iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** II **Month of publication:** February 2021

DOI: https://doi.org/10.22214/ijraset.2021.32964

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 II Feb 2021- Available at www.ijraset.com

Online Shaming via Social Media using Machine Learning: A Survey

Dige Vaishnavi A.[1], Gujar Sonali [2]
[1, 2]*Computer Engineering, Dr. Babasaheb Ambedkar Technological University, Lonere.*

Abstract: ***Social network sites involve billions of users around the world wide. User interactions with these social sites, like twitter have tremendous and occasionally undesirable impact implications for daily life. The major social networking sites have become a target platform for users to disperse a large amount of irrelevant and unwanted information. Twitter, it has become one of the most extravagant platforms of all time and, most popular micro blogging services which is generally used to share unreasonable amount of opinions. In this proposed work automate the task of public shaming detection in Twitter. Shaming tweets are categorized into nine types: abusive, comparison, passing judgment, religious, jokes on personal issues, vulgar, spam, non-spam, whataboutery and each tweet is classified into one of these types or as non-shaming. It is observed that out of all the participating users who post comments in a particular event, majority of them are likely to humiliate the victim. Interestingly, it is also the shaming whose follower counts increase faster than that of the non-shaming in Twitter.***
Keywords: Remove shammers, Online user behaviour, Tweet analysis, Public shaming, Tweet classification.

I. INTRODUCTION

It will be an online social network (OSN) defined as the use of dedicated websites applications that allow users to interact with other users or to find people with similar own interests Social networks sites allow people around the world to stay Touch each other regardless of age. The especially 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 Social networks like Facebook, LinkedIn, MySpace, Twitter are more popular lately. The offensive language detection is a processing activity of natural language that deals with find out if there are shamming (e.g. related to religion, racism, defecation, etc.) present in a given document and classify the file document accordingly. The document that will be classified in shamming word detection is in English text format that can be extracted from tweets, comments on social networks, movie reviews, political reviews, comments.

The work is divided into two parts:

A. Shaming Tweets are Categorized into Nine Types

1) Abusive
2) Comparison
3) Passing judgement
4) Religious
5) Sarcasm
6) Whataboutery
7) Vulgar
8) Spam
9) Non-spam

B. Tweet is Classified into one of These Types or as Non-shaming

Public shaming in online social networks has been increasing in recent years. These events has 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.

©IJRASET: All Rights are Reserved

II. RELATED WORK

A survey on hate speech detection. Given the steadily growing body of social media content, the amount of online hate speech is also increasing. Due to the massive scale of the web, methods that automatically detect hate speech are required. This survey describes key areas that have been explored to automatically recognize these types of utterances using natural language processing and in this paper also discuss limits of those approaches [1].

Guntur et al. [2] built a hate speech classification model using word representation with continuous bag of words (CBOW) and fast Text algorithm. This algorithm was chosen, because it is able to achieve a good performance, especially in the case of rare words by making use of character level information. Based on this result, we can see that there is no single, universal variations that outperform other. Chaya Libeskind [3] Aim to classify comments as abusive or non-abusive. In this paper develop a Hebrew corpus of user comments annotated for abusive language.

Then, we investigate highly sparse n-grams representations as well as denser character n-grams representations for comment abuse classification. Since the comments in social media are usually short, we also investigate four dimension reduction methods, which produce word vectors that collapse similar words into groups.

Mukul et al. [4] Kaggle's toxic comment dataset is used to train deep learning model and classifying the comments in following categories: toxic, severe toxic, obscene, threat, insult, and identity hate. The dataset is trained with various deep learning techniques and analyse which deep learning model is better in the comment classification. The deep learning techniques such as long short term memory cell (LSTM) with and without word Glove embeddings, a Convolution neural network (CNN) with or without Glove are used, and Glove pertained model is used for classification.

It presents an Indonesian abusive language detection system by tackling this problem as a classification task and solving it using the following classifiers: Naive Bayes, SVM, and KNN. In this paper also performed feature selection procedure based on Mutual Information value between words [5]. Aneta and Gareth [6] provide a comprehensive set of features based on users' attributes, as well as social-graph metadata. The former includes metadata about the account itself, while the latter is computed from the social graph among the sender and the receiver of each message. Attribute based features are useful to characterize user's accounts in OSN, while graph-based features can reveal the dynamics of information dissemination across the network. In particular, in this paper derive the Jaccard index as a key feature to reveal the benign or malicious nature of directed messages in Twitter. To the best of our knowledge, we are the first to propose such a similarity metric to characterize abuse in Twitter. Justin Cheng [7] Both negative mood and seeing troll posts by others significantly increases the probability of a user trolling, and together double this probability. A predictive model of trolling behaviour shows that mood and discussion context together can explain trolling behaviour better than an individual's history of trolling.

Pinkesh et al. [8] describe Hate speech detection on Twitter is critical for applications like controversial event extraction, building AI chatter bots, content recommendation, and sentiment analysis. In this paper define this task as being able to classify a tweet as racist, sexist or neither.

The complexity of the natural language constructs makes this task very challenging and this system perform extensive experiments with multiple deep learning architectures to learn semantic word embeddings to handle this complexity.

Guanjun and Surya [9] 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 viability of online social networks for youth today as spam once was to email in the early days of the Internet. Current work to tackle this problem has involved social and psychological studies on its prevalence as well as its negative effects on adolescents. While true solutions rest on teaching youth to have healthy personal relationships, few have considered innovative design of social network software as a tool for mitigating this problem. Mitigating cyber bullying involves two key components: robust techniques for effective detection and reflective user interfaces that encourage users to reflect upon their behaviour and their choices.

Hajime [10] characterize antisocial behaviour in three large online discussion communities by analysing users who were banned from these communities. In this paper find that such users tend to concentrate their efforts in a small number of threads, are more likely to post irrelevantly, and are more successful at garnering responses from other users. Studying the evolution of these users from the moment they join a community up to when they get banned, find that not only do they write worse than other users over time, but they also become increasingly less tolerated by the community. Further, in this paper discover that antisocial behaviour is exacerbated when community feedback is overly harsh. Our analysis also reveals distinct groups of users with different levels of antisocial behaviour that can change over time.

©IJRASET: All Rights are Reserved

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. These works are majorly differentiated by the algorithm for shamming detection systems.

IV. PROPOSED SYSTEM

In the existing 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:

A. Categorization and automatic classification of disgracing tweets.
B. Develop a web application for Twitter user to identify Shamers.

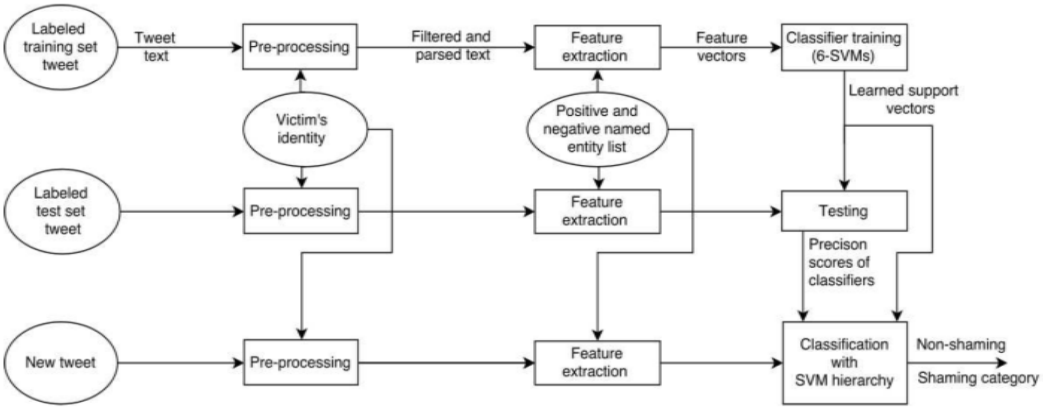


Fig. 1 A proposed block diagram for shaming detection.

The goal is classification of tweets automatically in nine categories. The main functional units are shown in fig 1. The labelled training set and test set for each category go through the pre-processing and feature extraction steps. A tweet is labelled non shame if all the classifiers label it as negative.

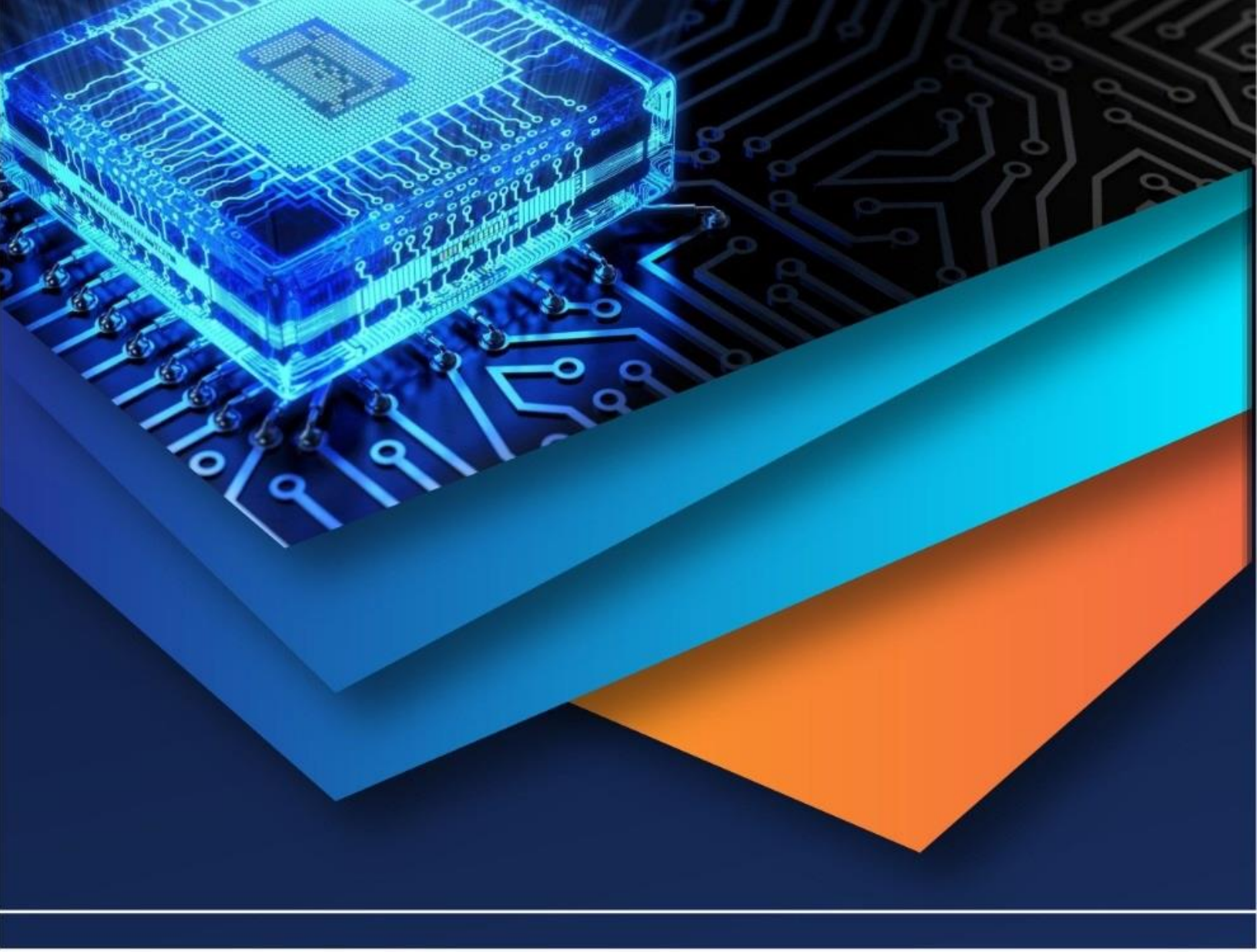
V. CONCLUSION

Shaming detection has lead to identify Shaming contents. Shaming words can be mined from social media. Shaming detection has become quite popular with its application. This system 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, APR 2019.
[2] Guntur Budi Herwanto, Annisa Maulida Ningtyas, Kurniawan Eka Nugrahaz, Nyoman Prayana Trisna" Hate Speech and Abusive Language Classification using fast Text" ISRITI 2019.
[3] Chaya Liebeskind, Shmuel Liebeskind" Identifying Abusive Comments in Hebrew Facebook" 2018 ICSEE.
[4] Mukul Anand, Dr. R. Eswari" Classification of Abusive Comments in Social Media using Deep Learning" ICCMC 2019.
[5] Dhamir Raniah Kiasati Desrul, Ade Romadhony" Abusive Language Detection on Indonesian Online.News Comments" ISRITI 2019.
[6] Alvaro Garcia-Recuero, Aneta Morawin and Gareth Tyson Trolls layer: crowd sourcing and Characterization of Abusive Birds in Twitter" SNAMS 2018
[7] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, Jure Leskovec , "Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions", ACM-2017
[8] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, Vasudeva Varma, "Deep Learning for Hate Speech Detection in Tweets", International World Wide Web Conference Committee-2017
[9] 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.
[10] 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.

©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)