|| Volume 6 || Issue 4 || April 2021 || ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH
AND ENGINEERING TRENDS

A SURVEY ON IDENTIFICATION OF ONLINE PUBLIC SHAMING USING MACHINE LEARNING FRAMEWORK

Miss. Priyanka D Nalawade [1], Prof.D.H.Kulkarni [2]

Student, Smt. Kashibai Navale College of Engineering Vadgaon, Pune[1]
Professor, Smt. Kashibai Navale College of Engineering Vadgaon, Pune[2]

Abstract: - Social network sites involve billions of users around the world wide. User interactions with these social sites, like twitter have a 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 microblogging 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 and what aboutery, 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 behavior, 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:

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

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.

II RELATED WORK

Rajesh Basak, Shamik Sural , Senior Member, IEEE, Niloy Ganguly, and Soumya K. Ghosh, Member, IEEE: This paper presents 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 author also discuss limits of those approaches.

Guntur Budi Herwanto, Annisa Maulida Ningtyas, Kurniawan Eka Nugrahaz, I Nyoman Prayana Trisna: Author built a hate speech classification model using word representation with continous bag of words (CBOW) and fastText algorithm. This algorithms was chosen, because it is able to achieve a good performance, specially 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 Liebeskind, Shmuel Liebeskind: In this study, author aim to classify comments as abusive or non-abusive. Author 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 Anand, Dr.R.Eswari: In this paper, 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 analyze 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 pretrained model is used for classification.

Dhamir Raniah Kiasati Desrul, Ade Romadhony: In this paper, author present 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. author also performed feature selection procedure based on Mutual Information value between words.

Alvaro Garcia-Recuero, Aneta Morawin and Gareth Tyson: Author 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, Author 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, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, Jure Leskovec: 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 behavior shows that mood and discussion context together can explain trolling behavior better than an individual's history of trolling.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, Vasudeva Varma: This Paper describe Hate speech detection on Twitter is critical for applications like controversial event extraction, building AI chatterbots, content recommendation, and sentiment analysis. Author 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 Lin,Sun, Surya Nepal, Jun Zhang,Yang Xiang, Senior Member, Houcine Hassan: This paper, Cyberbullying (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 cyberbullying involves two key components: robust techniques for effective detection and reflective user interfaces that encourage users to reflect upon their behavior and their choices.

HAJIME WATANABE, MONDHER BOUAZIZI , AND TOMOAKI OHTSUKI : In this paper, characterize antisocial behavior in three large online discussion communities by analyzing users who were banned from these communities. author 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, author discover that antisocial behavior is exacerbated when community feedback is overly harsh. Our analysis also reveals distinct groups of users with different levels of antisocial behavior that can change over time.

III GAP ANALYSIS

Sr No	Title	Author	Year	Description
1	Online Public Shaming on Twitter: Detection, Analysis, and Mitigation	Rajesh Basak, Shamik Sural , Niloy Ganguly, and Soumya K. Ghosh	IEEE 2019	Author proposed Shaming tweets are categorized into six types: abusive, comparison, passing judgment, religious/ethnic, sarcasm/joke, and whataboutery, and each tweet is classified into one of these types or as nonshaming using support vector machine.
2	Hate Speech and Abusive Language Classification using fastText	Guntur Budi Herwanto, Annisa Maulida Ningtyas, Kurniawan Eka Nugrahaz, I Nyoman Prayana Trisna	ISRITI 2019	Author built a hate speech classification model using word representation with continous bag of words (CBOW) and fastText algorithm. This algorithms was chosen, because it is able to achieve a good performance, specially 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.
3	Identifying Abusive Comments in Hebrew Facebook	Chaya Liebeskind, Shmuel Liebeskind	2018 ICSEE	In this study, author aim to classify comments as abusive or non-abusive. Author 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.
4	Classification of Abusive Comments in Social Media using Deep Learning	Mukul Anand, Dr.R.Eswari	ICCMC 2019	In this paper, 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 analyze 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 pretrained model is used for classification

5	Abusive Language Detection on Indonesian Online News Comments	Dhamir Raniah Kiasati Desrul, Ade Romadhony	ISRITI 2019	In this paper, author present 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. author also performed feature selection procedure based on Mutual Information value between words.
6	Trollslayer: Crowdsourcing and Characterization of Abusive Birds in Twitter	Alvaro Garcia-Recuero, Aneta Morawin and Gareth Tyson	SNAMS 2018	Author 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, Author 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.
7	Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions	Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, Jure Leskovec	ACM-2017	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 behavior shows that mood and discussion context together can explain trolling behavior better than an individual's history of trolling.
8	Deep Learning for Hate Speech Detection in Tweets	Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, Vasudeva Varma	International World Wide Web Conference Committee -2017	Hate speech detection on Twitter is critical for applications like controversial event extraction, content recommendation and sentiment analysis. Task to classify a tweet as racist, sexist or neither. The complexity of the natural language constructs makes this task very challenging.
9	Statistical Twitter Spam Detection Demystified: Performance, Stability and Scalability	Guanjun Lin,Sun, Surya Nepal, Jun Zhang,Yang Xiang, Senior Member, Houcine Hassan	IEEE TRANSACTIONS 2017	Due to the popularity of online social networks, cyber criminals are spamming on these platforms for potential victims. In this paper, performance of a wide range of mainstream machine learning algorithms are compared, aiming to identify the ones offering satisfactory detection performance and stability based on a large amount of ground truth data.

10	Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection	HAJIME WATANABE, MONDHER BOUAZIZI , AND TOMOAKI OHTSUKI	Digital Object Identifier – 2017	Hate speech refers to the use of aggressive, violent or offensive language, targeting a specific group of people sharing a common property, whether this property is their gender, their ethnic group or race or their believes and religion. Ternary classification of tweets into, hateful, offensive and clean.

IV 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.

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, I Nyoman Prayana Trisna" Hate Speech and Abusive Language Classification using fastText" 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" Trollslayer: Crowdsourcing 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