Online Public Shaming on Twitter: Detection, Analysis and Mitigation

Rajesh Basak, Shamik Sural, Niloy Ganguly, and Soumya K. Ghosh

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 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 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—Public Shaming, Tweet Classification, Online User Behavior, BlockShame

1 Introduction

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 checkability 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 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.

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 *"Going to Africa. Hope I don't get AIDS. Just kidding. 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 subjected to punishments disproportionate to the level of crime they have apparently committed. In Table 1, 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. 'Trigger' 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.

In the past, work (e.g., [2], [3], [4], [5]) 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*

- *The authors are with the Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, India.*
- *E-mails: {rajesh@sit, shamik@cse, niloy@cse, skg@cse}.iitkgp.ernet.in*

TABLE 1: Events with trigger and consequences considered in this study

Victim	Year	Trigger	Medium of triggering	Immediate consequences
Justine Sacco (JS) PR[1] officer	2013	Tweeted 'Going to Africa. Hope I don't get AIDS. Just kidding. I'm white!'	Twitter	Fired from her job
Sir Tim Hunt (TH) Eminent biologist	2015	Commented 'Three things happen when girls are in the lab. You fall in love with them, they fall in love with you, and when you criticize them, they cry'	News media	Resignation from fellow of Royal society
Dr. Christopher Filardi (CF) Field biologist	2015	Captured and killed a bird of a relatively unknown species for collecting scientific specimen	Facebook	Criticism from biologists and general public
Aamir Khan (AK) Bollywood actor	2015	Commented on rising intolerance in India and his wife's suggestion to leave the country	News media	Removed as brand ambassador of Snapdeal[2]
Hamish McLachlan (HM) TV journalist	2016	Hugged female Channel Seven[3] colleague during a live broadcast	Television	Criticism and subsequent apology
Leslie Jones (LD) Hollywood actor	2016	Acted in a lead role in the remake of the Hollywood movie 'Ghostbusters'	News media and Youtube	Left Twitter
Melania Trump (MT) Spouse of US President	2016	A Twitter user pointed out plagiarism in one of her campaign speech	Twitter	Criticism and negative media coverage
Priyanka Chopra (PC) Bollywood actor	2017	Wore a dress that did not cover her legs when meeting the Indian Prime Minister	Facebook	Criticism

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.

In this work, we propose 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 BlockShame that can be used by a Twitter user for blocking shamers

The rest of the paper is organized as follows. Section 2 discusses related work. We introduce a categorization of shaming comments based on an in-depth study of a variety of tweets in Section 3. A methodology for identification and prevention of such incidents is proposed in Section 4. Section 5 presents details of experiments and important results. The functionality and effectiveness of BlockShame are discussed in Section 6. Finally, we conclude the paper and provide directions for future research in Section 7.

2 Related Work

Efforts to moderate user generated content in the Internet started very early. Smokey [2] is one of the earliest computational work in this direction which builds a decision tree classifier for insulting posts trained on labeled comments from two web forums. Though academic research in this area started that early, it used different nomenclature including abusive, flame, personal attack, bullying, hate speech, etc., often grouping more than a single category under a single name [6]. Based on the content (and not the specific term used), we divide the related work into five categories- profanity, hate speech, cyberbullying, trolling and personal attacks.

Sood *et al.* [3] 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. Rojas-Galeano [4] 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 'α') 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!

1. Public relations
2. Major Indian e-commerce company www.snapdeal.com
3. An Australian television channel

comments and known anti-Semitic web page contents. A similar type of work has been done on anti-black hate speech on Twitter [9]. Burnap and Williams [10] collected tweets for two weeks after the Lee Rigby incident [11] and trained a classifier on typed dependency and hateful terms as features. Waseem and Hovy [12] released a public data set of sixteen thousand tweets labeled in one of the three categories- racist, sexist or none. They achieved an F1 score of 0.73 using character n-grams with logistic regression. Recently, Badjatiya *et al.* [13] reported F1 score of 0.93 using deep neural networks on the same data set.

Academic research on bullying was started by social scientists and psychologists with a special focus on adolescents [14], [15], [16]. Similarly, social studies on cyberbullying predate computational endeavors. Cyberbullying has three definite characteristics [14] borrowed from traditional bullying [17] - intentional harm, repetitiveness and power imbalance (e.g., anonymity in the Internet) which differentiates it from other forms of online attacks. Vandebosch and Cleemput [18] give a detailed analysis of cyberbullies, their victims and bystanders based on self reported experience of bullying, cyberbullying and Information and Communication Technology use by school children. Dinakar *et al.* [19] 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 cyber-aggression, Hosseinmardi *et al.* [21] 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 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.

Trolls disrupt meaningful discussions in online communities by posting irrelevant and provocative comments. Cheng *et al.* [22] contrast traits of users banned by moderators to users who are not banned in news websites. They observe differences in the quality of comments, number of replies received and use of positive words for the two groups. A classifier trained on such features in one community is also able to perform well in another. Cheng *et al.* [23] equate flagging of comments by community as instances of trolling and discover that a significant portion of users have very low flagged content earlier. They suggest that an ordinary user can behave like a troll depending on the mood of the user and the context of the discussion. Tsantarliotis *et al.* [24] introduce troll vulnerability metrics to predict likelihood of a post being trolled.

Personal attack is less rigorously defined and often holds all of the above categories in it. Such attacks can be directed towards the author of a previous comment or a third party. Sood *et al.* [25] show that using two classifiers- one for object of insult (previous author or third party) identification and another for insulting comment identification, boosts the overall accuracy of the system. A recent work [5] reports classification of personal attacks on Wikipedia author pages with accuracy comparable to annotation by a group of three human annotators.

In comparison with all of the above mentioned work, in this paper we study shaming comments on Twitter, which are part of a particular shaming event and hence they are related. Furthermore, when we consider a shaming event, the focus lies on a single victim. All the comments which are of interest should invariably be about that particular victim. Other comments are ignored. Most of the previous work 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.

3 Categorization of Shaming Tweets

After studying more than one thousand shaming tweets from eight shaming events on Twitter, we have come up with six categories of shaming tweets as shown in Table 2. A brief description of these categories along with their most common attributes is given below.

(a) **Abusive (AB)**
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 is a strong marker of this type of shaming.

(b) **Comparison (CO)**
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's[4] 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 becomes 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, etc.

TABLE 2: 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/YpcoKEPbIu Phil Robertson Vs. Gays Vs. Justine Sacco
Passing judgment (PJ)	CF	... Chris Filardi should be put down in the name of science to see what compels 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 - wonder why Hamish doesn't receive the...

A static database would be of little use as public perception about an entity can change frequently.

(c) **Passing Judgment (PJ)**
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.

(d) **Religious/Ethnic (RE)**
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, Jewish. Maligning any one of these group identities in reference to the victim constitutes a religious/ethnic shaming. In this work, we assume that we know the group identities to which a victim associates. For example, Justine Sacco is a US 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 (e.g., [26] uses a third party service called Face plus plus [27]).

(e) **Sarcasm/Joke (SJ)**
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 of [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, e.g.,

> "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, sudden change of sentiment, etc., are important attributes of this category.

(f) **Whataboutery (WA)**
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 category of comments are use of WH adverbs and past form of verbs.

4. Colonel Santosh Mahadik of the Indian army was killed in a terrorist encounter

It is worthwhile mentioning that in a work-in-progress version of this study published as a poster paper [29], we categorized shaming into ten broad categories including the six described above. However, after a more detailed scrutiny, in this work 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, etc.

4 Automated Classification of Shaming Tweets

Our goal is to automatically classify tweets in the aforementioned six categories. In Fig. 1, the main functional units involving automated classification of shaming tweets are shown. Both labeled training set and test set of tweets for each of the categories go through the pre-processing and feature extraction steps. The training set is used to train six support vector machine (SVM) classifiers. The precision scores of the trained SVMs are next evaluated on the test set. Based on these scores, the classifiers are arranged hierarchically. A new tweet, after pre-processing and feature extraction, is fed to the trained classifiers and is labeled with the class of the first classifier that detects it to be positive. A tweet is deemed non-shame if all the classifiers label it as negative.

We discuss the three steps of pre-processing, feature extraction and classification in detail below.

4.1 Pre-processing

We perform a series of pre-processing steps before feature extraction and classification is done. Named entity recognition, co-reference resolution and dependency parsing are performed using the Stanford CoreNLP library [30]. All references to victims including names or surnames preceded by salutations, mentions, etc., are replaced with a uniform victim marker after the dependency parsing step. We also remove user mentions, retweet marker, hashtags, URLs from the tweet text after dependency parsing and before parts of speech tagging with

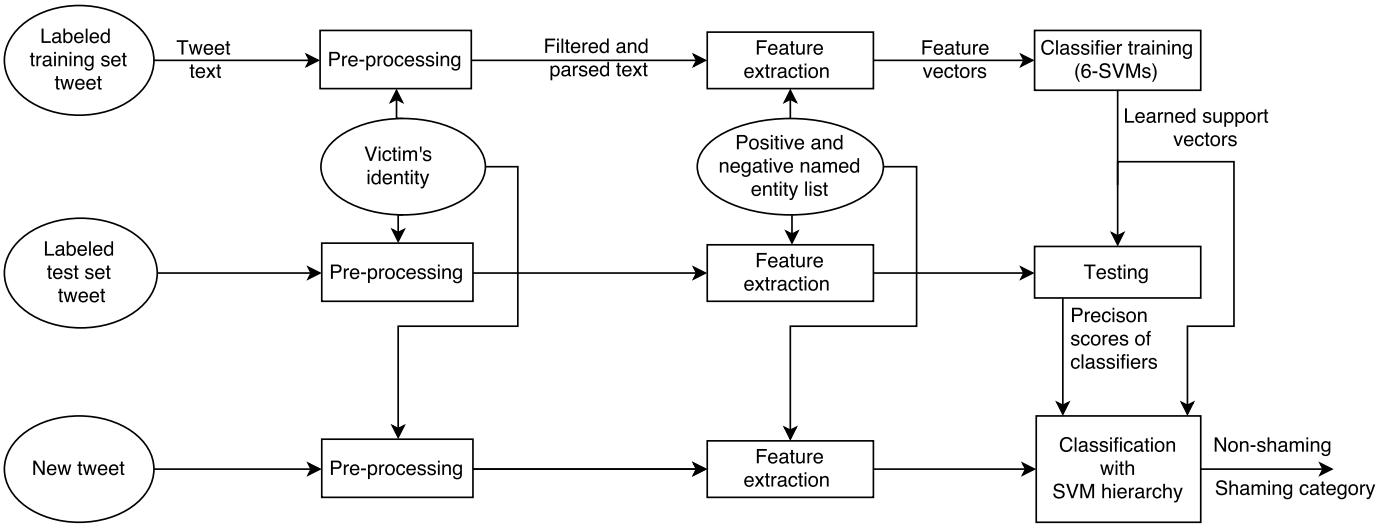

Fig. 1: Block diagram for shaming detection

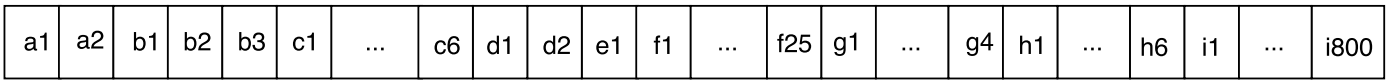

Fig. 2: Structure of the feature vector. a1 and a2: negative and positive words, b1 - b3: abusive, negative and positive association, c1 - c6: named entity associations, d1 and d2: authority, e1: group identities, f1 - f25: POS and others, g1 - g4: emojis, h1 - h6: sentiment features, i1 - i800: Brown cluster uni-grams

Stanford CoreNLP. If the event considered is a past event, current news source or search engine results would not be good indicators of a mentioned entity's polarity in that period. For those, a list is constructed based on historical news related to the mentioned entities. For recent events, search engine results can be relied upon.

4.2 Feature Extraction

We take into account a variety of syntactic, semantic and contextual features derived from the text of a tweet. The overall structure of the feature vector is given in Fig. 2. In the figure, a feature is represented by an index containing a letter followed by a number. Similar features are grouped together and they share a common letter in their indexes. The original features (with their respective indexes in parentheses) are described next with the help of the following example tweet from the event TH.

> "Boris Johnson is an embarrassing Roderick Spode wannabe, and his comments on Tim Hunt are even stupider than Hunt's original remarks."

This tweet belongs to the comparison shaming category.

Hereafter, by presence of a feature, we mean the feature value is in binary. Similarly, count of a feature is in integer while proportions are in floating point numbers.

(a) **Negative and positive words (a1 - a2)**
Shaming comments tend to contain more negative words than non-shaming ones do. Proportion of negative (a1) and positive words (a2) to all words in a tweet are taken as features. We use negative and positive words lexicon provided by Hu and Liu [31]. In the example tweet above, negative word count is 2 ('embarrassing' and 'stupider') which is divided by 21 (number of tokens separated by space) to give a value of 0.095 for a1. As there are no positive words in the tweet, the value of a2 is 0.

(b) **Abusive, negative and positive association (b1 - b3)**
We consider presence of negative (b1), positive (b2) and abusive (b3) words directly associated with the victim found from dependency relation as features. This additional information helps reduce the number of false negative decisions by the classifiers. In the example tweet above, there are no associations of the victim with abusive, negative or positive words. Thus, b1, b2 and b3 are set to false.

(c) **Association with named entities (c1 - c6)**
Mention of named entities (NE) other than the victim in a tweet is a good indicator of comparison shaming. To handle this, a list of NEs with their polarities (negative, neutral or positive) is used. Any NE which is not present in the list is also considered to be neutral. Count of mentions of these three polarities, i.e., number of positive mentions (c1), neutral mentions (c2) and negative mentions (c3) are used as features. Additionally, we use direct association of negative/positive words with NEs to get the number of

TABLE 3a: Feature values for the comparison shaming example tweet

Features	a1	a2	b1	b2	b3	c1	c2	c3	c4	c5	c6	d1	d2	e1	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10
Feature value	0.10	0	F	F	F	0	1	1	0	0	F	F	F	0	0.08	0.04	0	0.04	0.08	0.29	0	0.04	0	0.04

f11	f12	f13	f14	f15	f16	f17	f18	f19	f20	f21	f22	f23	f24	f25	g1	g2	g3	g4	h1	h2	h3	h4	h5	h6
0.04	0	0	0	0	0	0	0	0.04	0	0	0	0	1	0	F	F	0	0	1	0	0.24	0.74	0.02	0

i11	i12	i83	i97	i319	i347	i381	i437	i442	i468	i470	i473	i528	i530	i541	i574	i620	i650	i768
T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T

TABLE 3b: Feature values for an abusive shaming tweet

Feature values for the following abusive shaming tweet from the event JS are shown: *"This Justine Sacco is such a dumb bitch! SMH Uhh!!!"*. In this tweet, there is a dependency relation between the victim and the word 'bitch'. This word appears in both of our abusive words list and negative words list. Thus, b1 and b3 are set to true.

Features	a1	a2	b1	b2	b3	c1	c2	c3	c4	c5	c6	d1	d2	e1	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10
Feature value	0.2	0	T	F	T	0	0	0	0	0	F	F	F	0	0.17	0	0	0.08	0	0.33	0	0	0	0

f11	f12	f13	f14	f15	f16	f17	f18	f19	f20	f21	f22	f23	f24	f25	g1	g2	g3	g4	h1	h2	h3	h4	h5	h6
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.08	F	F	0	0	1	0	0.36	0.64	0	0

i54	i76	i85	i86	i303	i384	i437	i437	i531	i691	i721	i796
T	T	T	T	T	T	T	T	T	T	T	T

TABLE 3c: Feature values for a sarcasm/joke shaming tweet

Feature values for the following sarcasm/joke shaming tweet from the event MT are shown: *"Download the Melania Trump Pandora station. A mixture of 90s hip hop, 80s R&B, 70s Soul, 60s Rock and Roll, 50s Doo Wop, and country!* ☺☺*"*. Here, we observe that the overall sentiment (h1) of the tweet is 3 (i.e., positive) and it ends with two happy emojis (g1 equals true and g3 is set to 2). Both of these are indicative of the sarcasm/joke category.

Features	a1	a2	b1	b2	b3	c1	c2	c3	c4	c5	c6	d1	d2	e1	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10
Feature value	0.04	0.08	F	F	F	0	1	0	0	0	F	F	T	0	0	0	0	0.22	0.03	0.22	0	0	0	0

f11	f12	f13	f14	f15	f16	f17	f18	f19	f20	f21	f22	f23	f24	f25	g1	g2	g3	g4	h1	h2	h3	h4	h5	h6
0	0	0	0	0.03	0	0	0	0	0	0	0	0	0	0	T	F	2	0	3	0	0	0.75	0.25	0

i5	i83	i85	i92	i151	i179	i80	i440	i468	i470	i471	i531	i532	i564	i586	i616	i619	i680	i736	i754	i760
T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T

implied positive and negative mentions (c4 and c5) in a comment. Presence of direct association of an NE with the victim (by 'and', 'or', etc.) (c6), which is a stronger indicator of comparison as opposed to a mere presence of the NE, is taken as a feature. For the example tweet, the NE recognizer correctly identifies 'Boris Johnson' and 'Roderick Spode' as persons other than the victim. The first one is included in the NE list as a negative mention setting c3 to 1. c2 is also set to 1 as the second one is not present in the list. Values of c4 and c5 are both 0, as there are no dependency relationships between the mentioned entities and positive/negative words. 'Tim Hunt' is not directly associated with any of the NEs. So, c6 is set to false.

(d) **Authority (d1 - d2)**
Presence of a dependency relationship between the victim and certain auxiliary verbs, such as 'should', 'must' and 'ought' (d1), and tweet starting with a verb (d2) usually indicate authority, which is a feature of shaming utterances. d1 and d2 are set to false as these features are not present in the above mentioned tweet.

(e) **Group identities (e1)**
The victim's collective identities like religion, race, color, etc., are used to determine the count of negative words associations with these identities (e1), which is a strong indicator of religious/ethnic shaming. There are no negative word associations with Tim Hunt's collective identities. So, the value of e1 is set to 0 for the example tweet.

(f) **Parts of speech (POS) and others (f1 - f25)**
Proportion of POS tags in a tweet varies depending on the nature of the utterance, e.g., use of first and second person pronouns is more probable for subjective comments than objective ones. Shaming comments are primarily subjective in nature. The proportion of number of occurrences of a POS tag to all tokens is taken as a feature. We use the following tags from the Penn treebank [32] tagset- *JJ, JJR, JJS, NN, NNS, NNP, NNPS, POS, PRP, PRP$, RB, RBR, RBS, UH, VB, VBD, VBG, VBN, VBP, WDT, WP, WP$, WRB* (f1 to f23). Additionally, we consider the number of sentences (f24) and number of capital words (f25) in a tweet, which implies emphasis, as features. The values

of features from f1 to f23 are the number of each POS tag count divided by 21. The example tweet has a single sentence and there are no capital words. Hence, the value of f24 is 1 and f25 is 0.

(g) **Emojis (g1 - g4)**
Emojis constitute a popular means for expressing emotions. We divide common human face emojis in two groups, namely, happy and sad. Use of emojis from both the groups is often an indicator of sarcasm/Joke. Presence of happy (g1) and sad emojis (g2) along with count of those (g3 and g4) are used as features. These features are absent in the example tweet.

(h) **Sentiment features (h1 - h6)**
It is intuitive to assume shaming utterances to be in negative side of sentiment scale except in case of sarcasm/joke. We take the whole tweet sentiment (h1), which is an integer from 0 to 4, for five sentiment classes of very negative to very positive as a feature. For sarcasm/joke, the change of sentiment in a single tweet is also an important marker. So, we consider the proportion of non leaf nodes belonging to each of the five sentiment categories (h2 to h6) in the parse tree as features [33]. Sentiment of the example tweet is negative giving h1 a value of 1. Most of the non-leaf nodes in the parse tree of the example tweet are of neutral sentiment followed by negative sentiment.

(i) **Brown cluster uni-grams (i1 - i800)**
A typical tweet contains too few tokens from a huge vocabulary (comprised of dictionary words, hashtags, URLs, mentions, etc.) to create direct uni-gram features from it. As the resulting feature vector would be of very large dimension and sparse. To compensate for that, we use Brown cluster (a hierarchical clustering of words) uni-grams as features [34]. We consider a Brown cluster uni-gram list having 800 clusters (i1 to i800) produced from a corpus of about 6 million tweets [35]. It may be noted that, after tokenization, the given tweet produces 24 tokens including 2 punctuation marks (a comma and a period) and a special ''s' (from the word 'Hunt's'). However, ''s' is missing from the clusters and some tokens are from common clusters. For example, 'Borris', 'Roderick' and 'Tim' are from cluster index 12 while 'comments' and 'remarks' are from cluster index 650. The token 'Hunt' appears twice. Thus, only 19 cluster indexes out of the 800 have true values set for this particular tweet.

Considering all the above feature types, there are a total of 849 features (i.e., 800 uni-grams plus 49 other features described above), all derived from the texts of the tweets. The values of the features for the example tweet are shown in Table 3a. For Brown cluster uni-grams (i1 to i800), only the cluster indexes which have true value are shown. 'T' and 'F' in the table denote True and False values (1 and 0 in the feature vector), respectively. Tables 3b and 3c show feature values (rounded off to two places of decimal) for another two shaming tweets belonging to abusive and sarcasm/joke category, respectively.

4.3 Classification using Support Vector Machine (SVM)

Shaming classes are often found to be inherently overlapping, e.g., a comment is both RE and AB when it abuses a victim's ethnicity. For the sake of simplicity, we categorize each comment in only one class. Six one-vs.-all SVM classifiers [36] for each shaming category are constructed. While training a classifier, shaming comments from all other categories along with non-shame comments are treated as negative examples. Based on test set precision, the classifiers are arranged hierarchically placing one with higher precision above one with lower precision. The abusive classifier which has the highest precision (shown in Table 5) is placed on top.

For classification we use SVM with linear kernel from the java-ml library [37]. Linear kernel is chosen since it is known to perform well in classifying text data and is faster than nonlinear kernels. Equal number of tweets are sampled from all the shaming categories and the non-shaming category for each of the six classifiers to get balanced positive and negative examples in the training dataset.

5 Experimental Results

A large number of tweets belonging to a diverse set of shaming events occurring over years were collected using the Twitter 1% stream, Twitter search API and Topsy API (defunct at present). These were annotated by a group of annotators, who were instructed to label a tweet in one of the six shaming categories or label it as non-shaming (NS). Details of the collected shaming events are given in Table 4. In the table, '#Annotated' is the number of tweets manually labeled for each event. Note, for events LD, MT and PC, we do not have any annotated data. '#Unique tweets' is the number of collected unique tweets for an event. We do not include retweets explicitly in the dataset since a retweet is given the label of the original tweet.

5.1 Classification Performance

Performance scores for the six classifiers are shown in Table 5. True positive rate (TPR), true negative rate (TNR), false positive rate (FPR), false negative rate (FNR) and precision in percentage are reported in the table. Five-fold cross validation was performed for reporting this performance result. From the table, it is observed that the abusive shaming classifier has the highest precision and sarcasm/joke classifier has the lowest precision, which is consistent with our expectations.

As mentioned earlier, shaming categories are overlapping. It is, therefore, interesting to know which proportion of comments from a particular category is likely to get classified in other categories, i.e., labeled positive by a wrong classifier. This is illustrated in Table 6. In the first row of the table, out of the 319 manually annotated AB shaming category tweets, when each one is presented to all the trained classifiers one after another, the AB-classifier correctly outputs positive for 274 tweets, CO-classifier wrongly labels 12 tweets as positive, and so on. Finally, 20 tweets get negative labels from all the six classifiers, thus wrongly deciding these to be non-shaming. We observe that for all categories, a significant number of false

TABLE 4: Detailed breakup of tweets used for experiment

Events	JS	TH	CF	AK	HM	LD	MT	PC
#Annotated	453	306	18	407	44	none	none	none
#Unique tweets	29612	23696	100	5026	366	23472	179551	1644

TABLE 5: Performance of individual classifiers

Classifier Type	TPR%	TNR%	FPR%	FNR%	Prec%
Abusive (AB)	85.89	94.67	5.33	14.11	88.96
Comparison (CO)	81.96	92.78	7.22	18.04	85.02
Passing judgment (PJ)	69.49	86.00	13.98	30.51	71.30
Religious/Ethnic (RE)	77.33	92.00	8.00	22.67	82.86
Sarcasm/Joke (SJ)	60.00	83.75	16.25	40.00	64.86
Whataboutery (WA)	72.62	88.10	11.90	27.38	75.31

TABLE 6: Inter-category misclassification for individual classifiers

Shaming Type	AB	CO	PJ	RE	SJ	WA	NS
Abusive (AB)	274	12	17	18	25	14	20
Comparison (CO)	6	158	11	8	15	11	18
Passing judgment (PJ)	8	15	171	27	42	45	28
Religious/Ethnic (RE)	4	3	19	58	14	16	3
Sarcasm/Joke (SJ)	3	5	12	7	75	7	29
Whataboutery (WA)	1	4	14	13	9	61	12

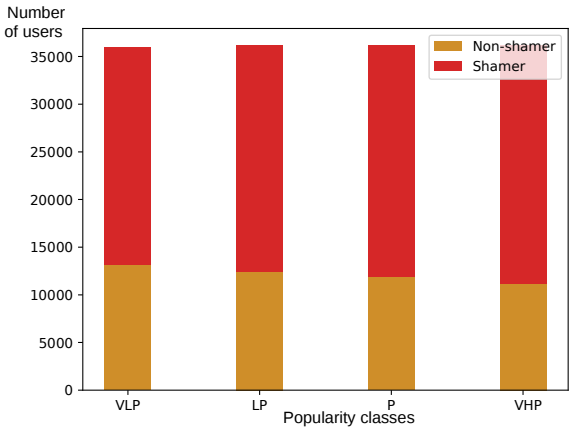

Fig. 3: Number of shamers and non-shamers in quartiles

negative decisions would end up in passing judgment category (these can also go to non-shame but only after the PJ-classifier outputs a negative label). This validates our decision to instruct annotators to label a tweet as PJ only when it does not fall in any other category but it is an instance of shaming. AB tweets have almost uniform tendency to get classified positive by other classifiers thus indicating that abusive words are used uniformly across all other categories. Sarcasm/joke and whataboutery comments are most often confused with non-shaming. This reflects the inherent difficulty in distinguishing these two categories from non-shaming when contextual information is limited or worse, absent.

After hierarchical arrangement, the precision and recall scores for the classifiers are given in Table 7. The final system has overall precision and recall scores of 72.69 and 88.08, respectively.

From the classified tweets, we have access to a large set of shamer and non-shamer users. The question we ask at this point is that whether these two categories of users are inherently different from one another. Also, there are two types of shamers: active- those who write an original shaming tweet, and passive- those who only retweet a shaming tweet (similar to bullies and bystanders in [18]).

The major findings of our work are given below.

TABLE 7: Hierarchical classification performance

Classifier Type	Precision%	Recall%
Abusive (AB)	80.89	92.20
Comparison (CO)	71.81	87.40
Passing judgment (PJ)	70.40	47.68
Religious/Ethnic (RE)	40.00	77.63
Sarcasm/Joke (SJ)	67.07	48.67
Whataboutery (WA)	34.19	25.32

5.2 Popularity and shaming

Follower count is an important indicator of a user's popularity (there can be others, e.g., number of retweets, likes his/her tweets get, etc.). Our event dataset contains a diverse set of users with respect to popularity having follower count ranging from zero to a few millions. To compare the tendency of shaming among these users, we divide them in equal size quartiles based on follower count- from very low popular (VLP) to very high popular (VHP). The intuition behind this is that, there are different classes of users in every OSN as also in real society in terms of popularity. For example, a celebrity or politician's Twitter attributes (like follower count, status count, etc.) are very unlikely to match that of a commoner. We observe in Fig. 3 that the number of shamers to that of non-shamers is almost double for each quartile increasing marginally with popularity. However, this small increase is due to the fact that in many cases, users have multiple comments and we mark them as shamers if any one of those is a shaming comment. Popular users are likely to comment more and they comment on multiple events increasing their chance of being labeled as shamers.

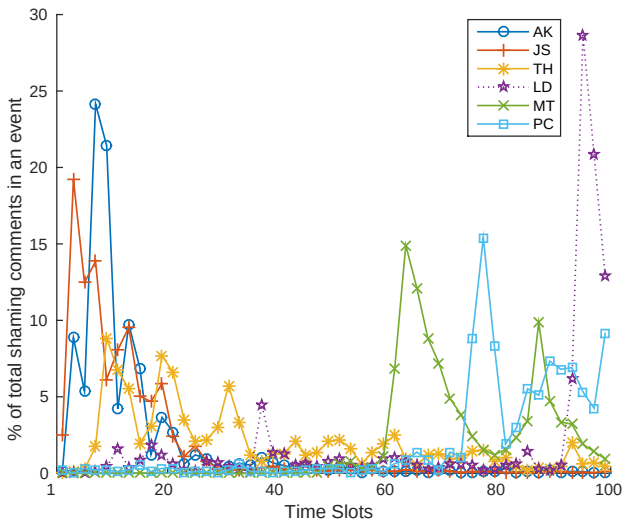

Fig. 4: Distribution of shaming comments with time

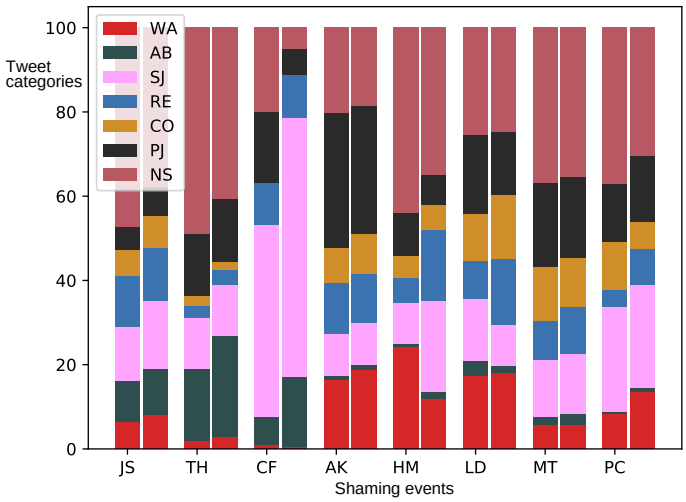

Fig. 5: Relative distribution of tweets and retweets in categories

TABLE 8: Avg. followers per month in popularity categories

Popularity(#followers range)	Shamer	FPM
VLP (0-179)	Yes	1.67
VLP (0-179)	No	1.62
LP (180-573)	Yes	6.08
LP (180-573)	No	5.87
P (574-1969)	Yes	17.41
P (574-1969)	No	16.27
VHP (1970-)	Yes	760.29
VHP (1970-)	No	495.81

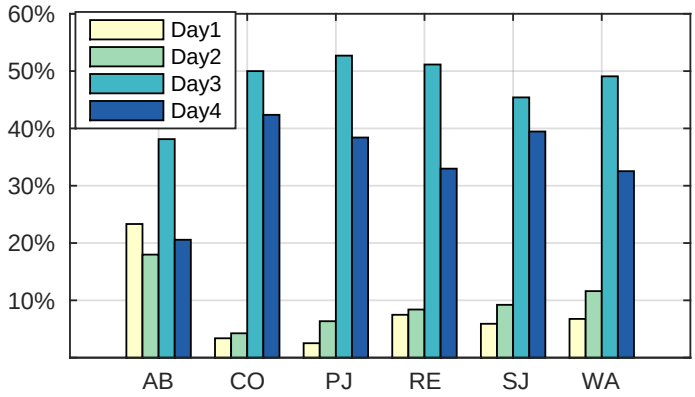

Fig. 6: Change in distribution of shaming categories across all events

5.3 Rewards for shamers

Negative discourse like public shaming also signifies emotional attachment and engagement of the users with the Twitter ecosystem. Hence, it is relevant to ask whether shamers get rewarded or not by such behavior. In this context, we define followers per month (FPM) to be the number of followers divided by the number of months spent in Twitter by a user. The intuition behind this is that a user who has acquired more followers than another user in the same period of time posts more engaging and interesting comments. Are shaming comments one of those? Comparing shamers with non-shamers, we find that the average FPM is 204 for shamers while it is only 119 for the latter. In Table 8, we list FPMs for shamers and non-shamers of the four classes separately. In all the popularity classes, shamers acquire more followers per month than the non-shamers do. Note, '#followers range' in parenthesis is the range of follower count for each quartile.

5.4 Dynamics of Shaming Events

In a bid to study the dynamics of shaming events, it was noted that their durations vary over a wide range. For ensuring a uniformity in representation, the entire duration for each event is divided into 100 time slots and the percentage of shaming comments (i.e, tweets and retweets) posted in each of these time slots for that particular event is plotted in Fig. 4. It may be observed from the figure that each of the six events has one major peak and several minor peaks. This indicates that the rate of shaming in Twitter is not uniform and usually occurs in bursts. Interestingly, only the events AK and LD, wherein both of the victims are popular television actors, have at least one prominent minor peak on the left of its major peak, i.e., smaller but significant bursts of shaming comments precede the major burst with respect to time.

Our chosen events are very diverse in terms of when these occurred, victim's profile and the nature of apparent violation of social norms. Despite these, in Fig. 5, we observe similarity in the distribution of tweets and retweets across all events. In the figure, proportions of tweet and retweet categories for the eight events are shown. For every two consecutive bars, the

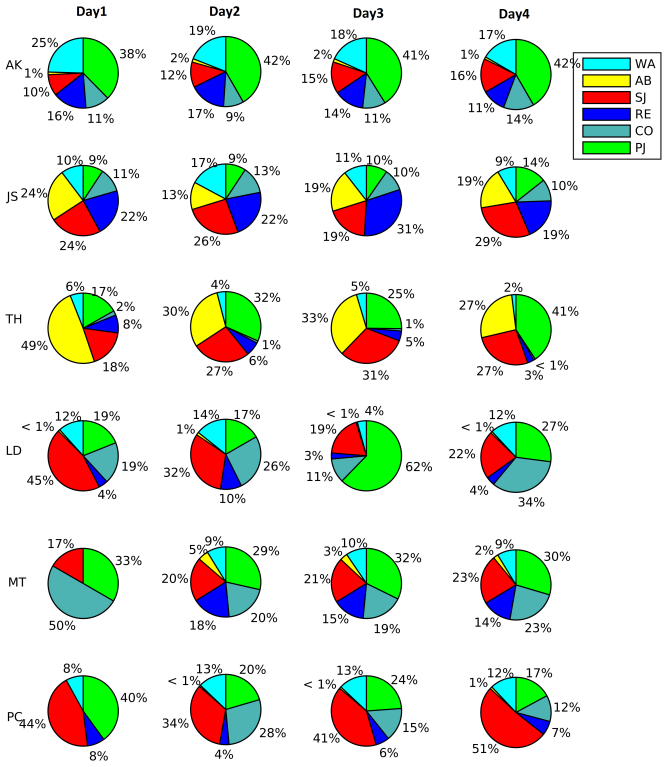

Fig. 7: Change in distribution of shaming categories for individual events

first bar denotes tweets and the second bar stands for retweets of an event. Though non-shame constitutes a major part of the bar, these are less likely to get retweeted. Sarcasm/jokes and passing judgments are popular means of shaming. Also, SJ tweets are very likely to get retweeted.

It was also observed that the distribution of the six categories of shaming tweets is not static and it changes over time as the shaming event progresses. Fig. 6 shows the proportion of posted shaming tweets in a category with respect to the total number of tweets in that category across all the shaming events. It is seen from the figure that all of the six categories peak on the third day and then goes down. However, the rise is not uniform. While the AB category rises moderately on the third day, the remaining five categories make big leaps from being very low on the first and second days. Thus implies that the abusive form of shaming of the victim starts early and its volume remains relatively steady as compared to the other types.

Fig. 7 shows this trend for six individual events over four days starting from the first shaming tweet's post date in our corpus. The remaining two events have too few number of shaming tweets to be divided into four days. As an event progresses, the share of SJ comments increases in most of the cases. We also notice that the share of RE comments for events JS and AK remains relatively larger for all days in comparison with other events. It may be concluded that the victim's original comment or action coupled with his or her social background have some influence on the type of shaming received. If the proportion of abusive comments are any approximation for the degree of outrage caused among Twitter users, then, in this respect, events JS and TH rank higher than the others.

6 Mitigation of Public Shaming in Twitter

There are two broad sets of controls available for users to counter inappropriate behavior in Twitter. The first consists of several tools for reporting tweets as well as accounts directly to Twitter for spam, harassment, abuse, etc. These measures are very effective in the sense that global actions can be taken by Twitter like deleting the offending tweet or even suspending the account of the offender altogether. However, the main problem with this approach is that action against a reported shaming tweet or account may take time. Twitter specifies the time to confirm the receipt of a report to be within 24 hours [38]. However, there is no commitment on the actual time needed to take action against the offender. As shaming events are viral in nature, delayed action would defeat any attempt aimed at protecting the victim.

The second set consists of three local controls, namely, 'mute'- which prevents tweets originating from the muted account from appearing in the user's feed, 'block'- which is similar to mute but it also unfollows/unfriends the blocked account and 'delete'- which deletes a direct message received by the user. Though limited in scope, these actions remove any tweet immediately from the victim's feed, thus, shielding him/her from shaming attacks.

Making use of the above-mentioned handles, we have designed an application named BlockShame [39] which proactively takes user defined actions (i.e., any one of the 'block', 'mute', 'delete' or none) for three kinds of interactions in Twitter, i.e., tweets, mentions and direct messages. Additionally, users have the freedom to choose certain shaming categories to be out of the purview of it.

The workflow of BlackShame includes the following steps:

(a) User authorizes BlockShame in Twitter from the application's website (see Fig. 8)
(b) User sets choice of actions along with (optionally) his/her group identities (see Fig. 9) for detecting and taking appropriate action on Religious/Ethnic type of shaming.
(c) User's recent tweets, mentions and direct messages are accessed from Twitter
(d) The obtained tweets are classified using pre-trained SVMs
(e) Actions are taken according to the choices set by the user in step (b)
(f) Steps (c) to (e) are repeated periodically at fixed short intervals until user revokes permission for BlockShame in Twitter

One of the ways to measure the effectiveness of a system like BlockShame is to count the average number of shaming tweets a shamer can post before he gets detected. To this end, we attempted to recreate a shaming event by directing a

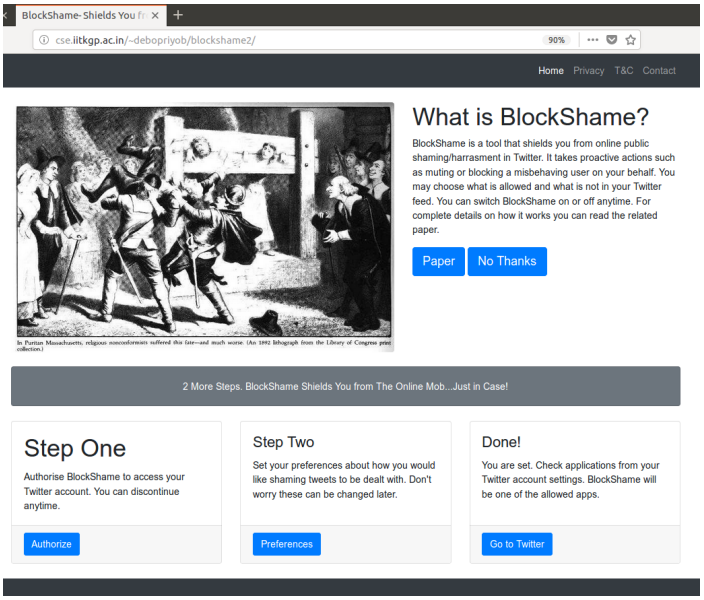

Fig. 8: BlockShame: home page

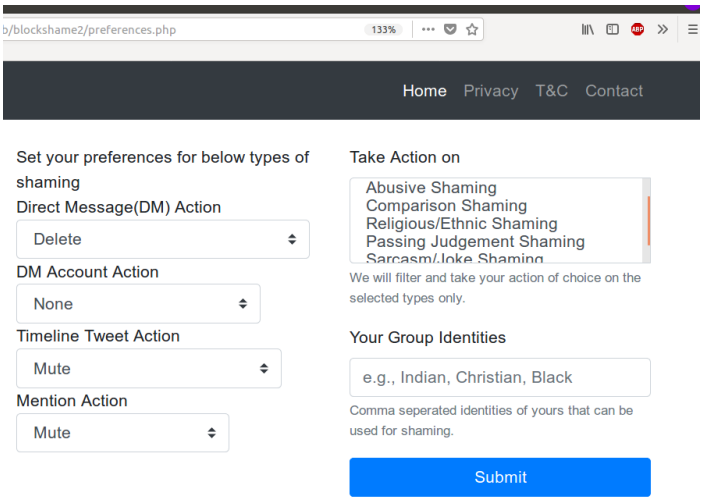

Fig. 9: BlockShame: setting preferred actions by users

set of withheld labeled shaming tweets to a Twitter account specifically created for this purpose. The account was made to subscribe to BlockShame. For the sake of this experiment, no action is actually taken on the shamer except for the fact that the sequence of labels predicted by BlockShame is stored. It may be noted that when a tweet is correctly classified as shaming, the shamer can be muted or blocked immediately. However, if a shaming tweet is miss-classified into non-shame, the victim can be potentially shamed by the same shamer again until he gets detected in one of his later attempts. Keeping these facts in mind, we define a *detection block* to be a sequence of

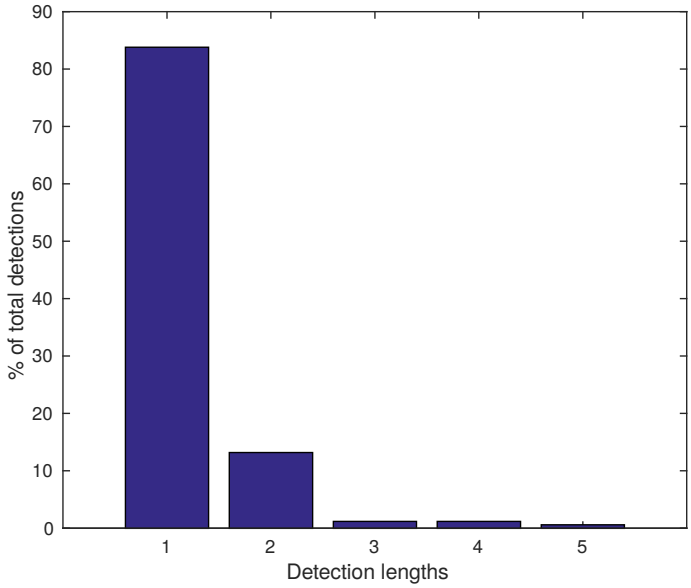

Fig. 10: BlockShame: number of tweets by shamers before detection

consecutive undetected shaming tweets followed by a single detected shaming tweet. *Detection length* is the number of tweets in a detection block. A detected shaming tweet which has no preceding undetected shaming tweet is of detection length one. For the exceptional case of one or more undetected shaming tweets appearing without a detected one, the detection length is taken to be the number of such tweets. From this perspective, the sequence of predictions by BlockShame for any shaming event can be viewed as a series of detection blocks, where each of the blocks corresponds to a shamer being detected. Fig. 10 shows the relative frequencies of detection lengths in percentage. It is observed that more than 80% of the detections blocks are of length 1 and about 13% are of length 2. This implies that a large majority of the shamers can be detected and action taken by BlockShame after their first two shaming posts. A negligible number of shamers remain undetected after their third shaming tweet.

After offending accounts have been muted or blocked by BlockShame, the victim may choose to report the accounts to Twitter for permanent action, if desired. The approach mentioned here can also be potentially deployed by Twitter itself for automating the process of taking appropriate action against repeated offenders.

7 Conclusion and Future Directions

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.

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 work has the potential for a user to choose to allow certain types of shaming comments (e.g., comments which are sarcastic in nature) giving her an opportunity for rebuttal, and block others (e.g., comments which 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 for another person.

Shaming is subjective in reference to shamers. For example, the same comment made by two different persons coming from different social, cultural or political background may have different connotations to the victim. We would like to include the attributes of the author of the comment as a contextual information when deciding if the comment is shaming or not. Moreover, in every event, we notice that after the initial outrage, the volume of apologetic or re-conciliatory comments gradually increases. A considerable proportion of users made multiple comments in a single event which contains both shaming and non-shame categories. We plan to investigate these behaviors further in future. The performance of individual classifiers are promising though there are scopes for improvement. We would like to repeat our experiments with an even larger annotated dataset to improve the performance further.

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.

[24] P. Tsantarliotis, E. Pitoura, and P. Tsaparas, "Defining and predicting troll vulnerability in online social media," *Social Network Analysis and 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," *Journal of the Association for Information Science and Technology*, 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," 2017.

[27] "Face++ cognitive services," retrieved February 20, 2018 from https://www.faceplusplus.com/.

[28] 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.

[29] 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.

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

[31] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 168–177.

[32] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of english: The penn treebank," *Computational linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[33] D. Bamman and N. A. Smith, "Contextualized sarcasm detection on twitter." in *ICWSM*, 2015, pp. 574–577.

[34] O. Owoputi, B. OConnor, C. Dyer, K. Gimpel, and N. Schneider, "Part-of-speech tagging for twitter: Word clusters and other advances," *School of Computer Science*, 2012.

[35] Brown-Clusters, "Twitter word clusters," retrieved July 24, 2017 from http://www.cs.cmu.edu/~ark/TweetNLP/.

[36] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[37] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[38] Twitter, "Report abusing behavior," retrieved February 07, 2018 from https://help.twitter.com/en/safety-and-security/report-abusive-behavior.

[39] "Blockshame shields you from the online mob...just in case!" retrieved February 07, 2018 from http://cse.iitkgp.ac.in/~rajesh.basak/blockshame/.

Rajesh Basak received his masters degree in Network and Internet Engineering from Pondicherry University, Pondicherry, India. Currently, he is a research scholar in the Department of Computer Science and Engineering, IIT Kharagpur, India. His research interests include online social networks and machine learning.

Shamik Sural is a professor in the Department of Computer Science and Engineering, IIT Kharagpur, India. He received the Ph.D. degree from Jadavpur University, Kolkata, India in 2000. He is a senior member of the IEEE and has previously served as the Chairman of the IEEE Kharagpur Section. He has published more than 200 papers in reputed international journals and conferences. A recipient of Alexander von Humboldt Fellowship for experienced researchers, his research interests include computer security and data science. He is an associate editor of the IEEE Transactions on Services Computing

Niloy Ganguly is a Professor at the Department of CSE, IIT Kharagpur, where he leads the Complex Networks Research Group (see http://cse.iitkgp.ac.in/resgrp/cnerg/). He has received his B.Tech. from IIT Kharagpur and his Ph.D. from BESU Shibpur, India, and worked as a post-doctoral fellow in Technical University, Dresden, Germany. His research interests are in complex networks, social networks, peer-to-peer networks, and information retrieval.

Soumya K Ghosh received the Ph.D. and M.Tech. degrees from the Department of Computer Science and Engineering, Indian Institute of Technology (IIT), Kharagpur, India. Presently, he is a Professor with the Department of Computer Science and Engineering, IIT Kharagpur. Before joining IIT Kharagpur, he worked for the Indian Space Research Organization. His research interests include spatial data science, spatial web services, and cloud computing. He has more than 200 research papers in reputed journals and conference proceedings. He is a member of the IEEE.

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.

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 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?'

We also identify several interesting discriminating user and tweet features related to shaming tweets.

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-

Table 2: Different forms of shaming tweet

Shaming Type	*Event*	*Example Tweet*
Whatabouterism (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?

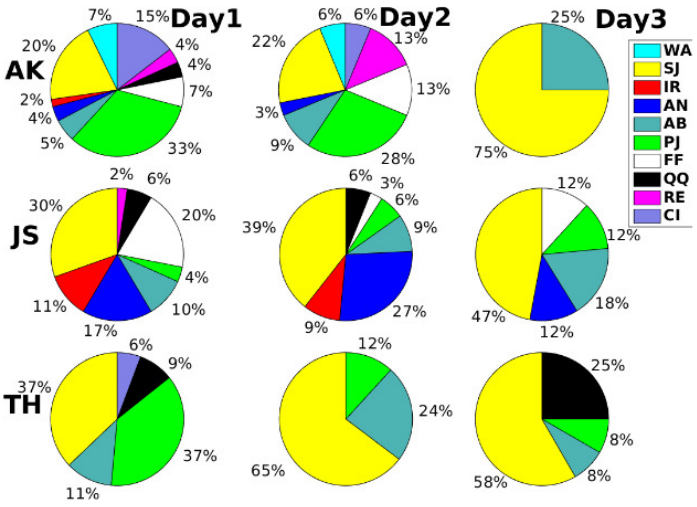

Figure 1: Shaming types for the first three days

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.