IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710


# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 9, Issue 12, December 2020

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 12, December 2020 ||

# Survey on Shaming Sentence Detection on Social Network

**Tailor Karan Rajesh[1], Navale Pradnya Pandurang [2], Bhatt Het Jaymin[3], Nehul Rutuja Vasant [4], Dhakulkar Bhagyashree [5]**

UG. Student, Dept. of Computer Engineering, Dr. D. Y. Patil School of Engineering and Technology, Pune, India. [1,2,3,4]

Professor, Dept. of Computer Engineering, Dr. D. Y. Patil School of Engineering and Technology, Pune, India. [5]

**ABSTRACT:** Twitter is an online social networking service that has more than 300 million users, generating a huge amount of information each day. Twitter's most important feature is its ability for users to tweet about events, situations, feelings, opinions, or even something new in real-time. There is no system of accuracy or reliability in place: Anyone can say just anything. It can be an easy way to attack your detractors for them to attack, the sort of Twitter war. In this survey, various applications of machine learning and hate speech detection to ease the detection of shammers and shamming tweets was included. With the increase of online social networks, and growth in public shaming events, voices against callousness of the site owners are growing stronger, there is a need to analyze the shaming tweets, classify shamming tweets into different categories, and to mitigate it by blocking them.

**KEYWORDS***:* Shaming, Social Network, OSN network, Performance metrics.

## I. INTRODUCTION

The limited knowledge of facts with the toxic nature of OSNs often translates into ignominy or financial loss or both for the victim. Unenthusiastic speech in the form such as hate speech, bullying, profanity, flaming, trolling, etc., in OSNs, is well studied in the literature. On the other hand, public shaming, which is the condemnation of someone who violates accepted social norms to arouse feelings of guilt in him or her, has not attracted much attention from a computational perspective. Nevertheless, these events are constantly being increasing for some years.

The immense volume of comments which is often used to shame an almost unknown victim speaks of the viral nature of such events. For example, when Justine Sacco, a public relations person for American Internet Company tweeted “Going to Africa. Hope I don’t get AIDS. Just kidding. I’m white!” she had just 170 followers. Soon, a barrage of criticisms started pouring in, and the incident became one of the most talked-about topics on Twitter and the Internet, in general, within hours. She lost her job even before she landed in South Africa. Jon Ronson’s “So You’ve Been Publicly Shamed” presents an account of many online public shaming victims. The observation that the author made from these diverse set of events about the victims that are subjected to punishments disproportionate to the level of crime they have committed. They have also formed a list of victims, the year in which the event happened, the action that triggered public shaming along with the triggering medium, and its immediate consequences for each studied event.

The trigger is the first action or word spoken by the victim responsible for initiating public shaming. “Medium of triggering” is the first communication media through which the general public became aware of the “Trigger.” The consequences for the victim, during or shortly after the event, are listed in “Immediate consequences.” Henceforth, the two-letter abbreviations of the victim’s name will be used to refer to the respective shaming event.

The author proposed a system for automating the task of shaming tweet detection from the perspective of victims and exploring two major aspects which are events and shamers. Further, the shaming tweets are categorized into six types namely abusive, comparison, passing judgment, religious/ethnic, sarcasm/joke, and whataboutery, and each tweet is classified into one of these types or as non-shaming. It is discovered that out of all the participating users who post comments in a particular shaming event, the majority of them are likely to shame the victim. Also, the count will increase faster for shamers than that of non-shamers.

## II. LITERATURE REVIEW

Deep Learning for Hate Speech Detection in Tweets: In this paper [1], the Author investigated the application of deep neural network architectures for the task of hate speech detection. The author found them to remarkably outperform the existing methods, which were able to classify a tweet as sexiest, racist, or neither. Embedded learning

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 12, December 2020 ||

from deep neural network models when merged with gradient boosted decision trees led to best accuracy values. The author, performed extensive experiments to learn complexity handling of semantic word embedding's with multiple deep learning architectures. In the future, The Author plans to explore the significance of the user network features for the task.

Smokey: Automatic Recognition of Hostile Messages This paper [2], describes some approaches to flame recognition, including a prototype system, Smokey. Author combined natural-language processing and sociolinguistic observation for identifying messages containing insulting words used in an insulting manner. Smokey builds a 47-clement feature vector based on the syntax and the semantics of every sentence, combining the vectors for the sentences within each message. A training set of 720 messages was used by Quinlans C4.5 decision-tree generator to determine feature-based rules that Author's ability to correctly sort 64% of the flames and 98% of the non-flames in another test set of 460 messages. Further techniques for greater accuracy and user customization are also discussed.

Design of a Scalable Data Stream Channel for Big Data Processing [3], In this paper author, tried to outline big data infrastructure for processing a huge amount of data streams. The project is based on a distributed stream computing platform. It provides cost-effective and large-scale big data services by developing a data stream management system. The author focused on scalable data stream channel, big data processing, big data infrastructure, data streaming process, and management. This research contributes to advancing the feasibility of massive processing for distributed, real-time computation even once they are overloaded.

Developing a Real-time Data Analytics Framework for Twitter Streaming Data [4], The proposed framework includes data ingestion, stream processing, and data visualization components with the Apache Kafka messaging system that is used to perform data ingestion tasks. Furthermore, Spark makes it feasible to perform sophisticated processing and machine learning algorithms in real-time. The key focus of the paper is on developing an analytical framework with the ability of in-memory processing to extract and analyze structured and unstructured Twitter data. It has even conducted a case study on tweets about the event of the Japan earthquake and how the people around the world reacted to it by analyzing the time and origin of tweets.

Detection of Behavior Patterns through Social Networks like Twitter [5], using Data Mining techniques as a method to detect Cyberbullying This research focuses on the detection and analysis of cyberbullying on pages and with pejorative terms in Spanish, taking profit of the author of classification of feelings through specialized tools. For the detection of cyberbullying, first, the efficiency of classification of every tool is measured, through a group of pejorative terms commonly wont to hurt people. In the analysis stage author used data mining techniques to generate a dictionary of pejorative terms that are related to cyberbullying and was thus be able to generate behavior patterns of these terms. This way provided better tools so that psychology specialists can optimize their work. The final result shows which platform is more flexible, also show which is best suited to the search of incidences of cyberbullying on Twitter.

A study on hate speech detection using natural language preparing: Shared Real-Time Sentiment Analysis for Big Data Social, Streams author performed a genuine test with ongoing stream information handling and found that it is difficult to store occasions of information, and consequently online logical calculations are used. To perform a constant examination, pre-handling of information should be acted such that a stand-alone outline of the stream is put away in the primary memory [6].

Detection of Cyberbullying Incidents on the Instagram Social, Network this paper makes the following major contributions: an appropriate definition of cyberbullying that incorporates both frequencies of negativity and imbalance power is applied in largescale labeling, and is differentiated from cyber aggression; cyberbullying is studied in the context of a media-based social network, incorporating both images and comments in the labeling; a detailed analysis of the distribution results of the labeling of cyberbullying incidents is presented, including a correlation analysis of cyberbullying with other factors derived from images, text comments, and social network metadata; multi-modal classification results are presented that comes with a spread of features to spot cyberbullying incidents[7].

Defining cyberbullying: A qualitative research into the understanding of youngsters [8] Data from 53 focus groups, which involved students from 10 to 18 years old, show that youngsters often interpret cyberbullying as Internet bullying and associate the phenomenon with a wide range of enactments. To be considered true cyberbullying, this enactment must meet several criteria. They should be intended to hurt (by the perpetrator) and recognized as hurtful (by the victim); be part of a repetitive pattern of negative offline or online actions; and be accomplished in a relationship described by a power imbalance (based on real-life power criteria, such as physical strength or age, and/or on ICT-related criteria like technological) know-how and anonymity.

Sarcasm detection on Twitter: A behavioral modeling approach [9] This paper aims to address a much critical task of sarcasm detection on Twitter by leveraging behavioral traits intrinsic to users expressing sarcasm. With this intentional uncertainty, sarcasm detection has always been a challenging task, even for humans. Author approaches to automatic sarcasm detection rely primarily on lexical and linguistic cues. The author identifies such traits using the users' past authored. The author course theories from behavioral and psychological studies to build a behavioral

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 12, December 2020 ||

modeling framework tuned for detecting sarcasm. The author approximates our framework and demonstrates its efficiency in identifying sarcastic authors.

Look before you shame: A study on shaming activities on Twitter [10] As online social networks (OSNs) are often flooded with scathing remarks against individuals or businesses on their perceived wrongdoing. This paper studies three such events to get insight into different aspects of shaming done through Twitter. An important contribution of our work is the categorization of shaming authors, which helps in understanding the dynamics of the spread of online shaming events. It also makes easier automated segregation of shaming authors from non-shaming ones.

## III. CONCLUSION

Overall, the application is a need of the hour as the increase of online social networks, and growth in public shaming events, voices against callousness on part of the site owners are growing stronger. This paper represents an analysis of various application of machine learning and hate speech detection to ease the detection of shammers and shamming tweets. After going through the literature survey a new system can be proposed which is capable of categorizing shaming comments into different categories and, if a tweet has been detected as shaming, particular tweets will be blocked.

## REFERENCES

[1] J. Ronson, So You've Been Publicly Shamed. London, U.K.: Picador, 2015.
[2] E. Spertus, "Smokey: Automatic recognition of hostile messages," in Proc. AAAI/IAAI, 1997, pp. 1058–1065.
[3] S. Sood, J. Antin, and E. Churchill, "Profanity use in online communities," in Proc. SIGCHI Conf. Hum. Factors Comput. Syst., 2012, pp. 1481–1490.
[4] S. Rojas-Galeano, "On obstructing obscenity obfuscation," ACM Trans. web, vol. 11, no. 2, p. 12, 2017.
[5] E. Wulczyn, N. Thain, and L. Dixon, "Ex Machina: Personal attacks are seen at scale," in Proc. 26th Int. Conf. World Wide web, 2017, pp. 1391–1399.
[6] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in Proc. 5th Int. Workshop Natural Lang. Process. Social Media Assoc. Comput. Linguistics, Valencia, Spain, 2017, pp. 1–10.
[7]Hate-Speech. Oxford Dictionaries. Accessed: Aug. 30, 2017. [Online]. Available: https://en.oxforddictionaries.com/definition/hate_speech
[8] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in Proc. 2nd Workshop Lang. Social Media, 2012, pp. 19–26.
[9] I. Kwok and Y. Wang, "Locate the hate: Detecting tweets against blacks," in Proc. AAAI, 2013, pp. 1621–1622.
[10] P. Burnap and M. L. Williams, "Cyberhate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making," Policy Internet, vol. , no. 2, pp. 223–242, 2015.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details