Online Public Shaming on Twitter: Detection, Analysis, and Mitigation

Rajesh Basak, Shamik Sural, *Senior Member, IEEE*, Niloy Ganguly, and Soumya K. Ghosh, *Member, IEEE*

***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.**

***Index Terms*— BlockShame, online user behavior, public shaming, tweet classification.**

I. 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: 1) a definite single target or victim; 2) an action committed by the victim perceived to be wrong; and 3) 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 I, we have listed the victim, year in which the event took place, action that triggered public shaming along with the triggering medium, and its immediate consequences for each studied event. "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 (see [2]–[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 victim's perspective. We consider a comment to be shaming only when it criticizes the target of the shaming event. For example, while "Justine Sacco gonna get off that international flight and cry mountain stream fresh white whine tears b" is an instance of shaming, a comment like *"Just read the Justine Sacco story lol*

Manuscript received March 2, 2018; revised October 25, 2018; accepted January 21, 2019. *(Corresponding author: Shamik Sural.)*

The authors are with the Department of Computer Science and Engineering, IIT Kharagpur, Kharagpur 721302, India (e-mail: rajesh@sit.iitkgp.ac.in; shamik@cse.iitkgp.ac.in; niloy@cse.iitkgp.ac.in; skg@cse.iitkgp.ac.in).

Digital Object Identifier 10.1109/TCSS.2019.2895734

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 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 BlockShame that can be used by a Twitter user for blocking shamers.

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

II. Related Work

Efforts to moderate user-generated content in the Internet started very early. Smokey [2] is one of the earliest computational works in this direction which builds a decision tree classifier for insulting posts trained on labeled comments from two web forums. Although academic research in this area started that early, it used different nomenclatures 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 homoglyph (characters that 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 (see [6]). Warner and Hirschberg [8] attempt to identify hate speech targeting Jews from a data set consisting of Yahoo! comments and known antisemitic web page contents. A similar type of work has been done on antiblack hate speech on Twitter [9]. Burnap and Williams [10] collected tweets for 2 weeks after the Lee Rigby incident [11] and trained a classifier on typed dependence and hateful terms as features. Waseem and Hovy [12] released a public data set of 16 000 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]–[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 used 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 the specific domain of interests, e.g., Lesbian, Gay, Bisexual, and Transgender (LGBT), cyberbullying, to detect comments that deviate from real-world beliefs and is a good indicator of subtler forms of bullying. For instance, asking a male which beauty salon he visits can be a case of bullying as OMCS tells that beauty salons are more likely to be associated with females. In addition, 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 15 comments of which more than 40% contain at least one profane word, to account for repetitiveness of bullying. Their best performing classifier uses unigram and trigram text features with image category (e.g., person, car, nature, etc.) and its metadata 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 has 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.

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

TABLE I
EVENTS WITH TRIGGER AND CONSEQUENCES CONSIDERED IN THIS PAPER

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

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. 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, which boosts the overall accuracy of the system. A recent work [5] reports the classification of personal attacks on Wikipedia author pages with accuracy comparable to annotation by a group of three human annotators.

Compared 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 that are of interest should invariably be about that particular victim. Other comments are ignored. Most of the previous works mentioned above do not make a distinction between acceptability and nonacceptability of a comment based on the presence or absence of a predefined victim.

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 II. 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's[1] 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

[1]Colonel Santosh Mahadik of the Indian army was killed in a terrorist encounter.

TABLE II
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...

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.

IV. 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 preprocessing 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 preprocessing 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 nonshame if all the classifiers label it as negative.

We discuss the three steps of preprocessing, feature extraction, and classification in detail in the following.

A. Preprocessing

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

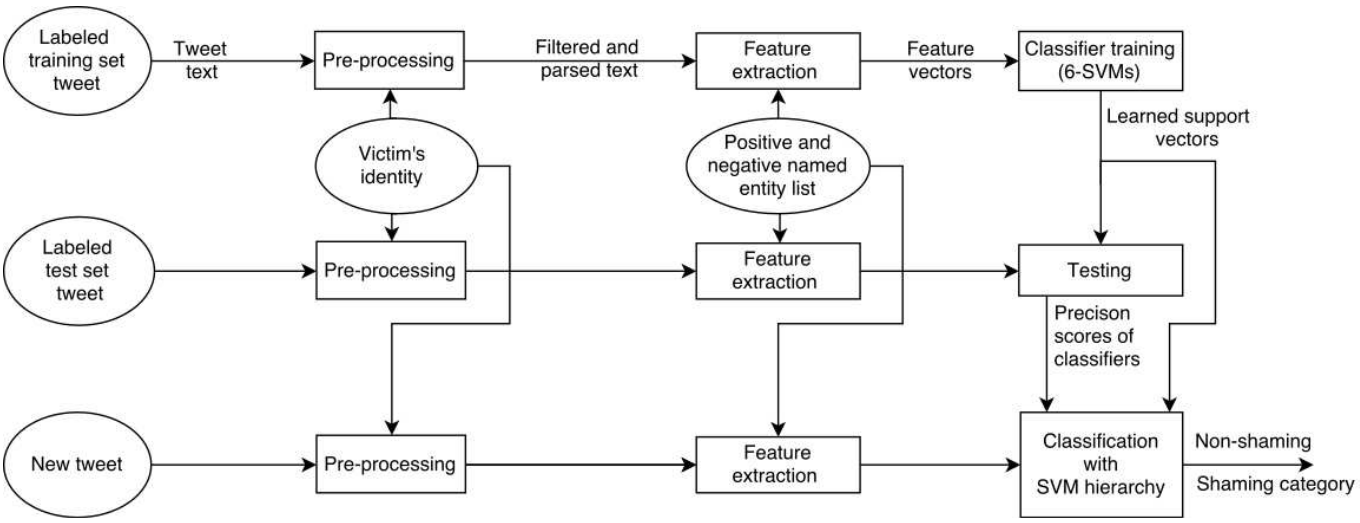

Fig. 1. Block diagram for shaming detection.

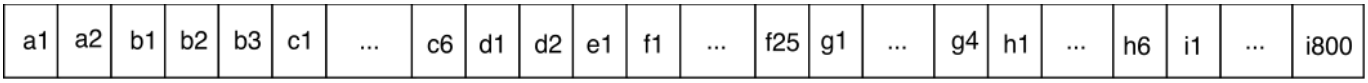

Fig. 2. Structure of the feature vector. a1 and a2: negative and positive words, b1–b3: abusive, negative and positive association, c1–c6: named entity associations, d1 and d2: authority, e1: group identities, f1–f25: POS and others, g1–g4: emojis, h1–h6: sentiment features, and i1–i800: Brown cluster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

B. 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 shown in Fig. 2. In this 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 the 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.

1) *Negative and Positive Words (a1–a2):* Shaming comments tend to contain more negative words than non-shaming (NS) 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 mentioned 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.
2) *Abusive, Negative, and Positive Association (b1–b3):* We consider the 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 to reduce the number of false negative decisions by the classifiers. In the example tweet mentioned above, there are no associations of the victim with abusive, negative, or positive words. Thus, b1, b2, and b3 are set to false.
3) *Association With Named Entities (c1–c6):* Mention of NEs 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 that 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. In addition, we use direct association of negative/positive words with NEs to get the number of 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. Therefore, c6 is set to false.
4) *Authority (d1–d2):* Presence of a dependency relationship between the victim and certain auxiliary verbs,

TABLE III

(A) FEATURE VALUES FOR THE COMPARISON SHAMING EXAMPLE TWEET. (B) FEATURE VALUES FOR AN ABUSIVE SHAMING TWEET. (C) FEATURE VALUES FOR A SARCASM/JOKE SHAMING TWEET

(a)

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

(b)

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

(c)

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

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.

5) *Group Identities (e1):* The victim's collective identities like religion, race, color, and so on, 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. Therefore, the value of e1 is set to 0 for the example tweet.
6) *Parts of Speech and Others (f1–f25):* Proportion of POS tags in a tweet varies depending on the nature of the utterance, e.g., use of the first and second person pronouns is more probable for subjective comments than objective ones. Shaming comments are primarily subjective in nature. The proportion of the 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$, RB, RBR, RBS, UH, VB, VBD, VBG, VBN, VBP, WDT, WP, WP$, and WRB* (f1 to f23). In addition, we consider the number of sentences (f24) and the 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.
7) *Emojis (g1–g4):* Emojis constitute a popular means for expressing emotions. We divide common human face emojis into 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.
8) *Sentiment Features (h1–h6):* It is intuitive to assume shaming utterances to be in a 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. Therefore,

TABLE IV
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

we consider the proportion of nonleaf 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 nonleaf nodes in the parse tree of the example tweet are of neutral sentiment followed by negative sentiment.

9) *Brown Cluster Unigrams (i1–i800):* A typical tweet contains too few tokens from a huge vocabulary (comprised of dictionary words, hashtags, URLs, mentions, etc.) to create direct unigram 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) unigrams as features [34]. We consider a Brown cluster unigram list having 800 clusters (i1 to i800) produced from a corpus of about 6million 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-mentioned feature types, there are a total of 849 features (i.e., 800 unigrams plus 49 other features described earlier), all derived from the texts of the tweets. The values of the features for the example tweet are shown in Table III(a). For Brown cluster unigrams (i1 to i800), only the cluster indexes that have true value are shown. "T" and "F" in the table denote True and False values (1 and 0 in the feature vector), respectively. Table III(b) and (c) shows feature values (rounded off to two places of decimal) for another two shaming tweets belonging to abusive and sarcasm/joke categories, respectively.

C. Classification Using Support Vector Machine

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-versus-all SVM classifiers [36] for each shaming category are constructed. While training a classifier, shaming comments from all other categories along with nonshame 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 that has the highest precision (as shown in Section V) 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 is sampled from all the shaming categories and the NS category for each of the six classifiers to get balanced positive and negative examples in the training data set.

TABLE V
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

V. 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 application programming interface (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 NS. Details of the collected shaming events are given in Table IV. 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 data set since a retweet is given the label of the original tweet.

A. Classification Performance

Performance scores for the six classifiers are shown in Table V. 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. Fivefold 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 VI. 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,

TABLE VI
INTERCATEGORY 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

TABLE VII
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

and so on. Finally, 20 tweets get negative labels from all the six classifiers, thus wrongly deciding these to be NS. We observe that for all categories, a significant number of false negative decisions would end up in passing judgment category (these can also go to nonshame 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 NS. This reflects the inherent difficulty in distinguishing these two categories from NS when contextual information is limited or worse, absent.

After hierarchical arrangement, the precision and recall scores for the classifiers are given in Table VII. 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 nonshamer users. The question we ask at this point is 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 in the following.

B. 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 data set contains a diverse set of users with respect to popularity having follower count ranging from zero to a few million. To compare the tendency of shaming among these users, we divide them into 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 nonshamers 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.

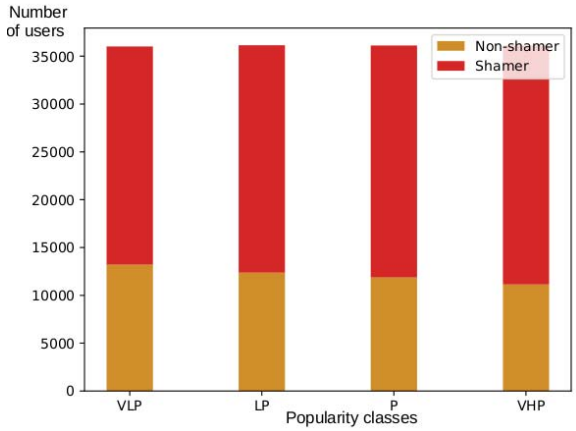

Fig. 3. Number of shamers and nonshamers in quartiles.

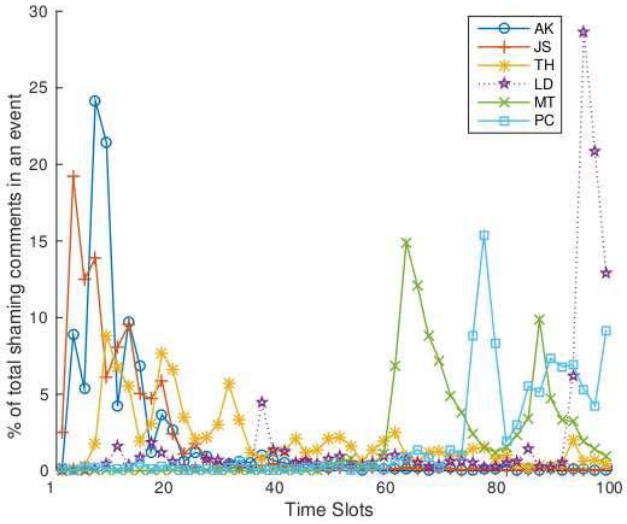

Fig. 4. Distribution of shaming comments with time.

C. 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.

TABLE VIII
AVERAGE FPM 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

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 nonshamers, we find that the average FPM is 204 for shamers while it is only 119 for the latter. In Table VIII, we list FPMs for shamers and nonshamers of the four classes separately. In all the popularity classes, shamers acquire more FPM than the nonshamers do. Note that "#followers range" in parenthesis is the range of follower count for each quartile.

While our analysis covers a fairly large number of users, yet further experiments are needed to claim that shaming increases follower count as a causal relationship. This is due to the fact that there could be other subtle characteristics of tweets (e.g., sense of humor) or even attributes of the user (e.g., education level) that can also potentially contribute to change in popularity.

Study of political polarization in Twitter has recently been gaining momentum [38]–[40]. Hong and Kim [40] show that among the members of the United States House of Representatives, all else being equal, those with more extreme views are likely to gather more Twitter followers than their moderate peers. Shaming can be construed as a form of extreme negative opinion as opposed to NS, which includes objective statements about the event as well as statements even supporting the victim. Thus, our results comparing average FPM for shamers and nonshamers point in a similar direction.

D. 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 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

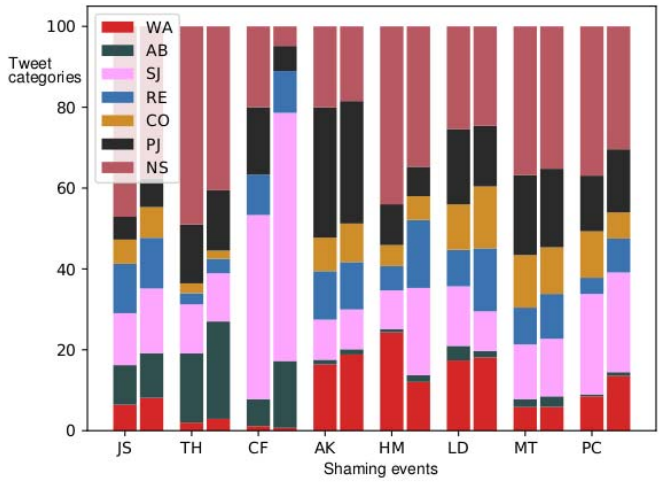

Fig. 5. Relative distribution of tweets and retweets in categories.

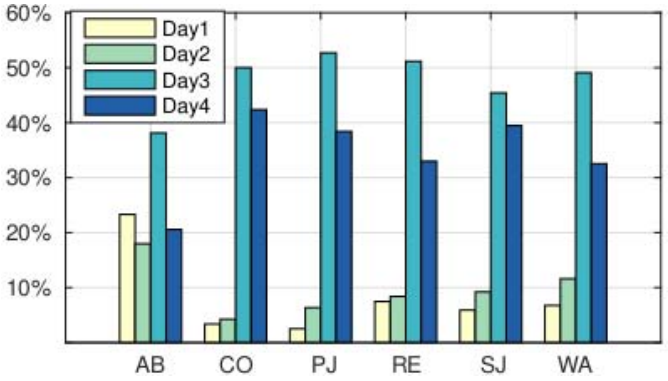

Fig. 6. Change in distribution of shaming categories across all events.

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 first bar denotes tweets and the second bar stands for retweets of an event. Although nonshame 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. This 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 4 days starting from the first shaming tweet's postdate in our corpus.

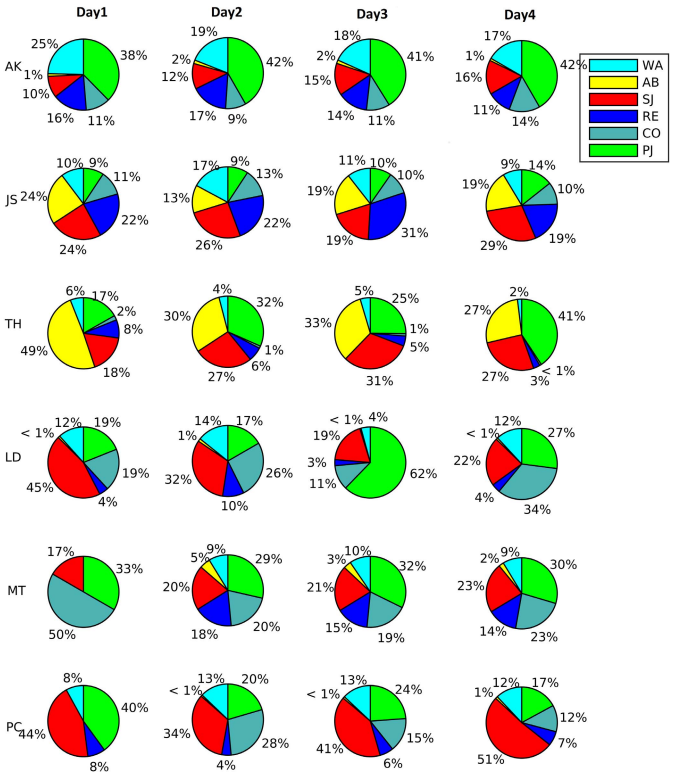

Fig. 7. Change in distribution of shaming categories for individual events.

The remaining two events have too few number of shaming tweets to be divided into 4 days. As an event progresses, the share of SJ comments increases in most of the cases. We also note 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.

VI. 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, and so on. 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 h [41]. 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 provided by Twitter API, namely, "mute" that prevents tweets originating from the muted account from appearing in the user's feed,

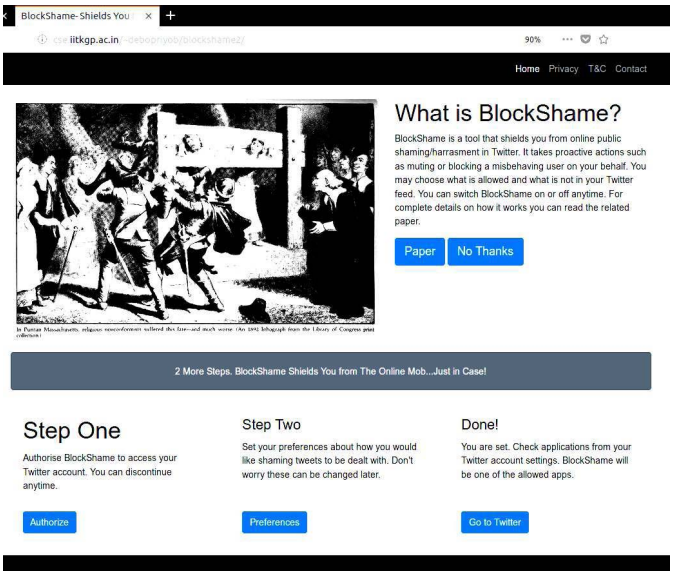

Fig. 8. BlockShame: home page.

"block" that is similar to mute but it also unfollows/unfriends the blocked account, and "delete" that deletes a direct message (DM) received by the user. Although limited in scope, these actions remove any tweet immediately from the victim's feed, thus, shielding him/her from direct shaming attacks. However, these tweets remain in the Twitter ecosystem to be viewed by users other than the victim.

Making use of the above-mentioned handles, we have designed an application named BlockShame [42] 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 DMs. In addition, 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.

1) User authorizes BlockShame in Twitter from the application's website (see Fig. 8).
2) 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.
3) User's recent tweets, mentions, and DMs are accessed from Twitter.
4) The obtained tweets are classified using pretrained SVMs.
5) Actions are taken according to the choices set by the user in step 2).
6) Steps 3)–5) 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 set of withheld labeled shaming tweets to a Twitter account specifically created for this purpose. The account was made

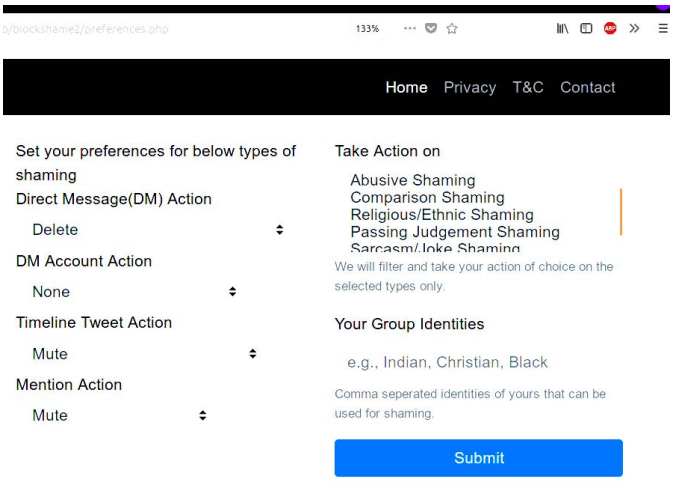

Fig. 9. BlockShame: setting preferred actions by users.

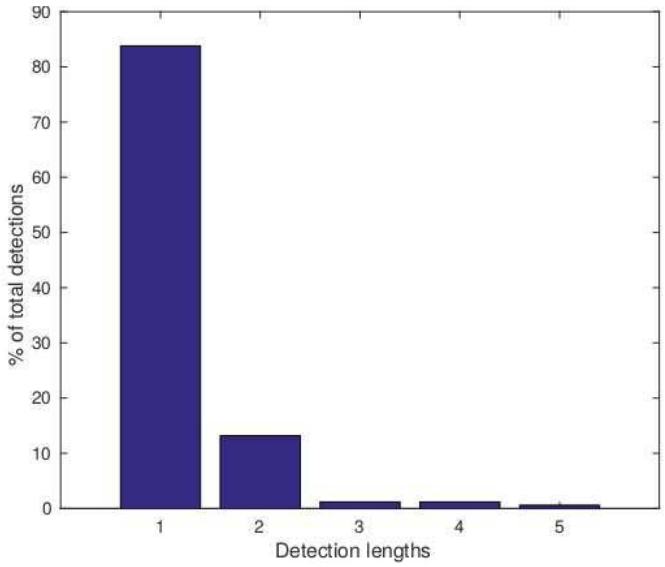

Fig. 10. BlockShame: number of tweets by shamers before detection.

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 missclassified into nonshame, 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 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 that 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 length1 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.

Once a tweet has been detected as shaming and the attacker is subsequently blocked/muted by BlockShame, further tweets from him/her, unless unblocked/unmuted by the victim, do not show up in the victim's feed even if those are NS. However, tweets of the same shamer to other users who have not been shamed are not blocked. In case of "delete," where a DM received gets deleted from the victim's account, future NS communications (DMs, Mentions, and Tweets) from the shamer appear normally in the victim's feed. This behavior is changed when the "DM Account Action" in BlockShame preferences is set to "block" or "mute" instead of "none" as shown in Fig. 9.

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.

VII. Conclusion

In this paper, 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 standalone utterances, we studied them to be part of certain shaming events. In doing so, we observe that seemingly dissimilar events share a lot of interesting properties, such as a Twitter user's propensity to participate in shaming, retweet probabilities of the shaming types, and how these events unfold 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 paper has the potential for a user to choose to allow certain types of shaming comments (e.g., comments that are sarcastic in nature) giving his/her an opportunity for rebuttal and block others (e.g., comments that attack her ethnicity) according to individual choices. Freedom to choose what type of utterances one would not like to see in his/her feed beforehand is way better than flagging a deluge of comments on the event of shaming. This also liberates moderators from the moral dilemma of deciding a threshold that separates acceptable online behavior from unacceptable ones, thus relieving themselves to a certain extent from the responsibility of fixing what is best for another person.

To ascertain whether shaming a victim in itself causes an increase in popularity in Twitter as mentioned in

Section V-C, the change in FPM of shamers and that of nonshamers before and after a shaming event need to be compared at an individual level. The caveat is that Twitter only provides the present follower count as public domain information and not the follower count history. One possibility to overcome this shortcoming is to track a large set of shamers and nonshamers for a long time before and after the respective events have taken place. We intend to collect and curate even longer term Twitter data to achieve this.

Shaming is subjective in reference to shamers. For example, the same comment made by two different persons coming from different social, cultural, or political backgrounds may have different connotations to the victim. We would like to include the attributes of the author of the comment as contextual information when deciding if the comment is shaming or not. This algorithm, once developed, will be included in BlockShame to enrich its functionality. Moreover, in every event, we note that after the initial outrage, the volume of apologetic or reconciliatory comments gradually increases. A considerable proportion of users made multiple comments in a single event that contains both shaming and nonshame categories. We plan to investigate these behaviors further in the future. The performance of individual classifiers is promising though there are scopes for improvement. We would like to repeat our experiments with an even larger annotated data set to improve the performance further.

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/uk-25298580

[12] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in *Proc. SRW HLT-NAACL*, 2016, pp. 88–93.

[13] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proc. 26th Int. Conf. World Wide Web Companion*, 2017, pp. 759–760.

[14] D. Olweus, S. Limber, and S. Mihalic, *Blueprints for Violence 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 instagram social network." [Online]. Available: https://arxiv.org/abs/1503.03909

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

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

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

[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.

[30] 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

[31] 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.

[32] 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.

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

[34] 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.

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

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

[37] 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.

[38] 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.

[39] M. Conover, J. Ratkiewicz, M. R. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, "Political polarization on Twitter," in *Proc. ICWSM*, vol. 133, 2011, pp. 89–96.
[40] S. Hong and S. H. Kim, "Political polarization on Twitter: Implications for the use of social media in digital governments," *Government Inf. Quart.*, vol. 33, no. 4, pp. 777–782, 2016.
[41] Twitter. *Report Abusing Behavior*. Accessed: Feb. 7, 2018. [Online]. Available: https://help.twitter.com/en/safety-and-security/report-abusive-behavior
[42] *Blockshame Shields you from the Online Mob Just in Case!* Accessed: Feb. 7, 2018. [Online]. Available: http://cse.iitkgp.ac.in/~rajesh.basak/blockshame/

Rajesh Basak received the M.Tech. degree in network and internet engineering from Pondicherry University, Pondicherry, India, in 2014.

He is currently a Research Scholar with the Department of Computer Science and Engineering, IIT Kharagpur, Kharagpur, India. His current research interests include online social networks and machine learning.

Shamik Sural (SM'06) received the Ph.D. degree from Jadavpur University, Kolkata, India, in 2000.

He is currently a Professor with the Department of Computer Science and Engineering, IIT Kharagpur, Kharagpur, India. He has authored or co-authored more than 200 papers in reputed international journals and conferences. His current research interests include computer security and data science.

Dr. Sural was a recipient of the Alexander von Humboldt Fellowship for experienced researchers. He served as the Chairman for the IEEE Kharagpur Section. He is an Associate Editor of the IEEE TRANSACTIONS ON SERVICES COMPUTING.

Niloy Ganguly received the B.Tech. degree from IIT Kharagpur, Kharagpur, India, in 1992, and the Ph.D. degree from Bengal Engineering and Science University, Shibpur, India, in 2004.

He was a Post-doctoral Fellow with the Dresden University of Technology, Dresden, Germany. He is currently a Professor with the Department of Computer Science and Engineering, IIT Kharagpur, where he leads the Complex Networks Research Group. His current research interests include complex networks, social networks, peer-to-peer networks, and information retrieval.

Soumya K. Ghosh (M'05) received the M.Tech. and Ph.D. degrees from the Department of Computer Science and Engineering, IIT Kharagpur, Kharagpur, India, in 1996 and 2002, respectively.

He was with the Indian Space Research Organization, Bengaluru, India. He is currently a Professor with the Department of Computer Science and Engineering, IIT Kharagpur. He has authored or co-authored more than 200 research papers in reputed journals and conference proceedings. His current research interests include spatial data science, spatial web services, and cloud computing.