TWITTER: ONLINE PUBLIC SHAMING ON ONLINE SOCIAL NETWORKS (OSNs)

ARCHANA CHALLA, Lecturer, Department of Computer Science, SriDurgaMalleswara Siddhartha MahilaKalasala, Vijayawada, Andhra Pradesh.

SUDHIR KONERU, Lecturer, Department of Computer Science, P. B. Siddhartha College of Arts and Science,Vijayawada, Andhra Pradesh.

ABSTRACT:

In this paper, we automate the task of public shaming detection in Twitter from the perspective of victims and explore primarily two aspects, namely, events and shamers. Shaming tweets are categorized into six types- abusive, comparison, passing judgment, religious/ethnic, sarcasm/joke 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 shaming event, majority of them are likely to shame the victim. Interestingly, it is also the shamers whose follower counts increase faster than that of the non-shamers in Twitter. Finally, based on categorization and classification of shaming tweets, an web application called BlockShame has been designed and deployed for on-the-fly muting/blocking of shamers attacking a victim on the Twitter.

KEYWORDS: Twitter, public shaming, victims.

1] INTRODUCTION:

Public shaming in online social networks and related online public forums like Twitter has been increasing in recent years. These events are known to have devastating impact on the victim's social, political and financial life. Notwithstanding its known ill effects, little has been done in popular online social media to remedy this, often by the excuse of large volume and diversity of such comments and therefore unfeasible number of human moderators required to achieve the task. 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 OSNs 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 is well studied in the literature. On the other hand, public shaming, which 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

 DOI : 10.36893.DRSR.2021.V08I14.101-106

far reaching impact on virtually every aspect of victim's life. Such events have certain distinctive characteristics that set them apart from other similar phenomena- (a) a definite single target or victim (b) an action committed by the victim perceived to be wrong (c) a cascade of condemnation from the society. In public shaming, a shamer is seldom repetitive as opposed to bullying. Hate speech and profanity are sometimes part of a shaming event but there are nuanced forms of shaming such as sarcasm and jokes, comparison of the victim with some other persons, etc., which may not contain censored content explicitly.

2] LITERATURE SURVEY:

Sood et al. [1] examine the effectiveness of list based profanity detection for Yahoo! Buzz comments. Relatively low F1 score (harmonic mean of precision and recall) of this approach is attributed to distortion of profane words with special characters (e.g., @ss) or spelling mistakes and low coverage of list words. The first caveat was partly overcome by considering words as abusive whose edit distance from a known abusive word equals the number of "punctuation marks" present in the word.

Galeano [2] solves the problem of intentional distortion of abusive words in order to avoid censorship by allowing homo-glyph (characters which are similar in appearance, e.g., 'a' and 'a') substitution to bear zero penalty in calculating edit distance between an abusive word and a distorted word, thereby increasing recall rate substantially. Hate speech, though well defined as- "Abusive or threatening speech or writing that expresses prejudice against a particular group, especially on the basis of race, religion, or sexual orientation" [7], is often used in several other connotations (e.g., in [6]). Warner and Hirschberg [8] attempt to identify hate speech targeting Jews from a data set consisting of Yahoo!

Dinakar et al. [3] employ Open Mind Common Sense (OMCS) [20], a common sense knowledge database, with custom built assertions related to specific domain of interests, e.g., LGBT cyberbullying, to detect comments which deviate from real world beliefs and is a good indicator of subtler forms of bullying. For instance, asking a male which beauty saloon he visits can be a case of bullying as OMCS tells that beauty saloons are more likely to be associated with females. Additionally, the authors propose several techniques to counter these incidents ranging from delaying posts, issuing explicit warnings, etc., to educating users about cyberbullying. Stressing the difference between cyberbullying and other forms of cyberaggression,

Hosseinmardi et al. [4] consider instagram pictures with a minimum of fifteen comments of which more than 40% contain at least one profane word, to account for repetitiveness of bullying. Their best

 DOI : 10.36893.DRSR.2021.V08I14.101-106

performing classifier uses uni-gram and tri-gram text features with image category (e.g., person, car, nature, etc.) and its meta data to achieve an F1 score of 0.87.

3] PROBLEM DEFINTION:

In the past, work on this topic has been done from the perspective of administrators who want to filter out any content perceived as malicious according to their website policy. However, none of these considers any specific victim. On the contrary, we look at the problem from the victims perspective. We consider a comment to be shaming only when it criticizes the target of the shaming event. For example, while "Justine Sacco gonna get off that international flight and cry mountain stream fresh white whine tears b" is an instance of shaming, a comment like "Just read the Justine Sacco story lol smh sucks that she got fired for a funny tweet. People so fuckin sensitive." is not an example of shaming from the perspective of Justine Sacco (although it contains censored words) as it rebukes other people and not her.

Disadvantages

- There is no accurate analysis lack of Classification using Support Vector Machine
- Only text classification and there is no sentiment analysis for different online public shaming.

4] PROPOSED APPROACH:

In the proposed system, the system proposes a methodology for the detection and mitigation of the ill effects of online public shaming. We make three main contributions in this work- (a) Categorization and automatic classification of shaming tweets

(b) Provide insights into shaming events and shamers

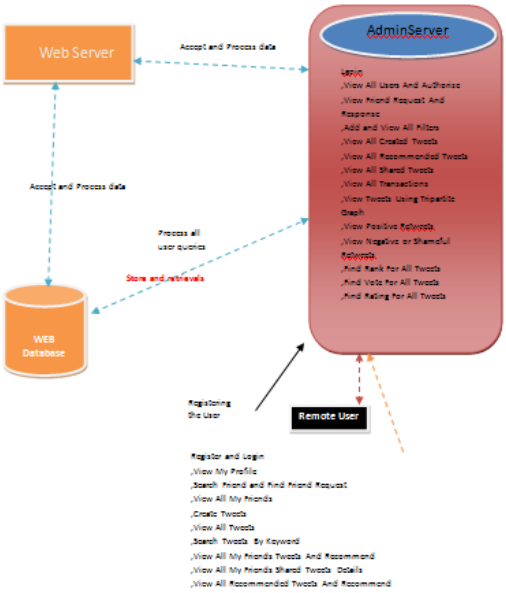
(c) Design and develop a novel application named Block Shame that can be used by a Twitter user for blocking shamers

Advantages

- The System is very effective due to AUTOMATED CLASSIFICATION OF SHAMING TWEETS.
- The System provides Analysis in the presence of Classification using Support Vector Machine.

 DOI : 10.36893.DRSR.2021.V08I14.101-106

5] SYSTEM ARCHITECTURE:



6] PROPOSED METHODOLOGY:

Admin

In this module, the Admin has to login by using valid user name and password. After login successful he can do some operations such as view all user and their details and authorize them, Add and View All Filters, View All Created Tweets, View All Recommended Tweets, View All Shared Tweets, View All Transactions, View Tweets Using Tripartite Graph, View Positive Retweets,View Negative or Shameful Retweets,Find Rank For All Tweets ,Find Vote For All Tweets, Find Rating For All Tweets

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

User

In this module, there are n numbers of users are present. User should register before doing some operations. After registration successful he has to wait for admin to authorize him and after admin authorized him. He can login by using authorized user name and password. Login successful he will do some operations like View My Profile, Search Friend and Find Friend Request, View All My Friends,

 DOI : 10.36893.DRSR.2021.V08I14.101-106

Create Tweets, View All Tweets, Search Tweets By Keyword, View All My Friends Tweets And Recommend ,View All My Friends Shared Tweets Details ,View All Recommended Tweets And Recommend.

Viewing Profile Details

In this module, the user can see their own profile details, such as their address, email, mobile number, profile Image.

Search Friends, Request, and View Friend Requests, View all Friend Details

In this, the user search for other users by their names, send requests and view friend requests from other users. User can see all his friend details with their images and personnel details.

Create Tweets

In this, the user can create their own tweets by providing tweet name, tweet description; tweet images and hash will be created based on tweet name.

View all your Tweets with Ranks

In this, the user can view all his created tweets with details along with tweet ranks.

View all Your Friends' tweets and Make Your Comment

In this, the user can view all his friends' created tweets and make your comment. If the user posts a comment more than once a day for particular tweet then the tweet rank will not increment for each comment. The Tweet Rank will be incremented only once even if user posts more comments on a day for particular tweet.

7] CONCLUSION:

In this work, 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. Instead of treating tweets as stand alone utterances, we studied them to be part of certain shaming events. In doing so, we observe that seemingly dissimilar events share a lot of interesting properties, such as, a Twitter user's propensity to participate in shaming, retweet probabilities of the shaming types and how these events unfold
in time.

DOI : 10.36893.DRSR.2021.V08I14.101-106

REFERENCES:

[1] J. Ronson, So You've Been Publicly Shamed. Picador, 2015.
[2] E. Spertus, "Smokey: Automatic recognition of hostile messages," in AAAI/IAAI, 1997, pp. 1058–1065.
[3] S. Sood, J. Antin, and E. Churchill, "Profanity use in online communities,"in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2012, pp. 1481–1490.
[4] S. Rojas-Galeano, "On obstructing obscenity obfuscation," ACM Transactions on the Web (TWEB), vol. 11, no. 2, p. 12, 2017.
[5] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," in Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017, pp. 1391–1399.
[6] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics, Valencia, Spain, 2017, pp. 1–10.
[7] Hate-Speech, "Oxford dictionaries," retrieved August 30, 2017 from https://en.oxforddictionaries.com/definition/hate speech.
[8] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in Proceedings of the Second Workshop on Language in Social Media. Association for Computational Linguistics, 2012, pp. 19–26.
[9] I. Kwok and Y. Wang, "Locate the hate: Detecting tweets against blacks." in AAAI, 2013.
[10] P. Burnap and M. L. Williams, "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making," Policy & Internet, vol. 7, no. 2, pp. 223–242, 2015.
[11] Lee-Rigby, "Lee rigby murder: Map and timeline," retrieved December 07, 2017 from https://http://www.bbc.com/news/uk-25298580.
[12] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on twitter." in SRW@ HLT-NAACL, 2016, pp. 88–93.
[13] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2017, pp. 759–760.
[14] D. Olweus, S. Limber, and S. Mihalic, "Blueprints for violence prevention, book nine: Bullying prevention program," Boulder, CO: Center for the Study and Prevention of Violence, 1999.
[15] P. K. Smith, H. Cowie, R. F. Olafsson, and A. P. Liefooghe, "Definitions of bullying: A comparison of terms used, and age and gender differences, in a fourteen–country international comparison," Child development, vol. 73, no. 4, pp. 1119–1133, 2002.
[16] R. S. Griffin and A. M. Gross, "Childhood bullying: Current empirical findings and future directions for research," Aggression and violent behavior, vol. 9, no. 4, pp. 379–400, 2004.
[17] H. Vandebosch and K. Van Cleemput, "Defining cyberbullying: A qualitative research into the perceptions of youngsters," CyberPsychology & Behavior, vol. 11, no. 4, pp. 499–503, 2008.
[18] H. Vandebosch and K. Van Cleemput, "Cyberbullying among youngsters: Profiles of bullies and victims," New media & society, vol. 11, no. 8, pp. 1349–1371, 2009.
[19] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," ACM Transactions on Interactive Intelligent Systems (TiiS), vol. 2, no. 3, p. 18, 2012.
[20] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu, "Open mind common sense: Knowledge acquisition from the general public," in OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". Springer, 2002, pp. 1223–1237.
[21] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Detection of cyberbullying incidents on the instagram social network," arXiv preprint arXiv:1503.03909, 2015.
[22] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial behavior in online discussion communities." in ICWSM, 2015, pp. 61–70.
[23] J. Cheng, C. Danescu-Niculescu-Mizil, J. Leskovec, and M. Bernstein, "Anyone can become a troll," American Scientist, vol. 105, no. 3, p. 152, 2017.

 DOI : 10.36893.DRSR.2021.V08I14.101-106