2581-4575

International Journal For Recent Developments in Science & Technology

Crossref

A Peer Reviewed Research Journal

A BRIEF ANALYSISANDDETECTION OF ONLINE PUBLIC SHAMING ON TWITTER

MRS. JASTI SWARUPA[1], SRAVYA ALAPARTHI[2], RETZ MAHIMA DEVARAPALLI[3]

#1 Assistant Professor,Dept Of IT, Vignan's Lara Institute Of Technology and Science, Vadlamudi, Guntur (dt), E-Mail Id :swarupa.jasti.77@gmail.com

#2 Assistant Professor,Dept Of IT, Vignan's Lara Institute Of Technology and Science, Vadlamudi, Guntur (dt),Email Id:sravyaalaparthi1714@gmail.com

#3 Assistant Professor,Dept Of IT, Vignan's Lara Institute Of Technology and Science, Vadlamudi, Guntur (dt),E-Mail Id:dretzmahima56@gmail.com

ABSTRACT

Online Public shaming rapidly increasing in social networks and related online public platforms like Facebook has been increasing in recenttimes .These events are known to possess a devastating impact on the victim's social, political, and financial life. Not with standing its known ill effects, little has been wiped out popular online social media to remedy this, often by the excuse of huge volume and variety of such comments and, therefore, unfeasible number of human moderators required to realize the task. Here In this paper, we automated the public shaming detection in Facebook from the attitude of victims and explore primarily two aspects, namely, events and shamers. Shaming tweets are differentiated intoseveral types abusive, comparison, passing judgment, religious/ethnic, sarcasm/joke, and whataboutery, and every tweet is assessed into one of these types or as non shaming. it's observed that out of all the participating users who post comments during a particular shaming event, majority of them are likely to shame the victim. Interestingly, it's also the shamers whose follower counts increase faster than that of the nonshamers in Twitter.

keywords— BlockShame, online user behavior, public sham- ing, tweet classification.

1.INTRODUCTION

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 someyears.Publicshamingeventshavefar-reachingimpact on virtually every aspect of victim's life. Such events

have certain distinctive characteristicsthatsetthemapartfromother similar phenomena: 1) a definite single target or victim; 2) an actioncommittedbythevictimperceived tobewrong;and3)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., whichmaynotcontaincensoredcontentexplicitly. The enormous 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*“GoingtoAfrica.HopeIdon’tget AIDS.Justkidding. 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 her plane landed in South Africa. Jon Ronson’s “So You’ve Been Publicly Shamed” [1] presents an account of several online public shaming victims. What is common for a diverse set of shaming events we have studied is that the victims are subjectedtopunishmentsdisproportionatetothelevelofcrime they have apparently committed. In Table I, we have listed the victim, year in which the event took place, action that triggered public shaming along with the triggering medium, and its immediate consequences for each studied event. “Trig- ger” is the action or words spoken by the “Victim” which initiated public shaming. “Medium of triggering” is the first communication media through which 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.

2.LITERATURE SURVEY

2.1 S. Rojas-Galeano, “On obstructing obscenity obfuscation,” *ACM Trans. Web*,vol.11,no.2,p.12,2017.

Grecian agora was the public place where citizens in ancient times gathered to debate current affairs and exercise rhetoric as a way to persuade audiences to follow a proposition for activity. These days computerized media, for example, interpersonal organizations stages, initially considered as basic virtual plug sheets to trade data among companions, have developed to become contemporary agoras, where any individual with an Internet–associated gadget may communicate their feelings and discussion them transparently and openly. Lamentably, the medium as well as the talk have changed, and talk contentions are currently every now and again dependent on feeling instead of reason, yielding conversations expected to disparage, contort or confound other's sentiment, staying away from genuine based discussion for supposition control by methods for counterfeit news, slander and individual or social gathering threatening vibe, a situation presently normally alluded as post–truth legislative issues [4]. Feeling guided contentions may lead effectively to radicalism in political, strict, ethnic, game or minorities sees, which thusly may bring about remarks shaded with individual animosity, badgering or cyberbullying [7, 3, 12]. In this direction, Google Counter-Abuse Technology Team has launched Perspective, a tool to identify toxicity of a written comment based on crowd–sourcing and machine learning models trained on large datasets of toxic conversations, as an attempt to provide safer places for online discussions [17]. Despite the remarkable efficacy of this tool to identify high–calibre language in diverse hot topics such as US Presidential election, Brexit and climate change, it has been suggested recently that its detection mechanism can be heavily defeated using adversarial strategies that corrupt the input text sequence with typographic or polarity manipulation, to such a degree that becomes unrecognisable to the trained model but remains readable by the human eye. For example, Hosseini et al. [6] has shown that the insulting statement “They are liberal idiots who are uneducated” (toxicity: 90%), becomes a mild comment when written as “They are liberal i.diots who are un.educated” (toxicity: 15%). Similarly, the rude sentence “It’s stupid and wrong” (toxicity: 89%), remains rude even if negated

2.2 E.Wulczyn,N.Thain,andL.Dixon,“Exmachina:Personalattacksseen at scale,” in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp.1391–1399.

The harm individual assaults cause to online talk rouses numerous stages to attempt to check the marvel. In any case, understanding the commonness and effect of individual assaults in online stages at scale remains shockingly troublesome. The commitment of this paper is to create and show a strategy that consolidates publicly supporting and AI to dissect individual assaults at scale. We show an assessment technique for a classifier as far as the accumulated number of group laborers it can rough. We apply our system to English Wikipedia, creating a corpus of over 100k excellent human-marked remarks and 63M machine-named ones from a classifier that is in the same class as the total of 3 group laborers, as estimated by the zone under the ROC bend and Spearman connection. Utilizing this corpus of machine-named scores, our philosophy permits us to investigate a portion of the open inquiries concerning the idea of online individual assaults. This uncovers most of individual assaults on Wikipedia are not the aftereffect of a couple of malignant clients, nor essentially the outcome of permitting unknown commitments from unregistered clients.

3.PROPOSED WORK

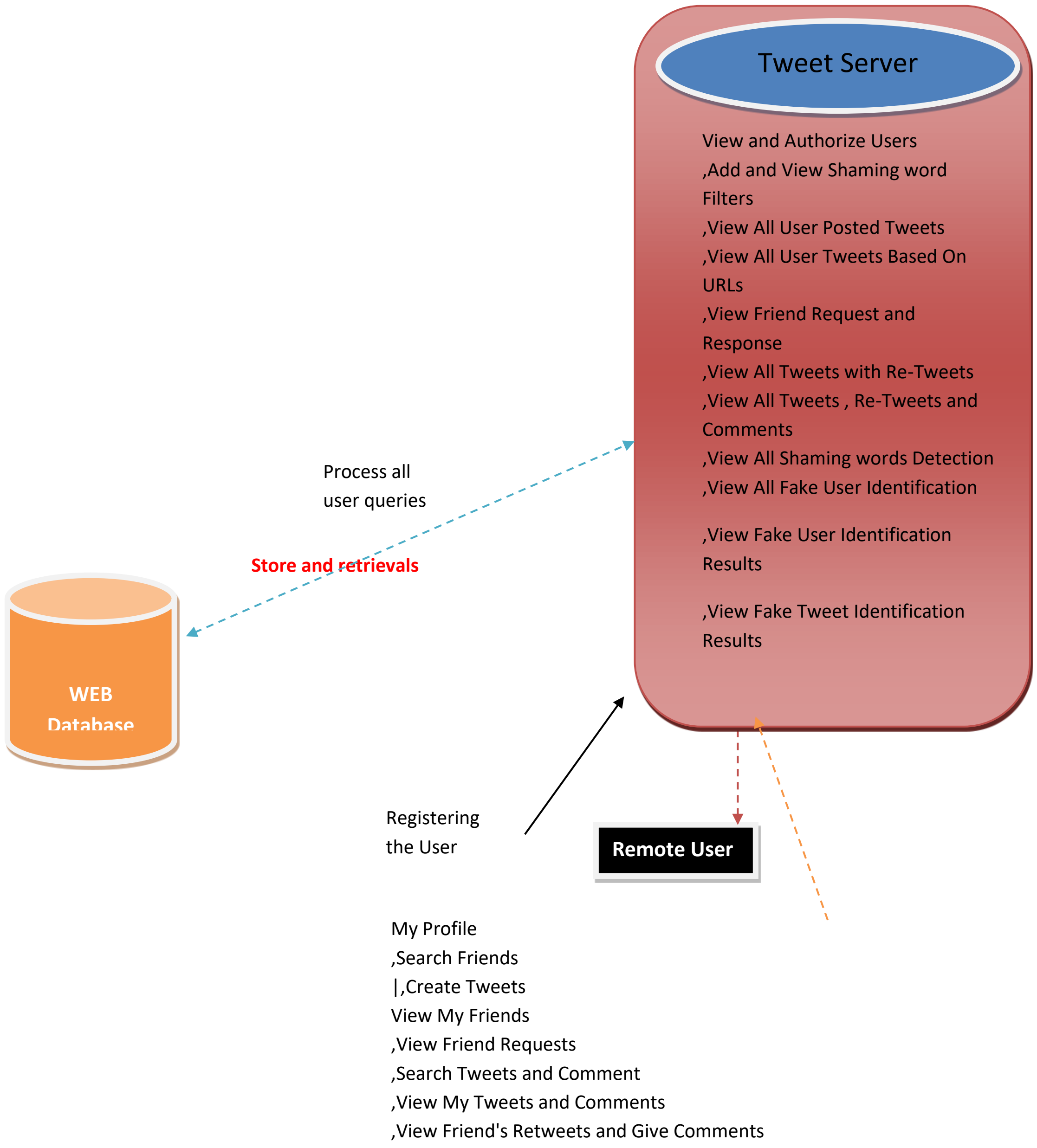

Tweet Server
View and Authorize Users
,Add and View Shaming word Filters
,View All User Posted Tweets
,View All User Tweets Based On URLs
,View Friend Request and Response
,View All Tweets with Re-Tweets
,View All Tweets , Re-Tweets and Comments
,View All Shaming words Detection
,View All Fake User Identification
,View Fake User Identification Results
,View Fake Tweet Identification Results
Process all user queries
Store and retrievals
WEB Database
Registering the User
Remote User
My Profile
,Search Friends
|,Create Tweets
View My Friends
,View Friend Requests
,Search Tweets and Comment
,View My Tweets and Comments
,View Friend's Retweets and Give Comments

Fig 1:Architecture

3.1Admin

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 and Authorize Users,Add and View Spam Filters ,View All User Posted Tweets,View All User Tweets Based On URLs,View Friend Request and Response,View All Tweets with Re-Tweets,View All Tweets , Re-Tweets and Comments,View All Spammers Detection,View All Fake User Identification,View Fake User Identification Results,View Fake Tweet Identification Results

3.2 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 My Profile, Search Friends ,Create Tweets, View My Friends,View Friend Requests,Search Tweets and Comment ,View My Tweets and Comments,View Friend's Retweets and Give Comments.

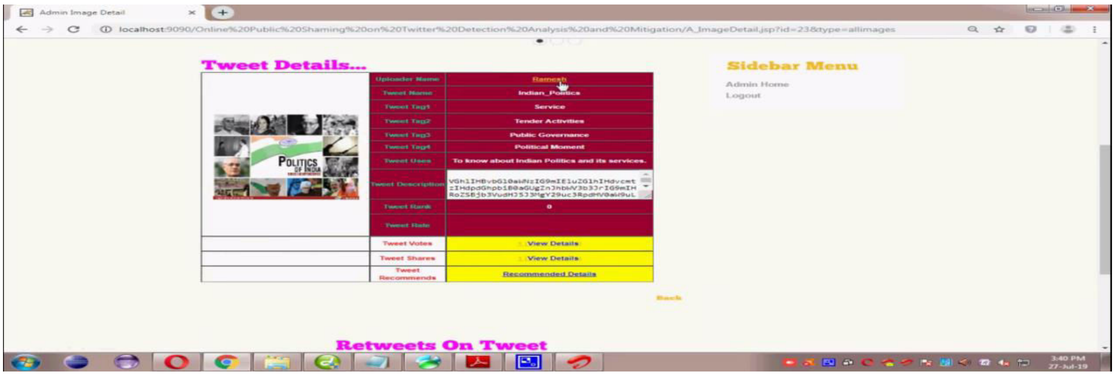
3.3 Friend Request & Response

In this module, the admin can view all the friend requests and responses. Here all the requests and responses will be displayed with their tags such as Id, requested user photo, requested user name, user name request to, status and time & date. If the user accepts the request then the status will be changed to accepted or else the status will remains as waiting.

3.4 Searching Users to make friends

In this module, the user searches for users in Same Network and in the Networks and sends friend requests to them. The user can search for users in other Networks to make friends only if they have permission.

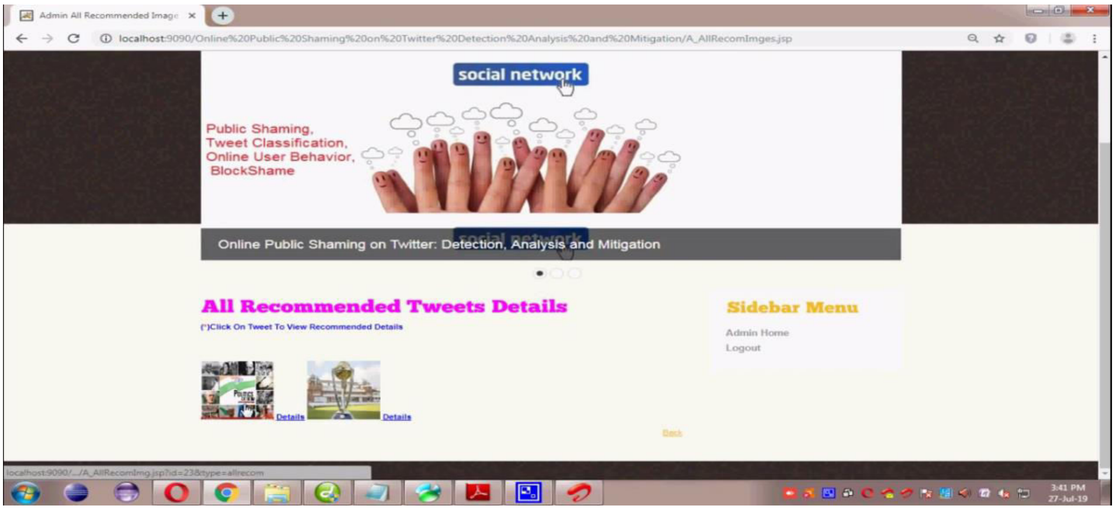
4.RESULTS AND DISCUSSION


Admin Image Detail
localhost:9090/Online%20Public%20Shaming%20on%20Twitter%20Detection%20Analysis%20and%20Mitigation/A_ImageDetail.jsp?id=23&type=allimages
Sidebar Menu
Admin Home
Logout
Back
Retweets On Tweet

Tweet Details...

Uploader Name	Ramesh
Tweet Name	Indian_Politics
Tweet Tag1	Service
Tweet Tag2	Tender Activities
Tweet Tag3	Public Governance
Tweet Tag4	Political Moment
Tweet Uses	To know about Indian Politics and its services.
Tweet Description	VGhlIHBvbGl0aWNzIG9mIEluZGlhIHdvcmt zIHdpdGhpbiB0aGUgZnJhbWV3b3JrIG9mIH RoZSBjb3VudHJ5J3MgY29uc3RpdHV0aW9uL
Tweet Rank	0
Tweet Rate	
Tweet Votes	View Details
Tweet Shares	View Details
Tweet Recommends	Recommended Details

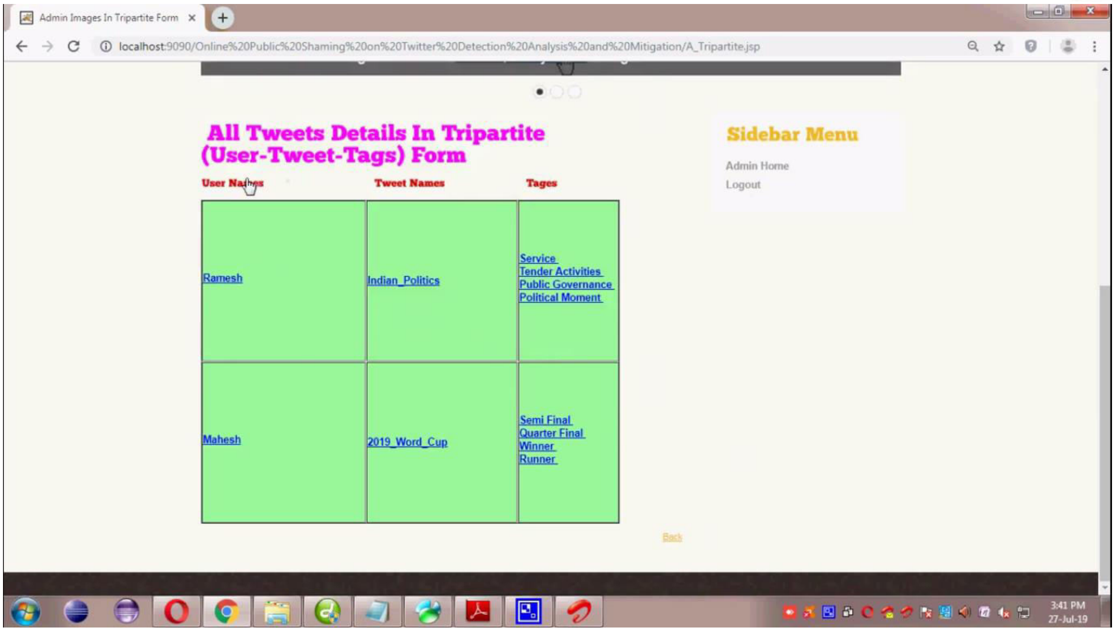
Fig 2:Tweet Details


Admin All Recommended Image
localhost:9090/Online%20Public%20Shaming%20on%20Twitter%20Detection%20Analysis%20and%20Mitigation/A_AllRecomImges.jsp
social network
Public Shaming,
Tweet Classification,
Online User Behavior,
BlockShame
Online Public Shaming on Twitter: Detection, Analysis and Mitigation
Sidebar Menu
Admin Home
Logout
Details
Details
Back
localhost:9090/.../A_AllRecomImg.jsp?id=23&type=allrecom
3:41 PM
27-Jul-19

All Recommended Tweets Details

(*)Click On Tweet To View Recommended Details

Fig 3:Recommeded Tweet Details


Admin Images In Tripartite Form
localhost:9090/Online%20Public%20Shaming%20on%20Twitter%20Detection%20Analysis%20and%20Mitigation/A_Tripartite.jsp
3:41 PM
27-Jul-19

All Tweets Details In Tripartite (User-Tweet-Tags) Form

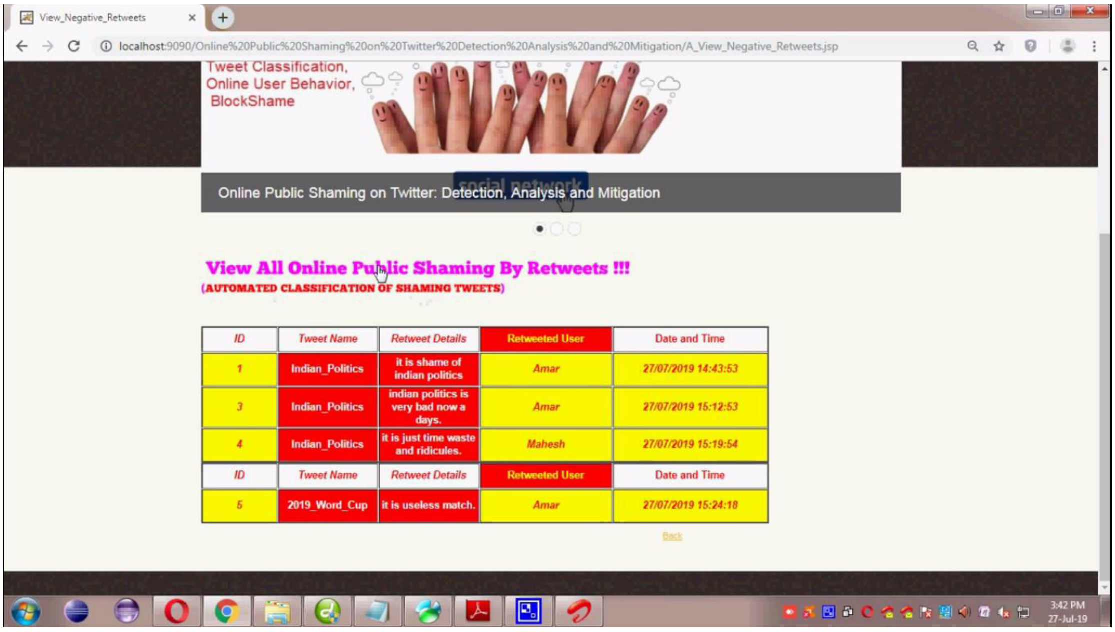
User Names	Tweet Names	Tages
Ramesh	Indian_Politics	Service Tender Activities Public Governance Political Moment
Mahesh	2019_Word_Cup	Semi Final Quarter Final Winner Runner

Sidebar Menu

Admin Home
Logout

Back

Fig 3:Tweets and Tags Details


View_Negative_Retweets
localhost:9090/Online%20Public%20Shaming%20on%20Twitter%20Detection%20Analysis%20and%20Mitigation/A_View_Negative_Retweets.jsp
Tweet Classification,
Online User Behavior,
BlockShame
social network
3:42 PM
27-Jul-19

Online Public Shaming on Twitter: Detection, Analysis and Mitigation

View All Online Public Shaming By Retweets !!!
(AUTOMATED CLASSIFICATION OF SHAMING TWEETS)

ID	Tweet Name	Retweet Details	Retweeted User	Date and Time
1	Indian_Politics	it is shame of indian politics	Amar	27/07/2019 14:43:53
3	Indian_Politics	indian politics is very bad now a days.	Amar	27/07/2019 15:12:53
4	Indian_Politics	it is just time waste and ridicules.	Mahesh	27/07/2019 15:19:54
ID	Tweet Name	Retweet Details	Retweeted User	Date and Time
5	2019_Word_Cup	it is useless match.	Amar	27/07/2019 15:24:18

Back

Fig 5:Shaming Words Detection

5.CONCLUSION

Here finally we concluded 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 standalone 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 intime.With the growth of online social networks and proportional rise in public shaming events, voices against callousness on part of the site owners are growing stronger. Categorization of shaming comments as presented in this paper has the potential for a user to choose to allow certain types of shaming comments (e.g., comments that are sarcastic in nature) giving his/her an opportunity for rebuttal and block others (e.g., comments that attack her ethnicity) according to individual choices. Freedom to choose what type of utterances one would not like to see in his/her feed beforehand is way better than flagging a deluge of comments on the event of shaming. This also liberates moderators from the moral dilemma of deciding a threshold that separates acceptable online behavior from unacceptable ones, thus relieving themselves to a certain extent from the responsibility of fixing what is best foranother person.

REFERENCES

[1] J. Ronson, *So You've Been Publicly Shamed*. London, U.K.: Picador, 2015.

[2] E.Spertus,"Smokey:Automaticrecognitionofhostilemessages,"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,andL.Dixon,"Ex machina:Personalattacksseen 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.2ndWorkshopLang.SocialMedia*,2012,pp.19–26.

[9] I.KwokandY.Wang,"Locatethehate:Detectingtweets againstblacks," in*Proc.AAAI*,2013,pp.1621–1622.

[10] P. Burnap and M. L. Williams, "Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decisionmaking,"*PolicyInternet*,vol.7,no.2,pp.223–242,2015.

[11] Lee-Rigby. *Lee Rigby Murder: Map and Timeline*. Accessed: Dec. 7, 2017. [Online]. Available: https://http://www.bbc.com/news/uk-25298580

[12] Z. Waseemand D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in *Proc. 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 *Proc. 26th Int. Conf. World Wide Web Companion*, 2017, pp.759–760.

[14] D. Olweus, S. Limber, and S. Mihalic, *Blueprints for Violence Pre- vention, Book Nine: Bullying Prevention*

Program. Boulder, CO, USA: Center for the Study and Prevention of Violence,1999.

[15] P. K. Smith, H. Cowie, R. F. Olafsson, and A. P. D. Liefooghe, "Definitions of bullying: A comparison of terms used, and age and gender differences, in a fourteen–country international comparison," *Child Develop.*, 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 Violent Behav.*, vol.9,no.4,pp.379–400,2004.

[17] H. Vandeboschand K. Van Cleemput, "Defining cyberbullying: A qualitative research into the perceptions of youngsters," *CyberPsychol. Behav.*,vol.11,no.4,pp.499–503,2008.

[18] H. Vandeboschand K. Van Cleemput, "Cyberbullying among young- sters: Profiles of bullies and victims," *New Media Soc.*, vol. 11, no. 8, pp. 1349–1371,2009.

[19] K.Dinakar,B.Jones,C.Havasi,H.Lieberman,andR.Picard,"Common sense reasoning for detection, prevention, and mitigation of cyberbully- ing," *ACM Trans. Interact. Intell. Syst.*, 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 *Proc. OTM Confederated Int. Conf. Move Meaningful Internet Syst.* Berlin, Germany: Springer, 2002, pp.1223–1237.

[21] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q.Lv, and S. Mishra. (2015). "Detection of cyberbullying incidents on the instagram social network."[Online]. Available: https://arxiv.org/abs/1503.03909

[22] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial behavior in online discussion communities," in *Proc. ICWSM*, 2015, pp.61–70.

[23] J. Cheng, C. Danescu-Niculescu-Mizil, J. Leskovec, and M. Bernstein, "Anyone can become a troll," *Amer. Sci.*, vol. 105, no. 3, p. 152,2017.

[24] P. Tsantarliotis, E. Pitoura, and P. Tsaparas, "Defining and predicting troll vulnerability in online social media," *Social Netw. Anal. Mining*, vol. 7, no. 1, p. 26,2017.

AUTHOR PROFILES

Mrs. JASTI SWARUPA completed her Bachelor Degree in Computer Science from VelagaNageswararao College of Engineering, Ponnur. She has completed Master Degree in Computer Science from Vignan's Lara Institute Of Technology and Science, Vadlamudi. Currently She is working as an Assistant Professor in IT department at Vignan's Lara Institute Of Technology and Science, Vadlamudi, Guntur (dt). Her areas of interests are Networks, and Image Processing.

SravyaAlaparthi is currently working as an Assistant Professor in Vignan’s Lara Institute of Technology and Science, Vadlamudi. She completed her master’s degree in computer science and engineering, Vignan’s Foundation for Science and Technology. Her current research interests include Machine Learning and Networks.

Retz MahimaDevarapalli is currently working as an Assistant Professor in Vignan's Lara Institute of Technology and Science, Vadlamudi. She completed her master’s degree in computer science and engineering, Vignan’s Foundation for Science and Technology. Her current research interests include Biomedical Image Processing and Machine Learning.