IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 12, December 2024

DOI: 10.17148/IJARCCE.2024.131241

# ELECTRONIC COMMUNITY HATRED ON TWITTER: DETECTION, ANALYSIS, AND DIMINUTION

**Aditya Singh[1], Sankalp Pandey[2]**

Computer science and engineering, SRM IST, Ghaziabad, Uttar Pradesh, India[1]

Computer science and engineering, SRM IST, Ghaziabad, Uttar Pradesh, India[2]

**Abstract:** There has been a spike in digital bullying in online groups and similar online public platforms such as Twitter in recent years. These incidents have been seen to have a negative impact on the victim's social, democratic, and economic well-being. Despite its well-documented adverse effects, leading online communities have done little to fix it, citing the sheer size and diversity of such comments and, as a result, the impractical number of human moderators required to achieve the task. we develop this automated digital bullying in online group identification on Twitter from the standpoint of the suspects, focusing on two factors- accidents and prudes. Bullying tweets are ordered into 5 types- offensive language, abusive language, ethnic, sarcasm and neither, In addition, each tweet is labeled as one of these kinds of non-shaming. Our objective is to automatically categorize tweets into the five types listed above. For each of the types, the data cleaning and feature-based steps are applied to both the named training set and the evaluation set of tweets. Finally, a web application for muting shamers attacking a victim on Twitter was designed and implemented based on the categorization and identification of bullying tweets.

## I. INTRODUCTION

ONLINE social networks (OSNs) are often filled with venomous comments directed at people or organizations for alleged misconduct. When any of these comments are focused on factual evidence about the case, a significant number of them seek to discredit the topic by making snap decisions based on misleading or partly truthful facts. The victim's ignominy or financial ruin, or both, is often the result of the victim's limited fact checkability combined with the virulent essence of OSNs. Hate speech, racism, profanity, flaming, trolling, and other forms of negative rhetoric in online social networks have also been researched extensively. On the other hand, from a computational standpoint, public shaming, which is the rejection of someone who violates agreed social norms to arouse feelings of shame in that person, has received little notice. Nonetheless, for many years, these incidents have been on the rise. The effects of public humiliation incidents can be felt in about any part of a victim's life. These cases have three distinct features that distinguish them from all related phenomena: 1) a single identifiable object or perpetrator; 2) an action taken by the victim that is considered to be wrong; and 3) a chain of societal disapproval. In contrast to bullying, a shamer is rarely repetitive when it comes to public humiliation. Hate speech and profanity are often part of a bullying case, but there are other types of shaming, such as sarcasm and jokes, comparisons of the perpetrator to other people, and so on, that do not contain specifically censored material. The massive number of remarks that are often used to ridicule an almost unknown survivor demonstrates the public significance of such incidents. When Justine Sacco, a public relations representative for American Internet Company, tweeted, "Going to Africa. I'm hoping to avoid contracting AIDS. Only joking. She only had 170 followers when she posted, "I'm blonde!" Within hours, the incident had been one of the most thought-about subjects on Twitter and the Internet in general, including a barrage of critiques. Even before her plane touched down in South Africa, she had lost her employment. The song "So You've Been Publicly Shamed" [1] by Jon Ronson tells the story of many people who have been publicly shamed online. What all of the shaming cases we've looked at have in common is that the victims are exposed to sentences that are disproportionate to the level of crime they've allegedly performed. For each studied case, we've identified the survivor, the year the event occurred, the behavior that caused public humiliation, as well as the triggering medium and its immediate effects. The "Victim's" action or words that triggered public humiliation are referred to as the "Trigger." The first contact medium by which the general public became aware of the "Cause" was the "Medium of Triggering." "Immediate effects" lists the consequences for the perpetrator that occur after or immediately after the incident. The two-letter abbreviations of the victim's name can also be used to refer to the shaming incident in question. A study on this subject has previously been conducted from the viewpoint of administrators who wish to root out any material that is deemed malicious according to their website policies (see [2]–[5]). None of these, though, recognizes a single survivor. On the opposite, we approach the issue from the victim's point of view.

  

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 12, December 2024

DOI: 10.17148/IJARCCE.2024.131241

And where a tweet criticizes the subject of the shaming case is it considered shaming? Although a message like "Justine Sacco going to get off the international flight and weep mountain stream fresh white moan tears b" is shaming, one like "Just read the Justine Sacco story lol smh sucks that she got fired for a funny tweet" is not. People are ridiculously sensitive." this is not an indication of shaming from Justine Sacco's view (despite the censored words) because it criticizes others rather than her.

We suggest a mechanism for detecting and mitigating the negative consequences of online public shaming in this article. In this article, we present three major contributions: 1) categorization and automated classification of shaming tweets; 2) offer insights into shaming incidents and shamers; and 3) plan and create a novel program that a Twitter user can use to block shamers.

## II. PROPOSED APPROACH

1. Data

We produced a dataset by combining three different datasets. The first dataset, which was edited and used in [13, 14], is publicly accessible on Crowdflower1.

This series of tweets has been categorized using one of the following instructions: "Hateful," "Offensive," or "Clean." The second dataset, comprised of tweets with the same instructions as the first, is also available on Crowdflower2.

This table has two columns: tweet-ID and class.

2. DATA PROCESSING

We combine the three data sets used in this study during the records pre-processing point. The responsibilities include removing useless columns from datasets and enumerating the classes. We retrieve the tweets referring to the tweet-ID gift inside the dataset for the 1/3 dataset. We convert the tweets to lowercase and exclude the following irrelevant content.

- Space Trends
- Twitter-Mention
- Retweet-Symbol
- Stop-words

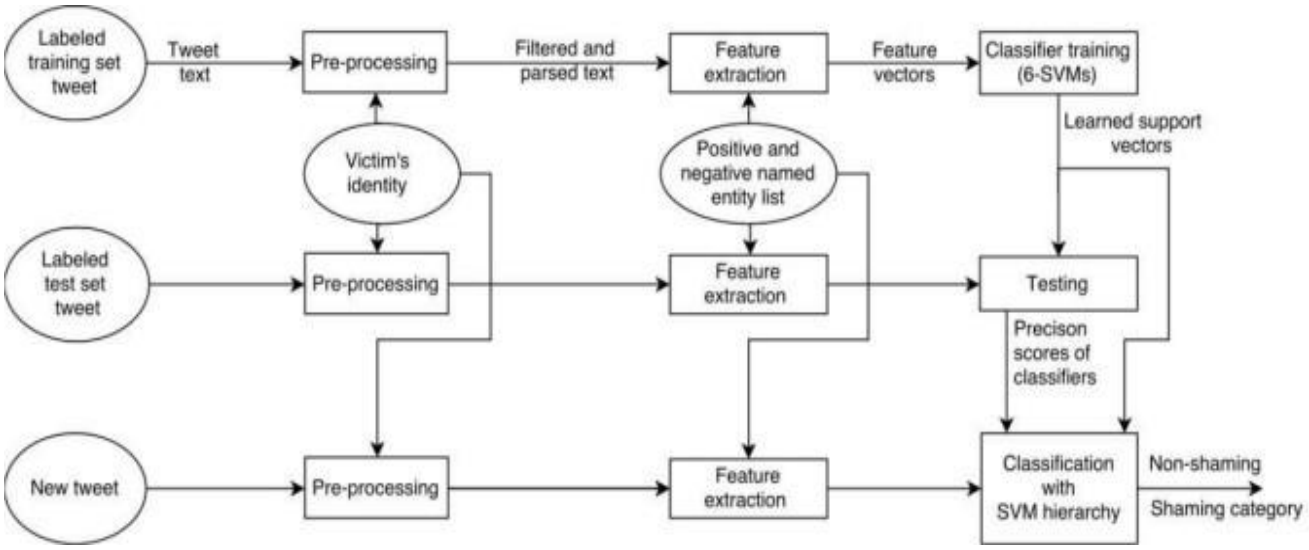


Fig. 1. Block diagram for shaming detection

To minimize the inflectional style of the words, we use the Porter-Stemmer set of rules. We arbitrarily shuffle and break up the dataset into bits after combining it in the correct format: teach the data set that includes 70% of the samples and look at a data set that includes 30% of the samples.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 12, December 2024

DOI: 10.17148/IJARCCE.2024.131241

3. Classification Of Bullying Tweets

We created five categories of tweets after examining over a thousand awkward tweets from eight embarrassing incidents on Twitter.

(a) Abusive Language (AB):

When the victim is humiliated by the shamer, a comment falls into this category. It should be noted that the existence of a list of swear language isn't adequate to establish abusive bullying. This is because a comment could contain offensive expressions but also be in the direction of the victim. However, offensive sentences about the victim discovered by dependency parsing of the remark are a strong indicator of this kind of bullying.

(b) Comparative analysis (CO)
The alleged accused's actions are compared and contrasted with that of another person in this type of bullying. The most important role here is to detect the concept of the person mentioned in the remark regularly to determine whether or not the evaluation is an example of bullying. The textual material may be lacking in clues, such as phrases with duality linked to the object. In these kinds of situations, the author of the remarks relies on the common memory of the public group's users to have the necessary meaning. This is so more often these days as the stated individual is included in various events.

(c) Passing-Judgment (PJ):
Fast decisions vilifying the sufferer may be skipped by sluts. Casting aspersions also crosses over into other groups. Only where a joke does not fall under one of the different groups is it PJ shaming. Casting aspersions frequently starts with a noun and progresses to using modal auxiliary verbs.

(d) Ethnic (RE):
There are frequently several companies with which a person knows. We consider three types of sufferer identities: citizenship (Indian, Chinese), skin color (black, white), and – anti (Christian, Jewish). – anti shaming occurs where one of these institutional identities is associated with the victim.

(e) Sarcasm:
In the Oxford Learner's Dictionary, sarcasm is described as "a nature of using statements that may be the polar opposite of what one way to be able to be offensive to another or to make humorous of them." This term is also used in a few more recent studies on semantic parsing in Twitter, such as [18].

(f) Sophistry (WA):
In sophistry, the shamers draw attention to the suspect's alleged subterfuge by claiming in prior activity in a situation similar to the current one. The use of WH adjectives and even beyond forms of verbs are significant determinants for certain types of input. It's worth noting that a model of this look was shown as a poster paper during progress. [19].

## III. EXPERIMENTAL RESULTS

The Twitter 1 percent tube, Twitter Seek API, and Topsy API were used to collect a large number of tweets from a variety of bullying incidents that unfolded over many years (defunct at present). This was labeled by a group of evaluators who had been instructed to mark each tweet in one of the five shaming groups or as non-shaming (NS).

Figure 2 shows the total number of bullying incidents.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 12, December 2024

DOI: 10.17148/IJARCCE.2024.131241

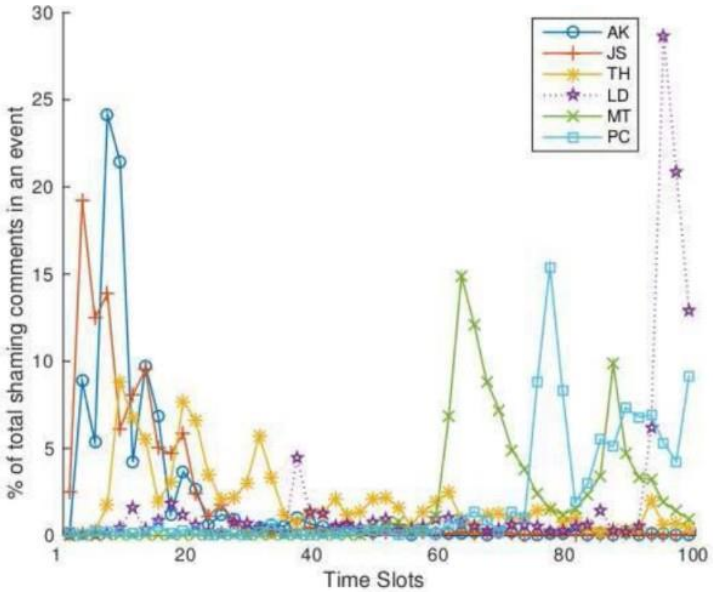


Fig.2. Composition of bullying messages with time.

From the table, '#Labelled' is the number of tweets that have been individually classified for each case. Notice that we no longer have any annotated data for the events. The number of precise tweets for an incident is referred to as "#Innovative tweets." Since a retweet is assigned the mark of the unique post, we don't have retweets specifically inside the dataset. Our most recent version is set up to interact with Twitter through the Twitter API, specifically to collect data tweets via the Twitter REST API. Tweepy, a library written in Python, makes uploading this functionality easily. APIs for Twitter, except primary data just like the tweet textual content, and consequently the creator of the tweet returns association consists of adding the data which can be wont to offer similar analysis. For every one hundred forty-person tweets, the System retrieves a JSON record with various metadata artifacts supplied as core and price sets, including identification notifications and textual information.

For this report, content is extremely important.
We also develop applications that work as a bridge between the user and Twitter. The software's configuration is seen in Figure 3. With the help of our module, we're capable of filtering out awful and insulting tweets published by using a person, as well as classifying tweets published on the consumer's domestic account, with the most straightforward issue being a 15-minute Twitter research request charge limiter.

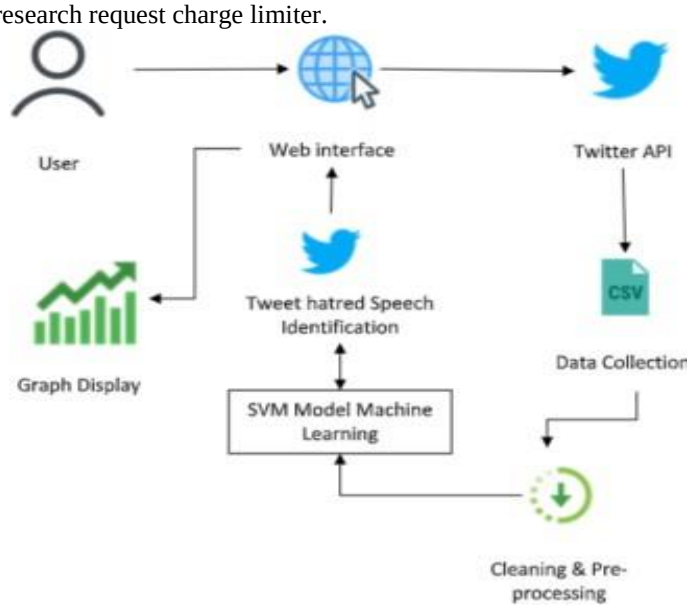


Fig.3. The system's architecture for interacting with Twitter via Twitter API.

  

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 12, December 2024

DOI: 10.17148/IJARCCE.2024.131241

## IV. CONCLUSION

In this journal, we suggest a machine-learning approach for detecting hateful speech and offensive language on Twitter using n-gram functions weighted with SVM values. We performed a comparative study of the Regression Model,On a variety of units of function values and version hyperparameters, Naive Bayes and SVM were used. For the L2 normalization of TFIDF, the results showed that the Regression model works better for the most relevant n-gram variety 1 to three.

We found 88.6 accuracies when comparing the version on check results. It was discovered that 4.8 percent of derogatory tweets were incorrectly labeled as hateful. This issue may be resolved by obtaining more samples of derogatory words that do not contain hateful expressions. The results can also be improved by increasing the remembering of the offensive elegance and accuracy of the hateful elegance. Also, it became clear that the edition no further takes into consideration derogatory terms in a statement. Incorporating linguistic functions may contribute to improvement in this area.

## REFERENCES

[1]. Zephoria.com, 2018. [Online]. Available: https://zephoria.com/top-15-valuable-facebook-statistics/. [Accessed: 22-Jun- 2018].

[2]. “Twitter Usage Statistics - Internet Live Stats”, Internetlivestats.com, 2018. [Online]. Available: http://www.internetlivestats.com/twitterstatis tics/. [Accessed: 22- Jun- 2018].

[3]. S. Hinduja and J. Patchin, “Bullying, Cyberbullying, and Suicide”, Archives of Suicide Research, vol. 14, no. 3, pp. 206-221, 2010.

[4]. H. Watanabe, M. Bouazizi and T.Ohtsuki, “Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection”, IEEE Access, vol. 6, pp. 13825-13835, 2018.

[5]. T. Davidson, D. Warmsley, M. Macy and I. Weber, “Automated Hate Speech Detection and the Problem of Offensive Language”, in International AAAI Conference on Web and Social Media, 2017.

[6]. S. Liu and T. Forss, “New classification models for detecting Hate and Violence web content,” 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), Lisbon, 2015, pp. 487-495.

[7]. P. Burnap and M. Williams, “Us and them: identifying cyber hate on Twitter across multiple protected characteristics”, EPJ Data Science, vol. 5, no. 1, 2016.

[8]. C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad and Y. Chang, “Abusive Language Detection in Online User Content”, Proceedings of the 25th International Conference on World Wide Web - WWW ’16, 2016.

[9]. E. Greevy and A. Smeaton, “Classifying racist texts using a support vector machine”, Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR ’04, 2004.

[10]. Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety”, 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, 2012.

[11]. D. Blei, A. Ng, M. Jordan, and J. Lafferty, “Latent dirichlet allocation”, Journal of Machine Learning Research, vol. 3, p. 2003, 2003.

[12]. D. Yin, Z. Xue, L. Hong and B. Davison, “Detection of harassment on Web 2.0,” in the Content Analysis in the Web 2.0 Workshop, 2009.Zephoria.com, 2018. [Online].Available:https://zephoria.com/top- 15- valuable-facebook-statistics/. [Accessed: 22- Jun- 2018].

[13]. - H. Watanabe, M. Bouazizi, and T. Ohtsuki, “Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions andPerform Hate Speech Detection”, IEEE Access, vol. 6, pp. 13825-13835, 2018.

[14]. - T. Davidson, D. Warmsley, M. Macy and I. Weber, “Automated HateSpeech Detection and the Problem of Offensive Language”, in International AAAI Conference on Web and Social Media, 2017.

[15]. - A. Rajadesingan, R. Zafarani, and H. Liu, "Sarcasm detection on Twitter: A behavioral modeling approach,” in Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. ACM, 2015, pp. 97–106.

[16]. R. Basak, N. Ganguly, S. Sural, and S.K. Ghosh, “Look before you shame: A study on shaming activities on Twitter," in Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee, 2016, pp. 11–12.

[17]. J. Ronson, So You’ve Been Publicly Shamed. London, U.K.: Picador, 2015.

[18]. A. 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

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 12, December 2024

DOI: 10.17148/IJARCCE.2024.131241

[19]. R. Basak, N. Ganguly, S. Sural, and S. K. Ghosh, "Look before you shame: A study on shaming activities on Twitter," in Proc. 25th Int. Conf. Companion World Wide Web, 2016, pp. 11–12.

[20]. C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, pp. 27:1–27:27, 2011.

[21]. E. Colleoni, A. Rozza, and A. Arvidsson, "Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data," J. Commun., vol. 64, no. 2, pp. 317–332, 2014.

[22]. M. Conover, J. Ratkiewicz, M. R. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, "Political polarization on Twitter," in Proc. ICWSM, vol. 133, 2011, pp. 89–96.

[23]. S. Hong and S. H. Kim, "Political polarization on Twitter: Implications for the use of social media in digital governments," Government Inf. Quart., vol. 33, no. 4, pp. 777–782, 2016.

[24]. Twitter. Report Abusing Behavior. Accessed: Feb. 7, 2018. [Online]. Available: https://help.twitter.com/en/safety-and-securit y/report-abusivebehavior

[25]. Blockshame Shields you from the Online Mob Just in Case! Accessed: Feb. 7, 2018.[Online].Available:http://cse.iitkgp.ac.i n/~rajesh. Basak/blockshame