropriate images, natural language processing technology has advanced to recognize insults. Through analysis of comments from a social news site, that current system is performing poorly and evaluate the cases on which they fail. They addressed community differences regarding creation/tolerance of profanity and suggest a shift to more contextually nuanced profanity detection systems.

Anna Schmidt, Michael Wiegand [3] presented a survey on hate speech detection. The steadily growing content of social media , the amount of online hate speech in networks is also increasing. Due to the massive scaling of the web contents, the automatic detection of hate speech is required. Our survey describes key areas that have been explored to automatically recognize these types of utterances using natural language processing Provides a short, comprehensive and structured overview of automatic hate speech detection, and focus on feature extraction in particular.

William Warner, Julia Hirschberg [4] presented an approach to detecting hate speech in online text, where hate speech is defined as abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation. In online social networks, hate speech against any group may exhibit some common characteristics, we have observed that hatred against each different group is typically characterized by the use of a small set of high frequency stereotypical words and the words may be used in either a positive or a negative sense, making our task similar to that of words sensing disambiguation.

Irene Kwok, Yuzhou Wang [5]studied that the social medium Twitter grants users freedom of speech, its instantaneous nature and retweeting features also amplify hate speech. Twitter has a sizeable black constituency, racist tweets against blacks are especially harming the Twitter community, though this effect may not be easily seen or understood against a backdrop of half a billion tweets a day.1 We apply a supervised machine learning approach, employing inexpensively acquired labelled data from diverse Twitter accounts to learn a binary classifier for the labels "racist" and "nonracist." The accuracy level of classifier is 76% accuracy on individual tweets, suggesting that with further improvements, the work can contribute data on the sources of anti-black hate speech.

Pete Burnap, Matthew L. Williams [6] concluded that a key contribution of is the production of a machine classifier that could be developed into a technical solution for use by policymakers as part of an existing evidence-based decision-making process. Further contributions of the paper are the identification of certain features of cyber hate on social media using a particular type of syntactic relationship within the text as a classification feature and ensemble machine classifier is applied to cyber hate. We include a section on how the classifier can be finely trained to suit the needs of policymakers, in order to minimize error and maximize confidence in results. We then demonstrate how the results of the classifier can be robustly utilized in a statistical model used to forecast the likely spread of cyber hate in a sample of Twitter data.

Zeerak Waseem, Dirk Hovy [7] provided a list of criteria founded in critical race theory, and use them to annotate a publicly available corpus of more than 16k tweets. We analyze the impact of various extra-linguistic features in conjunction with character n-grams hatespeech detection. We also present a dictionary based the most indicative words in our data.

Pinkish Badjatiya, Shashank Gupta, Manish Gupta, Vasudeva Varma [8] analyzed that hate speech detection on Twitter is critical for applications like controversial event extraction, building AI chatterbots, content recommendation, and sentiment analysis. The task is to classify a tweet as racist, sexist or neither. The complexity of the natural language constructs makes this task very challenging.Extensive experiments with multiple deep learning architectures to learn semantic word embedding to handle this complexity are performed.

Paraskevas Tsantarliotis, Evaggelia Pitoura , Panayiotis Tsaparas [9] identified troll vulnerable posts, that is, posts that are potential targets of trolls, so as to prevent trolling before it happens. They defined three natural axioms that a troll vulnerability metric must satisfy and introduce metrics that satisfy

III. PROPOSED ARCHITECTURE

The main functional units involving automated classification of shaming tweets are shown. Both labeled training set and a test set of tweets for each of the categories go through the preprocessing and feature extraction steps. The training set is used to train six Bayesian classifiers. The precision scores of the trained classifiers are next evaluated on the test set. Based on these scores, the classifiers are arranged hierarchically. A new tweet, after preprocessing and feature extraction, is fed to the trained classifiers and is labeled with the class of the first classifier that detects it to be positive. A tweet is deemed nonshame if all the classifiers label it as negative.Fig. 1. Explains the process of analysis and detection of online public shaming on Twitter using Bayes classifier.

A. Preprocessing

Preprocessing is a necessary step in data mining.Preprocessing involves the transformation of raw data into an understandable form. A series of preprocessing steps is performed before feature extraction and classification are done. All references to victims, including names or surnames preceded by salutations, mentions, and so on, are replaced with a uniform victim marker after the dependency parsing step. We also remove user mentions, repeated character, retweet marker, hashtags, URLs and all the text is converted to lower case from the tweet text after dependency parsing.

B. Feature Extraction

Feature extraction starts from an initial set of measured data and builds derived values (features intended to be informative and non redundant, facilitating the subsequent learning and generalizing step. A variety of syntactic, semantic, and contextual features are derived from the text of a tweet. a feature is represented by an index containing a letter followed by a number. Similar features are grouped together, and they share a common letter in their indexes.

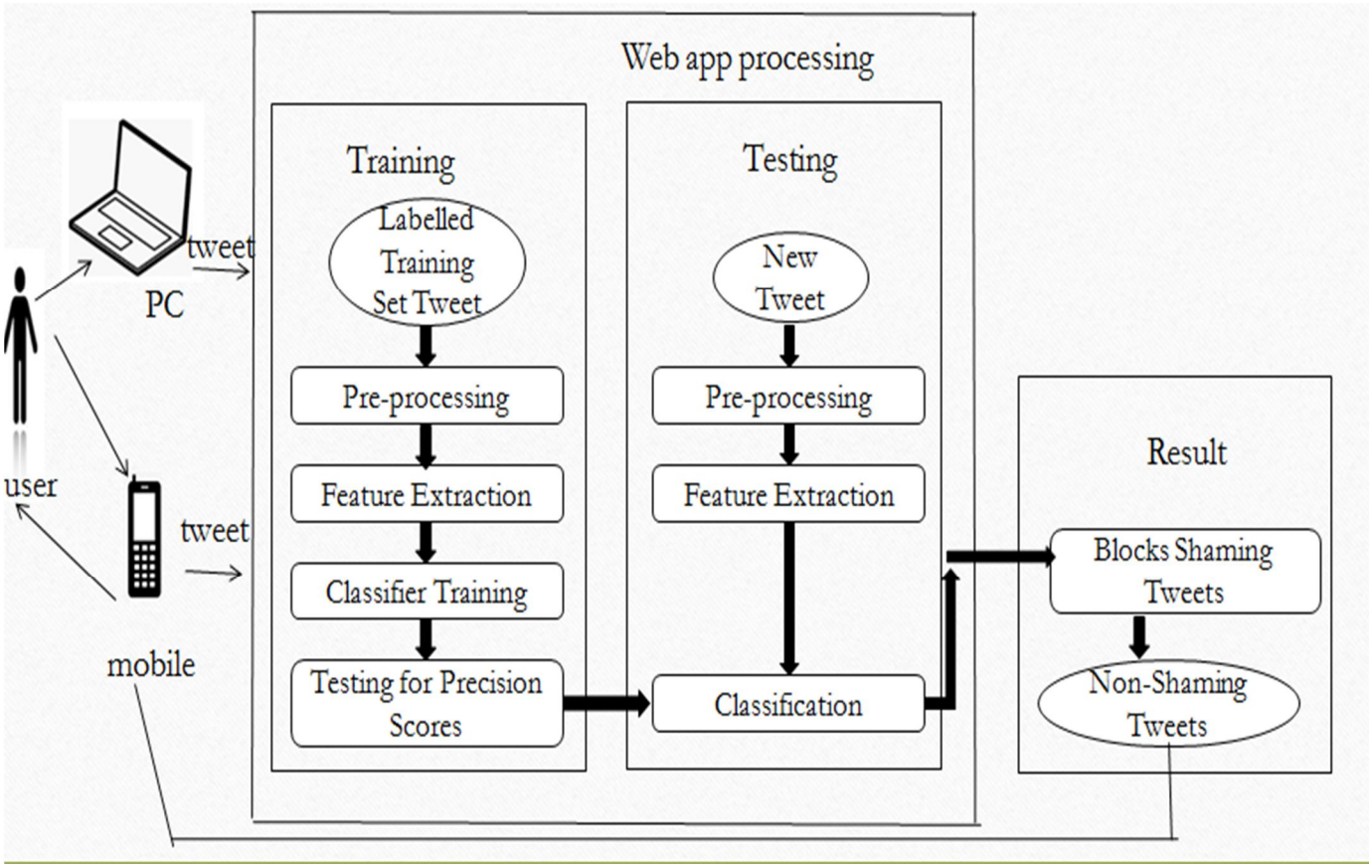


Fig. 1 Overall architecture Analysis and Detection of Online Public Shaming on Twitter Using Bayes Classifier

.

C. Classification using Bayes Classifier

Bayesian classifier is statistical classifier based on Bayes's theorem. The Bayesian classifier can predict the class membership probability such as: the probability that the given tuple belongs to a particular class. A simple Bayesian classifier known as a Naive Bayes classifier to be comparable in performance with decision tree and a neural network. All naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

While training a classifier, shaming comments from all other categories along with nonshame comments are treated as negative examples. Based on test set precision, the classifiers are arranged hierarchically placing one with higher precision above one with lower precision. The abusive classifier that has the highest precision is placed on top. Steps involved in classification of shaming tweets are

1) Build a vocabulary of all the words resident in the training data set.
2) Match tweet content against the vocabulary-word by word.
3) Build the word feature vector.
4) Plug the feature vector into the Naive Bayes classifier.

D. Mitigation of public shaming

These measures are very effective in the sense that global actions can be taken by Twitter like deleting the offending tweet or even suspending the account of the offender altogether. However, the main problem with this approach is that action against a reported shaming tweet or account may take time. However, there is no commitment to the actual time needed to take action against the offender. As shaming events are viral in nature, delayed action would defeat any attempt aimed at protecting the victim.

There are local controls provided by the Twitter API namely "block" that is similar to mute but it also unfollows/ unfriends the blocked account, and "delete" that deletes a direct message (DM) received by the user. Although limited in scope, these actions remove any tweet immediately from the victim's feed, thus, shielding him/her from direct shaming attacks.

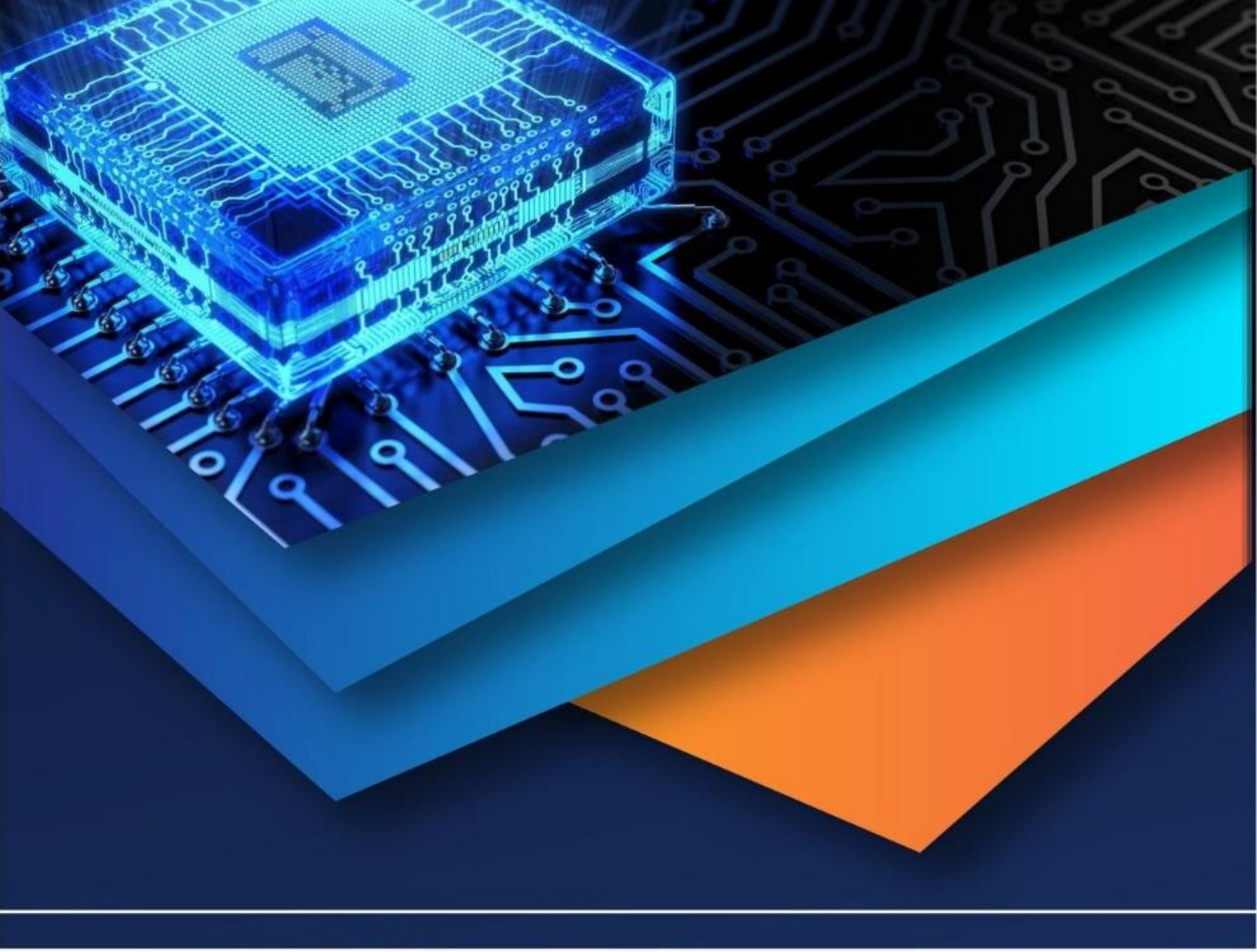
IV. CONCLUSION

In this paper, we proposed a potential solution for countering the menace of online public shaming in Twitter by categorizing shaming comments in six types, choosing appropriate features, and designing a set of classifiers to detect it.

With the growth of online social networks and a proportional rise in public shaming events, voices against callousness on part of the site owners are growing stronger. Categorization of shaming comments as presented in this paper has the potential for a user to choose to allow certain types of shaming comments (e.g., comments that are sarcastic in nature) giving his/her an opportunity for rebuttal and block others (e.g., comments that attack her ethnicity) according to individual choices. Shaming is subjective in reference to shamers. For example, the same comment made by two different persons coming from different social, cultural, or political backgrounds may have different connotations to the victim.

REFERENCES

[1] Karthik Dinakar, Birago. Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," ACM Trans. Interact. Intell. Syst., volume2, no. 3, p. 18, 2012.
[2] Sara Owsley Sood, Judd Antin, and Elizabeth F. Churchill, "Profanity use in online communities," in Proc. SIGCHI Conf. Hum. Factors Comput. Syst., 2012, p. 1481–1490
[3] Anna Schmidt and Michael Wiegand, "A survey on hate speech detection using natural language processing," in Proc. 5th Int. Workshop Natural Language Processing, Social Media Assoc. Comput. Linguistics, Valencia, Spain, 2017, p. 1–10.
[4] William Warner and Julia Hirschberg, "Detecting hate speech on the world wide Web," in Proc. 2nd Workshop Lang. Social Media, 2012, p. 19–26.
[5] Irene Kwok and Yuzhou Wang, "Locate the hate: Detecting tweets against blacks," in Proc. AAAI, 2013, p. 1621–1622.
[6] Pete Burnap and Matthew L. Williams, "Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making," Policy Internet, volume 7, no. 2, p. 223–242, 2015.
[7] Zeerak Waseem and Dirk Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in Proc. SRW HLTNAACL, 2016, p. 88–93.
[8] Pinkish Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma, "Deep learning for hate speech detection in tweets," in Proc. 26th Int. Conf World Wide Web Companion, 2017, p. 759–760,2017.
[9] ParaskevasTsantarliotis, Evaggelia Pitoura, and Panayiotis Tsaparas, "Defining and predicting troll vulnerability in online social media," Social Network Analysis & Mining, volume 7, no. 1, p. 26, 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY