Look Before You Shame: A Study on Shaming Activities on Twitter

Rajesh Basak, Niloy Ganguly, Shamik Sural, Soumya K Ghosh
Department of Computer Science & Engineering, Indian Institute of Technology Kharagpur
Kharagpur, India
{rajesh@sit, niloy@cse, shamik@cse, skg@cse}.iitkgp.ernet.in

ABSTRACT

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 various aspects of shaming done through twitter. An important contribution of our work is categorization of shaming tweets, which helps in understanding the dynamics of spread of online shaming events. It also facilitates automated segregation of shaming tweets from non-shaming ones.

1. INTRODUCTION

The relative ease with which opinion can be shared by almost anyone with little accountability in Twitter, often leads to undesirable virality. Spread of rumor in Twitter, for example, is well studied in the literature [1] [2]. Another fallout of negative virality - public shaming, although known to have far reaching impact on the target of shaming [3], has never been studied as a computational problem.

In this paper, we attempt to understand the phenomenon of public shaming over Twitter considering three (in)famous incidents, namely (i) In 2013, Justine Sacco (JS) faced the brunt of public shaming after posting a perceived racial tweet about AIDS and Africa (ii) In 2015, Nobel winning biologist Sir Tim Hunt's (TH) comments on women in science stormed OSNs resulting in his resignation from various academic and research positions and (iii) More recently, in November 2015, hugely popular Bollywood (Indian movie industry based in Mumbai, India) actor Aamir Khan (AK) had to face the *ire of Twitter* for commenting about his wife's alleged plans of leaving the country due to the prevalent intolerance. See Table 1 for details.

We categorize the shaming tweets in several classes based on the nature of their content against the target, like use of abusive language, making sarcastic comments, associating the target with negative characters, etc., as shown in Table 2. Such a categorization helps in understanding the trajectory of spread of shaming virality as presented next. We also identify several interesting discriminating user and tweet features related to shaming tweets.

Table 1: Comments that trigerred shaming

Justine Sacco	Going to Africa. Hope I dont get AIDS. Just kidding. I'm white!
Tim Hunt	Let me tell you about my trouble with girls. Three things happen when they are in the lab. You fall in love with them, they fall in love with you, and when you criticise them, they cry.
Aamir Khan	When I chat with Kiran at home, she says 'Should we move out of India?'

2. VARIATION IN SHAMING TYPE

For this study, shaming tweets for the three events were randomly selected from a downloaded collection of tweets and manually labeled by three annotators. They were instructed to label the tweets in one of the ten categories mentioned in Table 2. One hundred tweets from each event for which all three annotators agreed, were then analyzed.

Fig. 1 shows how the percentage of shaming categories for an event evolves as time progresses over the first three days since its start. It is observed that, *sarcasm or joke* is the most popular form of shaming in Twitter, followed by *passing judgment.* Further, the share of abusive tweets increased with time in all cases except only for the third day of the *Tim Hunt* event, where *questioning qualifications* is more popular, potentially due to the otherwise strong reputation of the target.

3. FEATURES OF SHAMING TWEETS

For automated identification of shaming tweets (across all the ten categories), we consider text features of tweet such as parts of speech, sentiment score, number of incomplete tweets, mentions, urls, hashtags as well as user features like count of status, friends, followers and favorited tweets. Some of these features are based on the LIWC [4] standard. Table 3 lists some of the features with respective mean values corresponding to non-shaming and shaming tweets. p-values for two-sample one tailed t-test are shown in the rightmost column indicating potential as a discriminating feature. Based on this data, the features with low p-values are used for classifying a tweet as shaming or non-shaming. However, these features are not discriminating enough to automatically classify a shaming tweet into one of the ten fine-grained cate-

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2889414.

Table 2: Different forms of shaming tweet

Shaming Type	*Event*	*Example Tweet*
Whataboutherism (WA)	AK	Wifey #AamirKhan Rao wasnt scared when - AR Rahman was threatened by the Muslim Ulemas
Sarcasm/Joke (SJ)	AK	Just in..Agarwal Packers and Movers has sent a Friend Request to #AmirKhan on Facebook...
Referring to religion, ethnicity (RE)	AK	trending #IStandWithAamirKhan reflects besides pseudo secular a particular community trying to malign the sovereignty of hindustan.
Associating with negative character (AN)	TH	I liked a @YouTube video `http://t.co/YpcoKEPbIu` Phil Robertson Vs. Gays Vs. Justine Sacco
Abuses (AB)	TH	Better headline: "Non-Nobel winning Biologist Calls Tim Hunt a dipshit."
Passing judgment (PJ)	TH	Tim Hunt along with all his nose hair needs to lock himself in the basement and rot there.
Comparison with ideal (CI)	TH	Tim Hunt wouldn't recognize a good scientist if Marie Curie, Jane Goodall, Shirley Ann Jackson, and Sally Ride all kickâĂę
Irrelevant past tweet (IR)	JS	I had a sex dream about an autistic kid last night. #fml
False fact-ing (FF)	JS	Isn't Justine Sacco's father a billionaire business man in South Africa?
Questioning qualifications (QQ)	JS	Justine Sacco clearly knows nothing about media and PR. So how did she become a top PR executive?

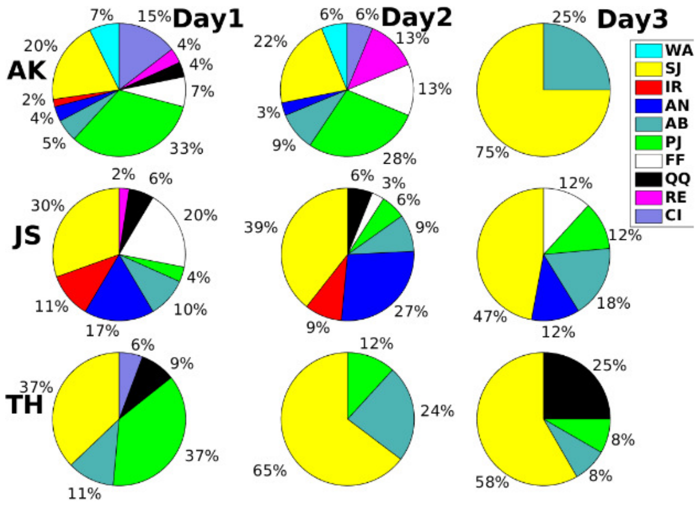

Figure 1: Shaming types for the first three days

gories - a problem that calls for more intricate use of NLP techniques and is left as future work.

4. DISCUSSION

Unlike rumors, whether detection and categorization of shaming tweets might be used to stop their spread is an open question as it could act as a two-edged sword - protecting the target from disproportionate punishment meted out without trial on OSN court vis-a-vis individual freedom of expression on OSN. Instead, we feel that our work can be used to study the nature of people who indulge in public shaming and determine their possible motive like one-upmanship, showing off righteousness, etc., based on past tweet history, number of followers, tendency to retweet and several other features that can be easily extracted. It can also find utility in the study of how a shaming target retaliates through his/her own tweets, be it in the form of apologies or by direct confrontation. All these are challenging computational problems that we plan to work on.

Table 3: Significant features with mean and p-values. HT: No. of hashtags, URL: urls, NNP: proper noun, PRP: personal pronoun, PRP$: possessive pronoun, VBG: verb present participle, WRB: "wh" adverbs, SC: status, FLC: follower, FVC: favorited count

Feature	*Non-Shaming Mean*	*Shaming Mean*	*p value*
HT	0.41	0.50	0.06
URL	0.64	0.30	<0.001
NNP	3.71	3.42	0.03
PRP	0.55	0.85	<0.001
PRP$	0.22	0.28	0.05
VBG	0.24	0.44	<0.001
WRB	0.10	0.15	0.02
SC	3.81×10^4	2.66×10^4	0.12
FLC	1.40×10^5	0.5×10^5	0.15
FVC	2.86×10^3	5.20×10^3	0.01

5. REFERENCES

[1] T. Takahashi and N. Igata. Rumor detection on Twitter. In *6th International Joint Conference on SCIS and ISIS*, pages 452–457. IEEE, 2012.

[2] Z. Zhao, P. Resnick and Q. Mei. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *24th International Conference on World Wide Web*, pages 1395–1405, 2015.

[3] J. Ronson. *So You've Been Publicly Shamed*. Picador, 2015.

[4] Y R Tausczik and J W Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.