International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, Pp: 3489-3495 www.ijaem.net ISSN: 2395-5252

Online Public Shamming On Twitter: Detection, Analysis, and Mitigation

Pooja S K, Shambhavi H, Arshad Pasha, SHALINI M S

Information science & engineering
Assistant professor
Mysore college of engineering and management, mysuru-570028

SUBMITTED: 15-07-2021 REVISED: 29-07-2021 ACCEPTED: 31-07-2021

ABSTRACT –

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 a 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. 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 whataboutery, and each tweet is classified into one of these types or as nonshaming. 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 nonshamers in Twitter. Finally, based on categorization and classification of shaming tweets, a 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,Internet, Task analysis, Companies, Distortion, Cultural differences

I. INTRODUCTION

Aim: The objective of this project is to develop a system to identify the tweet harbor hatred in micro blogs during disaster events.

Online shaming is a form of public shaming in which targets are publicly humiliated on the internet, via social media platforms (e.g. Twitter or Facebook), or more localized media (e.g. email groups). As online shaming frequently involves exposing private information on the Internet, the ethics of public humiliation has been a source of debate over internet privacy and media ethics. Online shaming takes many forms, including call-outs, cancellation (cancel culture), doxing, negative reviews, and revenge porn.

Online shaming is a form of public shaming in which internet users are harassed, mocked, or bullied by other internet users online. This shaming may involve commenting directly to or about the shamed; the sharing of private messages; or the posting of private photos. Those being shamed are perceived to have committed a social transgression, and other internet users then use public exposure to shame the offender.

People have been shamed online for a variety of reasons, usually consisting of some form of social transgression such as posting offensive comments, posting offensive images or memes, online gossip, or lying. Those who are shamed online have not necessarily committed any social transgression, however. Online shaming may be used to get revenge (for example, in the form of revenge pornography), stalk, blackmail, or to threaten other internet users.[1]

Privacy violation is a major issue in online shaming. Those being shamed may be denied the right to privacy and be subject to defamation. David Furlow, chairman of the Media, Privacy and Defamation Committee of the American Bar Association, has identified the potential privacy concerns raised by websites facilitating the distribution of information that is not part of the public record (documents filed with a government agency) and has said that such websites "just [give] a forum to people whose statements may not reflect truth.

II. EXISTING AND PROPOSED SYSTEM

We propose a machine learning based

classification algorithm to analyze the live tweet. In this system we are using a classifier to identify the words used in the tweet. The algorithm takes specific keywords list as input and the tweet to be identified as another input. The results were plotted in a graph.

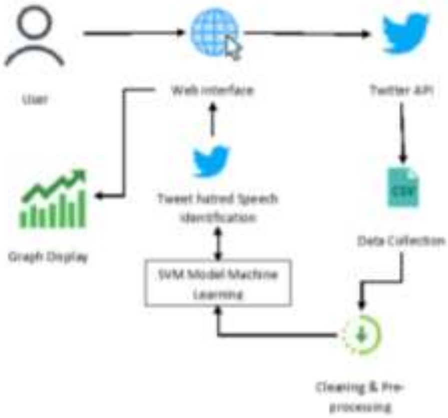

User
Web interface
Twitter API
Graph Display
Tweet hatred Speech identification
CSV
Data Collection
SVM Model Machine Learning
Cleaning & Pre-processing

Figure 1 existing system

DISADVANTAGES OF EXISTING SYSTEM:

- Most of the previous works mentioned above do not make a distinction between acceptability and non-acceptability of a comment based on the presence or absence of a predefined victim.
- Personal attack is less rigorously defined and often holds all of the above-mentioned categories in it. Such attacks can be directed toward the author of a previous comment or a third party.

PROPOSED SYSTEM:

- In this paper, we propose a methodology for the detection and mitigation of the ill effects of online public shaming. We make three main contributions in this paper:
- (1) categorization and automatic classification of shaming tweets;
- (2) provide insights into shaming events and shamers;
- (3) design and develop a novel application named Block-Shame that can be used by a Twitter user for blocking shamers.

ADVANTAGES OF PROPOSED SYSTEM:

- Our goal is to automatically classify tweets in the aforementioned six categories.
- Both labeled training set and test set of tweets for each of the categories go through the preprocessing and feature extraction steps

III. CATEGORIZATION OF SHAMING TWEETS

After studying more than 1000 shaming tweets from eight shaming events on Twitter, we have come up with six categories of shaming tweets as shown in Table.

A brief description of these categories along with their most common attributes is given in the following.

1) Abusive: A comment falls in this category when the victim is abused by the shamer. It may be noted that mere presence of a list of abusive words is not enough to detect abusive shaming, because a comment may contain abusive utterances but it can still be in support of the victim. However, abusive words associated with the victim as found from dependency parsing of the comment are a strong marker of this type of shaming.

2) Comparison: In this form of shaming, the intended victim's action or behavior is compared and contrasted with another entity. The main challenge here is to automatically detect perception of the entity mentioned in the comment so as to determine whether the comparison is an instance of shaming. The text itself may not contain enough hints, e.g., adjectives with polarity associated with the entity. In such cases, the author of the comment relies on the collective memory of the social network users to provide for the necessary context. This is true more often when the said entity appeared recently in other events, e.g., "#AamirKhan you have forgotten that acting is being appreciated only in cinema! Learn something from Mahadik's1 wife." This comment would be understood as shaming (Aamir Khan is the target) with little effort by anyone who has the knowledge that Mahadik is a positive mention. For someone who thinks Mahadik is a negative mention, the intent of the comment be-

comes ambiguous. Automatically predicting polarity of a mentioned entity in a comment in real time is a difficult task. An approximation would be average perception (sentiment score) about the entity in most recent comments, recent news sources, and so on. A static database would be of little use as public perception about an entity can change frequently.

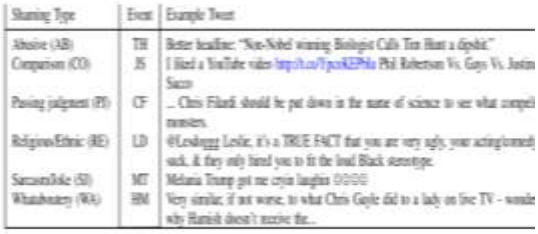
3) Passing Judgment: Shamers can pass quick judgments vilifying the victim. Passing judgment often overlaps with other categories. A comment is PJ shaming only when it does not fall in any of the other categories. Passing judgment often starts with a verb and contains modal auxiliary verbs.

4) Religious/Ethnic: Often, there are multiple groups which a person identifies with. We consider three types of identities of a victim- nationality like Indian, Chinese, ethnicity/race like black, white, and religious like Christian and Jewish. Maligning any one of these group identities in reference to the victim constitutes a religious/ethnic shaming. In this paper, we assume that we know the group identities to which a victim associates. For example, Justine Sacco is a U.S. citizen, white, and Christian. In actual scenario, this information can be inferred from the user's profile information on Twitter like name and location. In their absence, the display picture can potentially be used to predict a user's demographic information (see [26] uses a third party service called Face++ [27]).

5) Sarcasm/Joke: Sarcasm is defined as "a way of using words that are the opposite of what one means in order to be unpleasant to somebody or to make fun of them" in Oxford learner's dictionary. This definition is also used by some recent work on sarcasm detection in Twitter like that in [28]. We have tagged joke and sarcasm in the same category due to an inherent overlap between the two. A sarcasm/joke tweet is not shaming unless the subject of fun is the victim, for example, "Wow I remember last night seeing the Justine Sacco thing start, never thought it would get this big! Well played guys!" This tweet sarcastically criticizes Twitter users. Hence, it is not shaming. Presence of emojis and sudden change of sentiment are important attributes of this category.

6) Whataboutery: In whataboutery, the shamer highlights the victim's purported duplicity by pointing out earlier action/in-action in a past situation similar to the present one. Important indicators for these categories of comments are the use of Wh-adverbs (such as What, Why, How, etc.) and past form of verbs. It is worthwhile mentioning that in a work-in-progress version of this paper published as a poster paper [29], we categorized shaming into ten broad categories including the six described above. However, after more detailed scrutiny, in this paper, we have merged and omitted certain categories due to several reasons including sharing of features between two categories, low occurrences of comments in a category, and so on

Different Forms of Shaming Tweets



Shaming Type	Event	Example Tweet
Abusive (AB)	TH	Better headline: "Non-Nobel winning Biologist Calls Tim Hunt a dipshit."
Comparison (CO)	JS	I liked a YouTube video http://t.co/TposKEPhlu Phil Robertson Vs. Gays Vs. Justin Sacco
Passing judgment (PJ)	CF	... Chris Filardi should be put down in the name of science to see what compel monsters.
Religious/Ethnic (RE)	LD	@Lesdoggg Leslie, it's a TRUE FACT that you are very ugly, your acting/comedy suck, & they only hired you to fit the loud Black stereotype.
Sarcasm/Joke (SJ)	MT	Melania Trump got me cryin laughin 😂😂😂😂
Whataboutery (WA)	HM	Very similar, if not worse, to what Chris Gayle did to a lady on live TV - wonde why Hamish doesn't receive the...

IV. ARCHITECTURE OF PROPOSED SYSTEM

Framework engineering is the hypothetical plan that characterizes the structure and conduct of a framework. A design clarification is a legitimate depiction of a framework, sorted out in a mode that backings examination about the auxiliary properties of the framework. It characterizes the framework device or building squares and gives an arrangement from which items can be obtained, and frameworks built up, that will work commonly to actualize the for the most part framework. The System design is uncovered underneath.

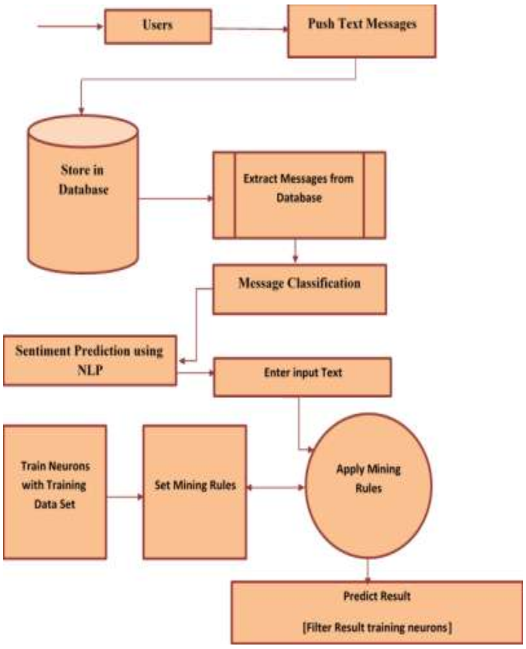

Users
Push Text Messages
Store in Database
Extract Messages from Database
Message Classification
Sentiment Prediction using NLP
Enter input Text
Train Neurons with Training Data Set
Set Mining Rules
Apply Mining Rules
Predict Result
[Filter Result training neurons]

Fig : System Architecture

V. RESULTS

Fig 1: the above figure shows that how the user can login to the platform like twitter if they are a new user they need to first fill this one and register otherwise if they are aold user they can directly login with their name and password.

Fig 2: This one shows that how the user interface of twitter where it contains a find friends ,view friends list and upload image, share photo, friends list, my_tweet, edit_profile, view profile.

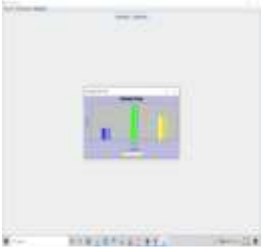
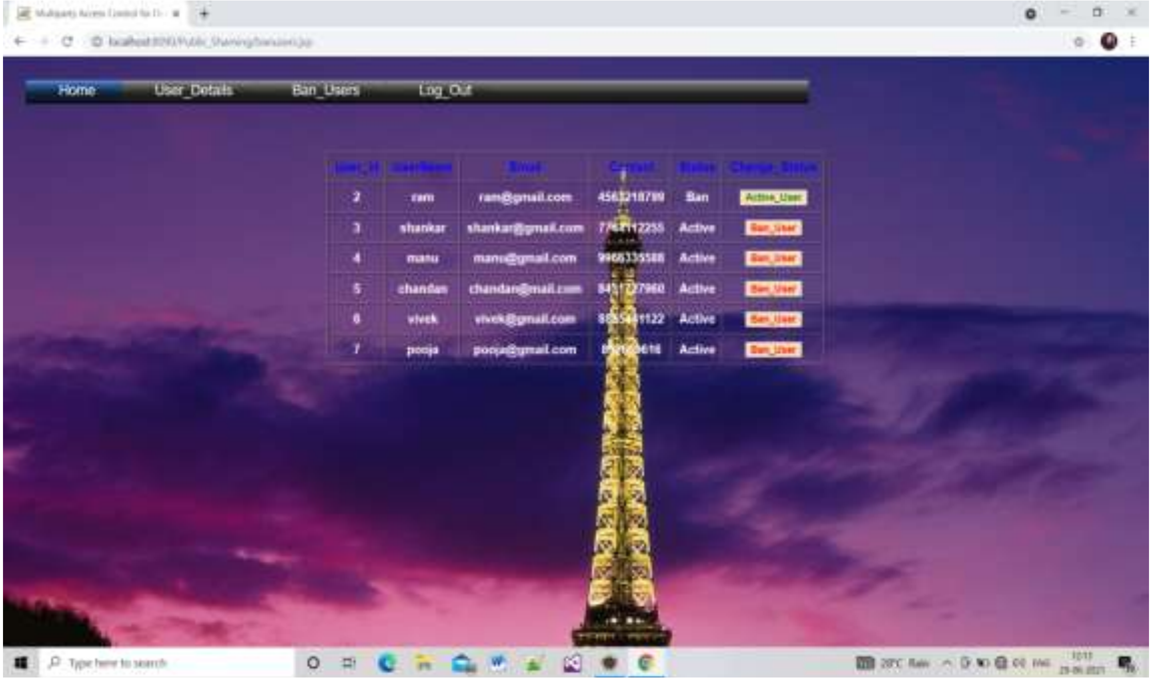
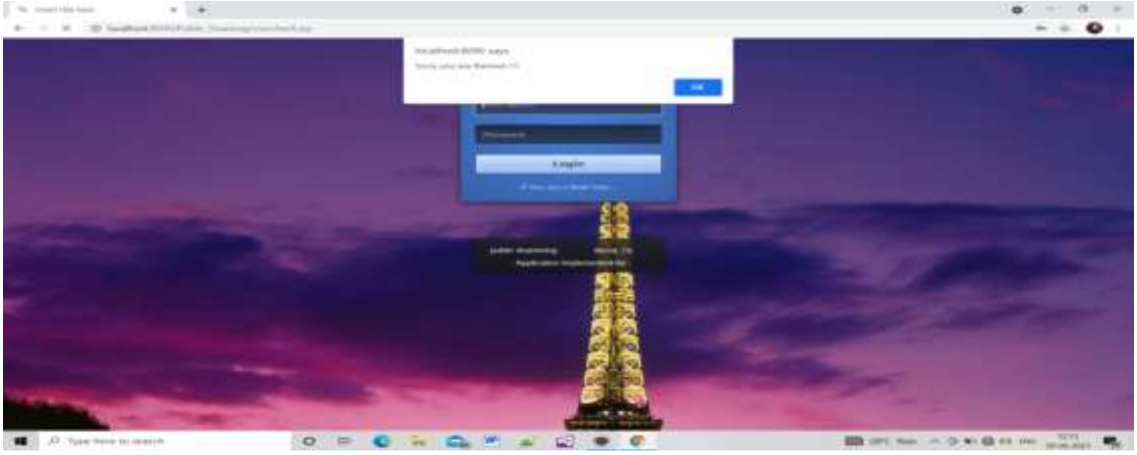


Fig 3: in this is the graph which identifies the sentiment using java filer master.



User_Id	UserName	Email	Contact	Status	Change_Status
2	ram	ram@gmail.com	4563218799	Ban	Active_User
3	shankar	shankar@gmail.com	[illegible]12255	Active	Ban_User
4	manu	manu@gmail.com	9966335588	Active	Ban_User
5	chandan	chandan@mail.com	94[illegible]27960	Active	Ban_User
6	vivek	vivek@gmail.com	98[illegible]41122	Active	Ban_User
7	pooja	pooja@gmail.com	[illegible]618	Active	Ban_User

Fig 4: in this the admin has the control to block a user who is using the bad words.



localhost:8090 says

Sorry you are Banned !!!

Fig 5:In this it shows that how the user is blocked by admin and he is not again to do such bad comments

VI. CONCLUSIONS

Public shaming is not new. It's been used as a punishment in all societies – often embraced by the formal law and always available for day-to-day policing of moral norms. However, over the past couple of centuries, Western countries have moved away from more formal kinds of shaming, partly in recognition of its cruelty.

Even in less formal settings, shaming individuals in front of their peers is now widely regarded as unacceptable behaviour. This signifies an improvement in the moral milieu, but its effect is being offset by the rise of social media and, with it, new kinds of shaming.

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 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, "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. Accessed: Dec. 7, 2017. [Online]. Available: https://http://www.bbc.com/news/uk25298580

[12]. Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in Proc. SRW HLTNAACL, 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 Prevention, 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. Vandebosch and 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. Vandebosch and K. Van Cleemput, "Cyberbullying among youngsters: 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, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," 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 in-

stagram 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.

[25]. S. O. Sood, E. F. Churchill, and J. Antin, "Automatic identification of personal insults on social news sites," J. Assoc. Inf. Sci. Technol., vol. 63, no. 2, pp. 270–285, 2012.

[26]. A. Chakraborty, J. Messias, F. Benevenuto, S. Ghosh, N. Ganguly, and K. Gummadi, "Who makes trends? Understanding demographic biases in crowdsourced recommendations," in Proc. 11th Int. AAAI Conf. Web Social Media, 2017, pp. 22–31.

[27]. Face++ Cognitive Services. Accessed: Feb. 20, 2018. [Online]. Available: https://www.faceplusplus.com/

[28]. 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. [29] 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.

[29]. C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in Proc. Assoc. Comput. Linguistics (ACL) Syst. Demonstrations, 2014, pp. 55–60. [Online]. Available: http://www.aclweb.org/anthology/P/P14/P14-5010

[30]. M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2004, pp. 168–177.

[31]. M. P. Marcus and M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of English: The Penn treebank," Comput. Linguistics, vol. 19, no. 2, pp. 313–330, 1993.

[32]. D. Bamman and N. A. Smith, "Contextualized sarcasm detection on Twitter," in Proc. ICWSM, 2015, pp. 574–577.

[33]. O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, and N. Schneider, "Part-of-speech tagging for Twitter: Word clusters and other advances," Ph.D.dissertation, School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, 2012.

[34]. Brown-Clusters. Twitter Word Clusters. Accessed: Jul. 2, 2017. [Online]. Available: http://www.cs.cmu.edu/~ark/TweetNLP/

[35]. C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.

[36]. 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.

[37]. 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.