Characterizing and Countering Communal Microblogs During Disaster Events

Koustav Rudra, Ashish Sharma, Niloy Ganguly, and Saptarshi Ghosh

***Abstract*— The huge amount of tweets posted during a disaster event includes information about the present situation as well as the emotions/opinions of the masses. While looking through these tweets, we realized that a large amount of communal tweets, i.e., abusive posts targeting specific religious/racial groups are posted even during natural disasters—this paper focuses on such category of tweets, which is in sharp contrast to most of the prior research concentrating on extracting situational information. Considering the potentially adverse effects of communal tweets during disasters, in this paper, we develop a classifier to distinguish communal tweets from noncommunal ones, which performs significantly better than existing approaches. We also characterize the communal tweets posted during five recent disaster events, and the users who posted such tweets. Interestingly, we find that a large proportion of communal tweets are posted by popular users (having tens of thousands of followers), most of whom are related to media and politics. Further, users posting communal tweets form strong connected groups in the social network. As a result, the reach of communal tweets is much higher than noncommunal tweets. We also propose an event-independent classifier to automatically identify anticommunal tweets and also indicate a way to counter communal tweets, by utilizing such anticommunal tweets posted by some users during disaster events. Finally, we develop a real-time service to automatically collect tweets related to a disaster event and identify communal and anticommunal tweets from that set. We believe that such a system is really helpful for government and local monitoring agencies to take appropriate decisions like filtering or promoting some particular contents.**

***Index Terms*— Anticommunal tweets, classification, communal tweets, disasters, microblogs, Twitter.**

I. Introduction

ONLINE social media (OSM) such as Twitter and Facebook are today seriously plagued by offensive and abusive content, such as trolling, cyberbullying, hate speech, and so on. A lot of research has been carried out in recent years for automatic identification of different types of offensive content [1]–[5]. Hate speech can come under several categories where people target various attributes such as religion, gender, sex, ethnicity, nationality, etc., of the target group [6].

Out of different types of hate speech, we in this paper focus on an especially harmful and potentially dangerous category—*communal tweets*, which are directed toward certain religious or racial communities such as "Hindu," "Muslims," "Christians," etc. Especially, we study communal tweets that are posted during times of *disasters* or emergency situations. A disaster situation generally affects the morale of the masses making them vulnerable. Often, taking advantage of such situation, hatred and misinformation are propagated in the affected region, which may result in serious deterioration of law and order situation. In this paper, we provide a detailed analysis of communal tweets posted during disaster situations—such as automatic identification of such tweets, analyzing the users who post such tweets—and also suggest a way to counter such content.

Earlier it has been observed that such offensive tweets are often posted during man-made disasters like terrorist attacks. For instance, Burnap and Williams [1] have shown that the U.K. masses targeted a certain religious community during Woolwich attack to which the attackers are affiliated. However, it is quite surprising that in certain geographical regions such as Indian subcontinent, communal tweets are posted even during natural disasters such as floods and earthquakes. Some examples of communal tweets are shown in Table I. Such kind of communal tweets help in developing hatred and agnosticism among common masses, which subsequently deteriorates communal harmony, law and order situation. In the midst of disaster, this kind of situation is really difficult for government to handle.

In this paper, we try to identify communal tweets, characterize users initiating or promoting such contents, and counter such communal tweets with anticommunal posts that ask users not to spread communal venom. Although there exist prior works on communal tweet identification, to our knowledge,

TABLE I
Examples of Communal Tweets Posted During Disaster Events

F**k these *Missionaries* who are scavenging frm whatever's left after the #NepalEarthquake Hav some shame & humanity.
Dear #kashmirFloods take away all rapist *muhammad's piglets* out of kashmir with you, who forced out kashmiri *Hindus* from their motherland!!
Radical Muslims want to behead u, moderate Muslims want radical Muslims to behead you n liberals want to save thm. result. #GurdaspurAttack
RT @polly: #HillaryClinton's reply when asked if war on terror is a war on *"radical Islam"* #DemDebate
Jesus *F***ing Christ* ... Active shooter reported in San Bernardino, California

Manuscript received April 19, 2017; revised September 23, 2017 and December 21, 2017; accepted January 21, 2018. This work was supported by the Information Technology Research Academy, DeITY, Government of India under Grant ITRA/15 (58)/Mobile/DISARM/05. The work of K. Rudra was supported by a fellowship from Tata Consultancy Services. *(Corresponding author: Koustav Rudra.)*

The authors are with the Department of Computer Science and Engineering, IIT Kharagpur, Kharagpur 721302, India (e-mail: koustav.rudra@cse.iitkgp.ernet.in; niloy@cse.iitkgp.ernet.in; saptarshi@cse.iitkgp.ernet.in; ashishshrma22@gmail.com).

This work is an extended version of the short paper: Rudra *et al.*, "Characterizing Communal Microblogs during Disaster Events," Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016.

Digital Object Identifier 10.1109/TCSS.2018.2802942

this paper is the first on characterizing communal tweets and users who post such tweets during disasters, and it tries to find out how social media platforms are used to spread communal content even during natural disasters in some regions. Our major contributions are listed as follows.

1) We develop a simple *rule-based classifier* using low-level lexical and content features to automatically separate out communal tweets from noncommunal ones (Section IV). Keeping in mind the limitations of previous works [1], [7], we develop an *event-independent* communal tweet classifier that can be directly used to filter out communal tweets during future events. Experiments conducted over tweet streams related to several disaster events with diverse characteristics show that the proposed classification model outperforms vocabulary-based approaches [1], [8].
2) After identifying communal tweets, we study the nature of communal tweets and the users who post them (Section V). Broadly, we have observed two categories of users: 1) *initiators,* who initiate communal tweets and 2) *propagators,* who retweet communal tweets posted by initiators or copy the content of other initiators and post their own tweet with minor changes. We observe that, alarmingly, a significant section of communal tweets are posted by some very popular users who belong to media houses or are in politics. Such communal tweets are retweeted more heavily compared with other kinds of tweets. These communal users are connected via a strong social bond among themselves.
3) Apart from communal tweets, in this paper, we observe that the tweets posted during disaster events follow certain specific traits, which can be exploited to counter adverse effects of communal tweets. After the first-level classification, we obtained communal and noncommunal tweets. Further analysis of the noncommunal tweet set reveals that a small number of users post anticommunal tweets which try to dissuade people from posting communal content. However, it is observed that such anticommunal posts are less retweeted and receive less exposure compared to communal tweets. Hence, a convincing way to counter the communal venom during disasters is to promote such anticommunal content. In the second step, we develop a classifier (Section VI) for automatically separating out anticommunal tweets from noncommunal tweets (identified in the first level). In this case also, we rely on some low-level lexical features to make this classifier event independent. This is the first study, to our knowledge, that looks at anticommunal tweets as a practical way of countering adverse effects of communal tweets.
4) Finally, we develop a system *DisCom* (http://www.cnergres.iitkgp.ac.in/projects/disaster_communal_identifier/) that collects tweet streams posted during disaster situations and identifies communal and anticommunal content in real time.

Note that, our communal tweet characterization approach was first proposed in a prior study [9]. This paper extends our prior work as follows. First, we have proposed a rule-based classifier using low-level lexical features to extract communal tweets and this classifier can be directly used over any future event without further training. Second, earlier we had classified users into two categories: 1) originators who post a tweet and 2) propagators retweeting the content of originators. In this paper, we not only rely on retweets but also explore similarity between tweets, their timestamps in order to identify initiators and propagators more accurately. Rest of the analysis is performed on these modified groups of users. Apart from that, we also analyze temporal patterns of the identified set of communal users to understand their outraging phenomenon. Third, we propose another rule-based classifier to detect anticommunal tweets and such tweets can be used to neutralize the effect of communal tweets. We have also developed a real-time system *DisCom* to automatically identify communal and anticommunal tweets posted during new disaster events. As a final contribution, we make the tweet-ids of the tweets related to all these disaster events and lexicons used in developing the classifiers publicly available to the research community at `http://www.cnergres.iitkgp.ac.in/disasterCommunal/dataset.html`.

II. Related Work

Microblogs, online forums, are increasingly being used by the masses to post offensive content and hate speeches. In recent times, researchers have put a lot of effort for automatic identification of such offensive content [1]–[5]. This section briefly discusses these studies, and points out how this paper is different from the prior works.

Several studies have attempted to identify online content that is potentially hate speeches or offensive in nature. For instance, Greevy and Smeaton [10] proposed a supervised bag-of-words (BOW) model to classify racist content in webpages. Along with words, context features are also incorporated to improve the classification accuracy in a later version [11]. Chen *et al.* [12] identified offensive content in Youtube comments using obscenities, profanities, and pejorative terms as features with appropriate weightage. Similarly, *cyberbullying* was identified by Dinakar *et al.* [13], using features like parts-of-speech tags, profane words, words with negative connotations, and so on. More recently, Burnap and Williams [1], [14] proposed hate term and dependence feature-based model to identify hate speeches posted during a disaster event (the Woolwich attack). Alsaedi *et al.* [15] proposed a classification and clustering-based technique to predict disruptive events like riot. Burnap and Williams [16] proposed a model to detect cyber hate on Twitter across multiple protected characteristics such as race, disability, sex, etc.

In recent times, researchers have analyzed hate speeches, abusive behavior from social media such as Facebook, Reddit, etc. Delgado and Stefancic [17] analyzed conditions such as national values, social contact, etc., which facilitate hate speeches on internet services such as e-mail, Facebook, and Youtube. Jaishankar [18] conducted case studies on several instances of hate on social networking platforms such as Orkut, Facebook, and Myspace, and analyzed the ways in which

Fig. 1. Word cloud of two events. (a) NEQuake. (b) PAttack.

these sources were misused. Schieb and Preuss [19] explored cases where counter speech to hate was successful and created a computational simulation model to find the effects that hamper or aid the influence of antihate speeches on Facebook. Chandrasekharan *et al.* [20] designed a novel framework for utilizing data from multiple online communities such as 4chan, Reddit, Voat, and MetaFilter to detect abusive posts targeted to a community.

The present attempt to identify and characterize communal content in Twitter is motivated by the following two perspectives. First, hate speech can come under various categories where people target specific characteristics of users such as gender, race, sex, nationality, religion, ethnicity, and so on. Prior studies [3] show that most prevailing hate speeches are targeted toward certain races, while religion-induced hate speech is very sparse. Hence, a general purpose hate speech identifier may fail to capture all the nuances of a rare category (say religion-based hate speech), especially, when tweets from the rare categories are posted in huge volumes for a short period of time. We actually demonstrate in this paper that the classifier proposed in [3] can hardly capture communal tweets. Consequently, in recent times, researchers focus on more granular levels of hate speech detection in Twitter. For example, Chaudhry [2] tried to track racism in Twitter and Burnap and Williams [1] detected religious hate speeches posted during the Woolwich attack.

Second, most of the prior studies on hate speech have focused on content posted in blogs or webpages [4], [7]. On the contrary, this paper focuses on Twitter, and it has been widely demonstrated that the standard Natural Language Processing-based methodologies, which have been developed for formally written text, do not work well for short, informal tweets [21]. Hence, new methodologies are necessary to deal with noisy content posted on Twitter.

Burnap and Williams [1], [14] detected hate speech (religious and racial) posted during the Woolwich attack using a BOW model, where n-grams containing specific hate terms and some dependencies like "det" (determiner) and "amod" (adjectival modifier) are considered as features. However, the BOW model has a known limitation—classifiers based on this model are heavily dependent on event-specific n-grams extracted from the training data, which might not be suitable for applying the classifier to different types of events. For instance, Fig. 1 shows the tag clouds of communal tweets posted during two different events—the Nepal earthquake (April 2015) and Paris terrorist attack (November 2015). It is evident from the figure that the religious community being targeted, and hence, the vocabularies are significantly different for these two events. As a result, a BOW-based classifier is unlikely to perform well if trained on one of these events and used on the other. Recently, Magdy *et al.* [8] used post event tweets to learn users stances toward Muslims and exploited preevent interactions, posted tweets to build a classifier to predict post event stances. However, we observe that overlap among the users who post communal tweets during multiple events is very low (Section V). Hence, such user-specific classifier has very low chance to perform well on future events. On the other hand, using low-level lexical and content features (instead of specific terms) can make the classifier's performance largely independent of specific disaster events considered for training as demonstrated in our prior work [22]. These findings motivated us to propose an *event-independent communal tweet classifier*.

The focus of almost all the prior works is on identifying offensive hate speech contents. However, very little efforts were made to characterize the users who post such contents. Recently, Silva *et al.* [3] tried to detect the sources and targets of hate speeches. However, detailed characterization of users who post offensive contents is necessary.

Preliminary version of this paper has been published in [9]. In this paper, we have extended it as follows: 1) we have developed our own classifier to identify communal tweets; 2) characterize users who posted communal tweets during disaster events in more detail (Section V); and 3) finally, we propose a method to detect anticommunal tweets and show how such tweets can be used to neutralize the harmful effect of communal tweets.

III. Data Set

This section describes the data sets used for the study, and various types of tweets present in the data sets.

We considered tweets posted during the following recent disaster events:

1) *NEQuake:* a destructive earthquake in Nepal;
2) *KFlood:* floods in the state of Kashmir in India;
3) *GShoot:* three gunmen dressed in army uniforms attacked the Dina Nagar police station in Gurudaspur district of Punjab, India;
4) *PAttack:* coordinated terrorist attacks in Paris;
5) *CShoot:* a terrorist attack consisting of a mass shooting at the Inland Regional Center in San Bernardino, CA, USA.

Note that, the first two events are natural disasters, and the last three events are man-made disasters. Additionally, we have considered events occurring in different geographical regions so that this paper would not get influenced by any kind of demographics. Tweet-ids of these tweets are made publicly available to the research community at `http://www.cnergres.iitkgp.ac.in/disasterCommunal/dataset.html`.

We applied keyword-based matching to retrieve relevant tweets using the Twitter Application Programming Interface (API) [23] during each event. For example, to identify the tweets related to the NEQuake event, we search tweets with keywords like "#NepalEarthquake," "Nepal," and "earthquake," etc. For each keyword, we collected *all* the tweets returned by the Twitter Search API.

TABLE II
STATISTICS OF DATA COLLECTED

Event	# Tweets	# Distinct users
NEQuake	5,05,077	3,26,536
KFlood	14,922	8,367
GShoot	53,807	29,293
PAttack	6,48,800	5,77,888
CShoot	2,93,483	1,64,276

TABLE III
GOLD STANDARD—NUMBER OF TWEETS IN DIFFERENT DISASTER EVENTS

NEQuake	KFlood	GShoot	PAttack	CShoot
247	112	203	201	152

Further, we consider only English tweets based on the language identified by Twitter.

For each event, we report the number of tweets collected and the number of distinct users who posted them in Table II. We describe our communal tweet identification step in Section IV.

IV. IDENTIFYING COMMUNAL TWEETS

This section focuses on extracting communal tweets from rest of the tweets, by developing a rule-based classifier.

A. Establishing Gold Standard

To understand the patterns, specific traits of communal tweets and evaluate the proposed classifier, we require gold standard annotation for a set of tweets. For each of the events stated in Section III, we randomly sampled 4000 tweets (after removing duplicates). These tweets were independently observed by three human volunteers, all of whom have a good knowledge of English. The volunteers were asked to identify whether a tweet is communal or not.

There was an unanimous agreement for 81% tweets, while we consider the majority decision for the rest. By this process, a total of 915 tweets were identified as communal. Table III shows the number of tweets in gold standard across five disaster events. From the rest of the tweets, we randomly sampled the same number of noncommunal tweets to build gold standard data set.

B. Features for Classification

As stated earlier, we want our classifier to be event independent, i.e., the classifier should be such that it can be directly used over tweets posted over later events. Hence, we take the approach of using a set of lexical and content features for the classification task, which is known to make the classifier's performance largely independent of the events considered for training [22].

We use the first three data sets, i.e., NEQuake, KFlood, and GShoot as *training set*. In other words, the tweets from these three data sets are used to identify discriminating features and develop our classifier. The other two data sets, i.e., PAttack and CShoot, are used as *test set*, to check the performance of our proposed classifier over future disaster events. Next, we describe the features used for classification.

1) Presence of Communal Slang Phrases: In order to develop the classifier, we needed a lexicon of religious terms and antagonistic hate terms about religion and related nationality. For this, we considered the terms in a standard lexicon of religious terms `http://www.translationdirectory.com/glossaries/`. However, all these terms are *not* hate terms; rather, the lexicon contains many general religion-related terms as well. Hence, we employed three human annotators (the same who judged the tweets) to mark the terms in the lexicon as hate terms or normal religious term. We obtain an unanimous agreement for 84% of the terms, and for the rest, we follow majority verdict. Similarly, we collected all the hate terms related to religion and nationality from a repository of terms frequently used in hate speeches—`www.hatebase.org`.

2) Presence of Religious/Racial Negated or Hate Terms: We detect the presence of any strongly negative term or slang term in the vicinity of neutral religious terms such as "Muslim" or "Christian." We use a subjectivity lexicon developed in [24] to identify strongly negative terms, and we obtain a standard list of slang terms from `www.noswearing.com`. Then, we check whether such terms appear within a left and right word window of size two each with respect to a religious term. Thus, presence of phrases like "bastard missionaries," "islamic scoundrels," "jesus f***tards" are identified.

3) Presence of Communal Hashtags: We observed that some specific hashtags are explicitly used across various events to curse certain religious communities, such as "#SoulVultures," "#evangelicalvultures," "#WeAreThanklessMuslims," and "#TweetlikeSecularJamat." Such hashtags are mostly present in communal tweets. We ourselves developed a lexicon of such communal hashtags. These lexicons can be downloaded and used for research purposes.[1] Note that these hashtags were identified by the annotators only from the training set, i.e., the NEQuake, KFlood, and GShoot data sets.

4) Presence of Religious Terms With wh-Words/Intensifiers: Sometimes wh-words/intensifiers with neutral religious terms such as "Muslim" or "Christian" are used to target certain religious communities sarcastically specially in disaster scenario (e.g., "Why do all the Muslim guys barking endian endian?? If u dnt knw hw to write english jst dnt write.. #GurdaspurAttack"). Sometimes we also observe that a tweet that appears to be a normal tweet in the general scenario can actually become communal in the context of a disaster (e.g., "Why do Christians pray," which is a sarcastic comment on the religious habits of a religious group). We use a list of intensifiers (so, too, really, $\cdots\cdots\cdots$) collected from Wikipedia.[2]

C. Evaluating Classification Performance

We compare the performance of our proposed set of features under two scenarios: 1) *in-domain classification*, where the

[1] `http://www.cnergres.iitkgp.ac.in/disasterCommunal/dataset.html`

[2] `https://en.wikipedia.org/wiki/Intensifier`

classifier is trained and tested with the tweets related to the *same event* using a 10-fold cross validation and 2) *cross-domain classification*, where the classifier is trained with tweets of one event, and tested on another event. In this case, all the annotated tweets of a particular event are used to train/develop the model and then it is tested over all the tweets of rest of the events.

Selection of Classification Model: Performance of a classifier is heavily dependent on the appropriate model selection. We now attempt to select the most appropriate model for our proposed set of features based on some specific criteria. We consider seven state-of-the-art classification models for the above set of features: 1) SVM with default RBF kernel and $\gamma = 0.5$ (SVMG); 2) SVM with RBF kernel (SVM); 3) Random Forest (RF); 4) SVM with linear kernel (LSVC); 5) Logistic regression (LR); 6) Naive Bayes (NB); and 7) Rule-based classifier (RL)—here we follow a simple approach—if any of the above-mentioned features is present in a tweet, we mark that tweet as communal; otherwise noncommunal.

For each of these models (except rule based), we use the Scikit-learn [25] package. To judge the performance of these models on the above-mentioned feature sets, we set the following evaluation criteria. Each criterion is computed and averaged over the three training data sets.

1) *Average In-Domain Accuracy:* Average accuracy of the classifier across the three events in the training set, as in in-domain scenario.
2) *Average Cross-Domain Accuracy:* Average accuracy of the classifier in different cross-domain scenarios among the three events in the training set. In this case, we have six different cross-domain settings.
3) *Average Precision for Communal Tweets:* Detection of communal tweets with high precision is a necessary requirement for the classifier. Hence, we consider average precision across the three training data sets.
4) *Average Recall for Communal Tweets:* The classifier should ideally capture all the communal posts, i.e., have high recall. Hence, we consider the recall averaged over the three training data sets.
5) *Average F-Score for Communal Tweets:* F-score of the classifier indicates the balance between coverage/recall and accuracy/precision. F-score is calculated as a harmonic mean of precision and recall using the following equation:

$$\text{F-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}. \quad (1)$$

We report the performance of different classification models on the proposed set of features in Table IV. From Table IV, it is clear that the proposed rule-based classification model shows slightly better performance compared to other models.[3] These results clearly reveal the benefit of working with event-independent features. All the subsequent results are produced using the rule-based model.

[3]Though the proposed model performs better than the other classifiers, the improvement is not statistically significant in most cases.

TABLE IV
SCORE OF DIFFERENT EVALUATING PARAMETERS FOR SEVEN DIFFERENT CLASSIFICATION MODELS USING PROPOSED FEATURES

Classifier	In-domain accuracy	Cross-domain accuracy	Precision	Recall	F-score
SVMG	0.9295	0.9308	0.9504	0.9106	0.9284
SVM	0.9267	0.9308	0.9502	0.9087	0.9274
RF	0.9304	0.9308	0.9404	0.9117	0.9258
LSVC	0.9295	0.9308	0.9504	0.9106	0.9284
LR	0.9254	0.8919	0.9513	0.8530	0.8941
NB	0.9267	0.9117	0.9509	0.8817	0.9112
Rule-based	0.9308	0.9308	0.9494	0.9117	0.9291

D. Comparison of Proposed Approach With Baselines

We use the following state-of-the-art communal tweet detection approaches as our baselines:

1) BUR: A religious and racial hate speech detection approach proposed by Burnap and Williams [1] using n-grams(1–5), hateful terms (`http://www.rsdb.org/`) and Stanford typed dependencies like "determiner" and "adjectival modifier."

2) USR: Recently, Magdy *et al.* [8] have shown that the past tweet history of users can be used to detect communal tweets. This method used to preevent interactions (mentions and replies), contents/tweets (unigrams and hashtags) posted by users to predict post event stances of these users.

Note that both the baseline methods are supervised, hence they require training. However, our proposed method is rule based (unsupervised) and can be used directly over future events. For training and testing of baseline methods, we have used the SVM classifier—specifically, the Scikit-learn package [25] with the linear kernel.

Performance of Baseline Classifiers: Table V shows the performance of the baseline classifiers when trained and tested on NEQuake, KFlood, and GShoot events.

a) In-domain classification: Here, tweets from the same event are used to train and test the baseline classifiers and accuracy is measured using 10-fold cross validation. The results are shown in the diagonal entries in Table V. The BUR method performs quite well in case of in-domain scenario and achieves around 83% accuracy averaged over all the three events. Given that the USR method does not perform well, it is evident that users' past history is not helpful in predicting future stances.

b) Cross-domain classification: In this case, tweets of one event are used to train the baseline classifier and then it is tested over tweets of another event. Results are shown in the *nondiagonal* entries in Table V, where the left-hand side event is used as the training event, and the event stated at the top represents the test event. In this case, the performance of the baseline models is often as low as that by random chance (accuracy 50%). Only in some cases, where the same community was targeted in both training and test event, the BUR model achieves around 69% accuracy.

Performance of Proposed Classifier: Table VI shows the performance (precision, recall, F-score, and accuracy) of the proposed rule-based classifier on the same three events. Averaged over the three data sets, our proposed rule-based classifier

TABLE V

CLASSIFICATION ACCURACIES (AC), RECALL (R), AND F-SCORES (F) FOR COMMUNAL TWEETS,USING BASELINE MODELS (BUR, USR). DIAGONAL ENTRIES REPRESENT IN-DOMAIN CLASSIFICATION, WHILE THE NONDIAGONAL ENTRIES REPRESENT CROSS-DOMAIN CLASSIFICATION

Train set	Test set																	
	NEQuake						KFlood						GShoot					
	BUR			USR			BUR			USR			BUR			USR		
	AC	R	F	AC	R	F	AC	R	F	AC	R	F	AC	R	F	AC	R	F
NEQuake	0.8659	0.84	0.8609	0.7013	0.8402	0.7321	0.6043	0.4086	0.5081	0.5654	0.7289	0.6265	0.55	0.3850	0.4638	0.55	0.6666	0.5970
KFlood	0.5260	0.5812	0.5631	0.5647	0.6352	0.5934	0.8488	0.7916	0.8362	0.5922	0.7654	0.6409	0.7075	0.4950	0.6285	0.5388	0.5888	0.5608
GShoot	0.5140	0.5307	0.4989	0.5235	0.5999	0.5573	0.6826	0.3826	0.5465	0.5654	0.6355	0.5938	0.7950	0.7750	0.7908	0.5222	0.6888	0.5796

TABLE VI

CLASSIFICATION SCORES (PRECISION, RECALL, AND F-SCORE) FOR COMMUNAL TWEETS AND OVERALL ACCURACY USING RULE-BASED CLASSIFIER WITH PROPOSED FEATURES, FOR THE EVENTS IN THE TRAINING SET

Event	Precision	Recall	F-score	Accuracy
NEQuake	0.9698	0.9000	0.9336	0.9360
KFlood	0.9173	0.9652	0.9406	0.9391
GShoot	0.9613	0.8700	0.9133	0.9175

TABLE VII

CLASSIFICATION SCORES (PRECISION, RECALL, AND F-SCORE) FOR COMMUNAL TWEETS AND OVERALL ACCURACY USING RULE-BASED CLASSIFIER WITH PROPOSED FEATURES, FOR FUTURE EVENTS

Event	Precision	Recall	F-score	Accuracy
PAttack	0.9336	0.9849	0.9586	0.9575
CShoot	0.9006	0.9666	0.9324	0.9300

achieves 94% precision and 91% recall in communal tweet detection. Thus, it is clear from Table VI that our proposed method performs significantly better compared to baseline techniques. This improvement is 17% over method proposed by Burnap (BUR). Note that, since we define a set of rules that are independent of the vocabularies used in an event, no separate training is required for the proposed classifier.

E. Further Analysis of Proposed Classifier

1) Application Over Future Events: As stated in Section I, our objective is to make the communal tweet classifier independent of the vocabularies used during a specific disaster. We used NEQuake, KFlood, and GShoot events as a training set, to learn the patterns of communal tweets. In this section, we apply the classifier over other two events (PAttack and CShoot). Table VII reports recall, F-score, and accuracy of the classifier for these two events. The proposed classifier achieves very high performance over these two future events as well. Hence, we see that people follow more or less similar patterns in targeting different religious communities during various disaster scenarios.

Note that out of the four features used for the classification (described in Section IV-B), only one is dependent on terms derived from the training set—presence of communal hashtags. The other three features are based on our observations from the training events, but do not use any information specific to the training events. Additionally, the feature ablation experiments (reported later in Table IX) show that the presence of communal hashtags is not very important for the classification. Hence, the performance of the classifier would not be significantly affected even if we change the training and test events.

TABLE VIII

MISCLASSIFIED COMMUNAL TWEETS POSTED DURING DISASTERS

"Allah ho Akbar" battle cry was raised by pigs killed in #GurdaspurAttack #presstitutes media will not show,because they r funded by Saudi's
Huh, its a Muslim behind California attack
Threat frm a kashmiri muslim not a terrorist. Every1 shd keep this as proof

2) Analyzing Misclassified Tweets: For our proposed method, we have also analyzed different types of errors i.e., how many times a communal tweet is marked as a noncommunal tweet or vice versa. Table VI reflects that we achieve precision of 0.94 over the three training data sets, which indicates around 6% noncommunal tweets are marked as communal tweets. On the other hand, an average recall score is 0.91. 9% of communal tweets is misclassified as noncommunal tweets. Similarly, for these two test events (Table VII), we achieve precision of 0.93 and 0.90 for PAttack and CShoot, respectively. In other words, around 7% and 10% noncommunal tweets are marked as communal ones. However, recall is relatively high and only 2%–4% communal tweets are missed out by the classifier.

Marking a communal tweet as noncommunal is a more serious problem compared to classifying a noncommunal tweet as communal. Table VIII shows some examples of misclassified communal tweets. Almost in every case, tweets are posted in a sarcastic way, i.e., particular communities are targeted in roundabout fashion. In this paper, we have tried to capture some part of sarcasm by checking the presence of wh-words, intensifiers along with religious terms. However, in the future, we will try to capture more sarcastic patterns present in communal tweets considering event/vocabulary-independent models [26].

3) Feature Ablation: Finally, we attempt to judge the importance of individual features in the classification, through feature ablation experiments. One feature is dropped at a time, and the degradation of the classifier performance (as compared with the performance using all features) gives an idea of the importance of the dropped feature. Table IX reports the accuracy, recall, and F-score of the communal tweet classifier for feature ablation experiments, averaged over all the data sets. Presence of communal slangs and religious/racial negated terms appear to be the most determining factors. However, all

TABLE IX

FEATURE ABLATION EXPERIMENTS FOR THE PROPOSED CLASSIFIER. NONE REPRESENTS THE CASE WHEN ALL FEATURES WERE USED

Ablated Feature(s)	Accuracy	Recall	F-score
NONE	0.9360	0.9373	0.9357
Religious negated terms	0.8687	0.7744	0.8665
Communal slangs	0.7595	0.5518	0.6852
Communal hashtags	0.9112	0.8852	0.9048
Religious terms with wh-words / intensifiers	0.9101	0.8846	0.9061

the features together help in increasing the overall accuracy of communal tweet classifier.

The above results indicate that communal and noncommunal tweets can be effectively classified based on low-level content-based features.

V. CHARACTERIZING COMMUNAL TWEETS AND ITS USERS

In this section, we try to understand and characterize communal tweets and the users who post them. We apply our proposed classifier described in Section IV, over the data sets; we refer to the tweets that were categorized as communal by our classifier as *communal tweets* (60 000), and the users who posted them as *communal users* (48 000). Specifically, we compare the set of communal tweets and communal users during a particular event with an equal number of randomly sampled noncommunal tweets (as judged by our classifier) and the users who posted them (referred to as *noncommunal users*) during the same event.

A. Characterizing Communal Tweets

1) Which Communities Are Targeted?: It is observed that during disaster scenario, people post communal tweets targeting specific religious communities. Examples of some communal tweets and communities targeted via those tweets are given in Table X. We observe that these targeted communities do not remain same across different disasters. During man-made disasters, like terrorist attacks, common masses mainly target that community to which attackers are affiliated. Along with that, some other communities are also targeted. For example, during Paris attack, Islamic people were the main targets but Christians were also targeted side by side.

It is interesting to note that users post communal tweets targeting specific religious communities like Christian missionaries, Muslims, etc., even during natural disasters like NEQuake and KFlood. During natural disasters, most of the people target core communities of the affected place which have been causing harm to the sentiments of other communities. For example, during Kashmir Floods, Muslims were targeted as some of the Muslim residents of Kashmir had maligned a temple of lord Shiva (Hindu mythological figure) before the disaster occurred. However, in some cases (e.g., Nepal earthquake), people have specific reasons for targeting a community due to the behavior and exertion of certain people of that community during the post disaster scenario.

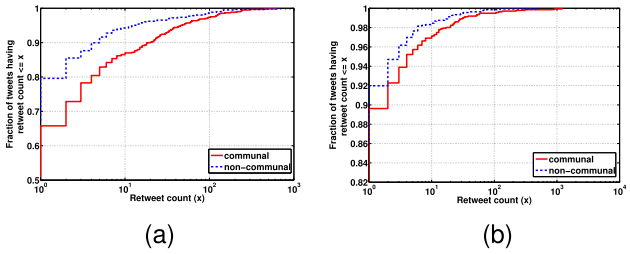

Fig. 2. CDF of retweet count of communal and noncommunal tweets. Communal tweets are retweeted more. (a) NEQuake. (b) CShoot.

2) Popularity of Communal Tweets: In this section, we check whether communal tweets receive large attention from people. To measure the popularity of a tweet, we consider retweet count of a tweet which is a standard metric to determine its exposure.[4] We show the distribution of retweet counts of communal and noncommunal tweets, for the two events, NEQuake and CShoot, in Fig. 2. For this study, we discarded retweets and only considered original tweets. From Fig. 2, we can see that communal tweets become more popular compared to noncommunal ones. We observe a similar pattern for other events.

B. Characterization of Communal Users

We next analyze the users who post communal tweets during the disaster events. For this, communal users are divided into following two categories: 1) *initiators*, users who initiate communal tweets and 2) *propagators*, users who retweet the communal tweets posted by initiators or some other propagators or who copy the content of some initiator and post their own tweet with minor changes.

We next describe the construction procedure of initiator and propagator sets and study the properties of initiators and propagators separately.

1) Construction of Initiator and Propagator Set: For dividing users into initiators and propagators, we need to find the set of retweets Y of a particular tweet x. The users in set Y would then be classified as propagators, while the user who posted x will become the initiator. As per the prototype, tweet x of user u is said to be propagated by tweet y of user v if y is formed by copying x, preceding it with RT and addressing u with @. However, due to the 140-character limitation on twitter and user's personal formatting preferences, a significant number of retweets do not follow this prototype [27]. Users like to add their own comments and sometimes even skip acknowledging the original users. As a consequence, some of the retweets lack distinguishable markers and patterns which makes their identification difficult [28]. Thus, in order to get a near-accurate classification of users into initiators and propagators, there is a need to incorporate the inconsistent syntax a significant number of users follow while retweeting. We attempt to minimize error in this classification and try to find *true initiators and propagators*. We first compute normalized *phrasal overlap measure* [29] between all pair of tweets in

[4]All the tweets are recrawled after several months from the date of the events, and hence, such tweets contain more or less final retweet count.

TABLE X
COMMUNITIES TARGETED DURING DISASTER EVENTS

Event	Communities Targeted	Sample communal tweets
NEQuake	Christian	Meanwhile cheap *Christians* r busy spreading Christianity, Idiot morons #NepalEarthquake
	Muslim	Y shd Allah waste his time killing *muslims* in #NepalEarthquake when de demselves r killing each others #SoulVultures [url]
KFlood	Muslim	I wish equal no of *muslims* perish & equal no are forced 2 leave their homes like KPs.. Then only they'll understand pain.. #kashmirFloods
GShoot	Muslim	#GurdaspurAttack is jihad on heart of Punjab ,brave sikhs will never forgive these *muslim* pigs for their coward act #Gurdaspur
PAttack	Persecuted Christian	RT @USER: If u think bringing any "persecuted Christians" into America from Syria and no terrorists will slip through, you're a fu
	Muslim	Slaughter. Like shooting fish in barrel. Now tell me what should be done with radical *Muslims* ?
CShoot	Muslim	https://t.co/xQ2Xo7WbPa via @USER- Take care of the radical *Islam* terrorists first Sir

our corpus. This measure is based on the Zipfian relationship between the length of phrases and their frequencies in a text collection and is defined as follows:

$$\text{phrasal\_overlap\_norm}(t_1, t_2) = \tanh\left(\frac{\sum_{i=1}^{n} m(i) * i^2}{|t_1| + |t_2|}\right) \quad (2)$$

where $m(i)$ is the number of i-gram phrases which match in tweets t_1 and t_2, n represents the highest n-gram considered for computing phrasal overlap, and $|t_1|$ is the length of tweet t_1. In (2), higher n-grams get more weight which also helps in capturing the context as opposed to comparison of unigrams. We then cluster together the tweets t_1 and t_2 having $\text{phrasal\_overlap\_norm}(t_1, t_2) \geq \text{similarity\_threshold}$ using Hierarchical Clustering Algorithm. We define the representative of each cluster as the tweet which was posted first on twitter among all the tweets of the cluster (i.e., tweet having the smallest timestamp). Phrasal overlap between two clusters is defined as the overlap between the representative tweets of those clusters.[5] For a cluster of size k, one tweet (tweet having the smallest timestamp) is representative tweet and the rest of the $k-1$ tweets are retweets of that tweet. The users corresponding to representative tweets become initiators and those corresponding to retweets become propagators. For our purposes, we take the value of n as 3 and similarity_threshold as 0.8.[6]

2) Popularity of Initiators and Propagators: We next investigate popularity of users who post communal tweets during disaster. Popularity of a user works as a major driving force in determining the popularity of tweets [30]. We observe a uniform phenomenon across all the five disaster events—both common masses (27% having less than 100 followers) and popular users (10% having more than 10 000 followers) involve themselves in initiating and propagating communal content. Especially, some popular communal users belonging to media houses and politics have several tens or hundreds of thousands of followers. We provide examples of some such popular communal initiators and propagators in Table XI.

[5]Please note that we remove "RT @user" from the tweet, if present, before finding the overlap similarity.

[6]We have tried different values but this setting provides best result.

TABLE XI
SAMPLE POPULAR USERS WHO POSTED COMMUNAL TWEETS

Role	Screen_name	Follower count	Bio of the user
Initiator	abhijitmajumder	69.5k	Journalist. Managing editor, Mail Today. Views are personal, retweets are not necessarily endorsements
	HinduRajyam	11.9k	Om Namo Venkateshaya Namaha.Establishing Hindu Rashtra shud be the immediate goal of every hindu! Follow @noconversion
Propagator	SanghParivarOrg	133k	http://t.co/cF4rB7S56v is an independent initiative by Swayamsevaks. @RSSOrg is official Twitter Handle for RSS
	mediacrooks	98k	changing the way we consume news.... rts do not imply endorsements..

TABLE XII
OVERLAP SCORE BETWEEN INITIATORS AND PROPAGATORS FOR COMMUNAL AND NONCOMMUNAL TWEETS ACROSS DIFFERENT EVENTS

Tweet type	NEQuake	KFlood	GShoot	PAttack	CShoot
Communal	0.21	0.20	0.14	0.11	0.11
Non-communal	0.32	0.32	0.29	0.23	0.30

3) Do Initiators Also Work as Propagators?: Next, we try to figure out whether during a disaster event communal tweet initiators also play the role of propagators during the same event. For this, during each event, the *Szymkiewicz-Simpson similarity* score [31] between initiator set and propagator set is computed. Table XII shows the overlap score obtained across five disaster events for both communal and noncommunal tweets. For communal tweets, we obtain a low similarity score of 0.15 averaging over all the events. Thus, communal tweet initiators hardly involve themselves in retweeting others contents; rather they are interested in posting their own views. Interestingly, this overlap score for natural disasters (NEQuake and KFlood) is twice the score of man-made disasters (GShoot, PAttack, and CShoot). Generally, in case of man-made disasters, common masses become angry and they raise their voice. Hence, initiators hardly involve themselves in propagating

such tweets. In case of natural disasters, communal sentiment among the common masses is not instinctive. Thus, initiators also play the role of propagators in order to activate communal belief among the people.

However, the overlap between communal initiators and propagators is less than that of noncommunal initiators and propagators. For this overlap, we do not observe any significant difference between natural and man-made disasters.

4) User Overlap Across Different Events: We investigate whether a common set of users involved themselves in initiating/propagating communal tweets during multiple events. For this, we considered events that occurred in the same geographical region (e.g., NEQuake, KFlood, GShoot, and all of which occurred in the Indian subcontinent). We found a small set of common users who posted tweets across all the three events. For instance, communal tweets are posted during all these three events by initiators as "simbamara," "RamraoKP_," and propagators like "IndiaAnalyst," "HinduRajyam." In general, overlap among the communal users of three events is low (about 5%). This overlap score is three times higher (about 15%) in case of noncommunal tweets. We define such common set of communal users as *core communal users*.

It is observed that only 22% of the core communal users are initiators and rest of the users help in propagating communal content during disaster. We also analyze the influence of such core users. Around 10% followers of these core users are popular (having more than 10 000 followers) and popularity of these users can help in getting wide exposure of communal content posted by core users. Again 5% of these core communal users are popular i.e., these users have more than 10 000 followers. If such users post communal content then they have high probability of getting large number of retweets and exposure.

5) Topical Interests of Communal Users: In this section, we try to infer topical interests of communal users. Specifically, we attempt to match the interests of communal users to one of seven broad topics: 1) Media and Journalism (News); 2) Politics; 3) Movies and Entertainment; 4) Writers/Authors; 5) Sports; 6) Religion; and 7) Business. We collected specific keywords from online sources,[7] which help in characterizing the above-mentioned broad topics. Users whose topics of interest do not fit into any of the above-mentioned categories are marked as others.

To perform this analysis, users are divided into two categories: 1) common users, having < 5000 followers and 2) popular users, having $\geq 10\,000$ followers. Twitter account bio is used to infer the topical interest of communal users. We check whether the keywords corresponding to any of the broad topics stated above are present in their biographies. For popular users, we not only rely on their biographies but also use our prior method [32] which can infer topical interest of popular users. Finally, we match the topical characteristics with the keywords corresponding to any of the broad topics.

We show the distribution of topical interests of popular and common initiators in Table XIII. We notice a similar phenomenon for propagators. Most of the popular initiators belong to news media and politics. Interest of common masses is distributed across multiple topics such as news, sports, politics, religion, etc.

TABLE XIII

DISTRIBUTION OF TOPICS OF INTEREST OF COMMON AND POPULAR INITIATORS OF COMMUNAL TWEETS

User	Broad topic of interest							
	Media	Politics	Sports	Religion	Writing	Entertainment	Business	Others
Popular users	50%	33%	2%	5%	2%	4%	1%	3%
Common users	21%	25%	12%	19%	6%	8%	2%	7%

TABLE XIV

COMPARING THE PROFILE BIO AND TWEETS POSTED BY USERS WHO POSTED COMMUNAL TWEETS AND OTHER USERS

Most frequent words in bio	
communal	religion, india, hindu, life, endorsement
non-communal	fan, indian, music, lover, life
Most frequent words in tweets	
communal	hindu, religion, congress, media, muslim
non-communal	govt, india, life, people, movie

For active users, their profile and the past history can also be useful in characterizing them. Thus, for further analysis, we process[8] the posted tweets and account bio of communal and noncommunal users to infer their interest and behavior.

For each category of users, we show top 5 words that appear in their account bio and posted tweets in Table XIV. As expected, we find the presence of religion and politics-related words in the bio and tweet of communal users. However, we do not find any topic-specific alignment with the most occurring words in the bio and tweet of noncommunal users. Such words are either normal chat words or they represent positive sentiment.

6) Are Common Communal Users Provoking Popular Users?: Mentioning popular users to improve visibility of tweets is a common phenomenon on Twitter. Traditional communication theory states that a minority of users, called the *influentials*, excel in persuading others [33]. Thus, mentioning these influentials in the network helps in achieving a large-scale chain-reaction of influence driven by word-of-mouth [30], [34]. Popular users, i.e., users having a large number of followers on twitter, are influential, and a retweet by popular users can help improve the visibility of a tweet [35]. Thus, common users, i.e., users with small number of followers on twitter, often mention popular users in their tweets to increase the reachability and effectiveness of tweets. Table XV shows the percentage of times a common user mentioned a popular user out of the total mentioning instances in communal and noncommunal tweets, respectively, in case of natural (NEQuake and KFlood) and man-made (GShoot, PAttack, and CShoot) disasters.

We found that the percentage of cases in which a common user ($<$5000 followers) mentions a popular user ($\geq$10 000 followers) is larger for communal tweets than noncommunal

[7] goo.gl/p4CPyX, goo.gl/Iqxo9T

[8] case-folding, stopwords removal, etc.

TABLE XV

% OF TIMES COMMON USERS MENTION POPULAR USERS IN COMMUNAL AND NONCOMMUNAL TWEETS

Event	Communal Tweets	Non-communal Tweets
Natural	63.74%	**69.06%**
Man-made	**74.07%**	68.89%

TABLE XVI

RECIPROCITY AND DENSITY OF THE MENTION AND FOLLOW NETWORKS AMONG DIFFERENT GROUPS OF USERS

Event	User group	Mention Network		Follow Network	
		Reciprocity	Density	Reciprocity	Density
NEQuake	communal	4.20%	0.0037	25.05%	0.0099
	non communal	3.31%	0.0002	16.88%	0.0007
GShoot	communal	4.74%	0.0047	26.14%	0.0133
	non communal	3.78%	0.0012	16.92%	0.0038

tweets in the case of man-made disasters and smaller for natural disasters. While computing these results, we had used the number of followers of a user as his/her measure of influence and popularity. It is clear that in the case of man-made disasters when people are already angry toward some communities, people try to provoke popular users by mentioning them in their communal posts. On the other hand, such trend is less in the case of natural disasters.

C. Interactions Among the Users

In this section, we check the interaction pattern of noncommunal and communal users among themselves. In Twitter, user u can interact with user v mostly in the following two ways: 1) v can be mentioned (@mention) by user u in her tweet and 2) u can subscribe to the content posted by v by following v.

Two types of interaction networks are constructed among users: 1) *mention network*—if user u has mentioned v, we add a link $u \rightarrow v$ and 2) *follow network*—if user u follows the content posted by v, we add a direct link $u \rightarrow v$. To quantify the level of interaction among the users, two structural properties of the above-mentioned networks are measured: 1) density, fraction of number of links present in a network and all possible links that can be present in a network and 2) reciprocity, what fraction of directed links are reciprocated, i.e., $v \rightarrow u$ and $u \rightarrow v$ both present in the network. Mutual friends generally have a high probability to share reciprocal links.

We report reciprocity and density values for mention and follow networks among different groups of users in Table XVI. A similar trend is observed across all the disaster events. Here, we report the result for two disaster events—NEQuake and GShoot. From Table XVI, we can see that communal users form a more dense network among themselves compared to noncommunal users. Apart from density, we also observe that reciprocity of both the networks is higher for communal users. It indicates that a large fraction of communal users are mutual friends. Thus, there is a significant interaction among communal users and strongly-tied communities formed by them in social network.

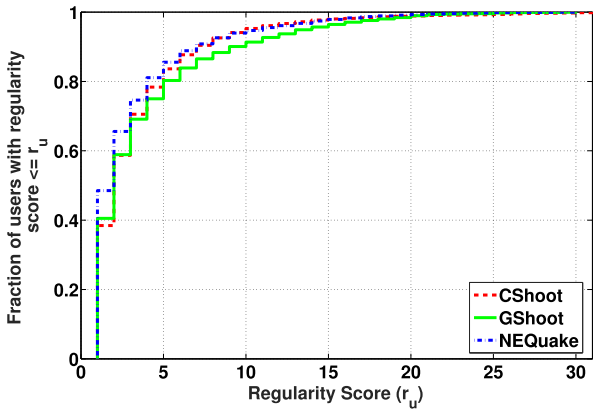

Fig. 3. CDF of regularity score of communal users.

D. Are the Users Getting Outraged Suddenly?

Previous studies argue that a significant rise is observed in communal hate online following "trigger" events like disaster [1], [36], [37]. According to them, these trigger events work as activators to wake up the old feelings of hatred and negative sentiments toward suspected perpetrators and related groups. In this section, we check if such a sudden rise exists in the case of disasters and attempt to quantify it. We are also interested in finding out whether there exist users who have a general tendency to post communal tweets irrespective of the event and situation. In order to perform this analysis, we study the nature of tweets posted by the communal users for a particular time period surrounding the disaster which encompasses general as well as event-specific behavior of the communal users. Let a user u in our data set first posted a communal tweet on day d. We define $TimeWindow(u, d)$, corresponding to a communal user u as a list of 31 days, comprising of 15 days before d and 15 days after d. For each communal user u, we scrapped all the tweets posted by her on $\forall\ d \in TimeWindow(u, d)$. We used Twitter Advanced Search[9] utility that can retrieve tweets posted by a user, given her screen name and a particular $TimeWindow(u, d)$. Our communal tweet detection algorithm is applied on these tweets which marked the retrieved tweets as communal and noncommunal. Based on the classification, we define a vector v for each user u as follows:

$$v[i] = \begin{cases} 1 & \text{if user u posted a communal tweet on } d+i \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where $i \in [-15, 15]$.

Next, for each user u, we find her *regularity score*, $r_u = \sum_{i=-15}^{15} v_u[i]$, where r_u defines the number of days user u posted a communal tweet in her $TimeWindow(u, d)$. Fig. 3 shows the cdf of regularity score for NEQuake, GShoot, and CShoot. From Fig. 3, we observe the following two interesting phenomena.

1) Most of the users (80–90%) have regularity score <5.
2) There are a small fraction of users (10–20%) having large values of regularity score (≥ 5).

[9] https://twitter.com/search-advanced

TABLE XVII
EXAMPLES OF ANTICOMMUNAL TWEETS POSTED DURING DISASTERS

Event	Tweet text
NEQuake	Sad commentary of our times that people bring religion even into the devastating
GShoot	A terrorist has no religion. No need 2 specifically mention d religion of a terrorist anywhere
PAttack	Tears & blood know no religion. All they know is pain. It's not just in Paris. It's everywhere. We are the killers & we are the victims.
CShoot	#California So sorry to hear abt the shooting & Killings of innocent people. There is no religion which allows that.

Thus, a large fraction of users only get outraged at the time of disaster and do not express their hatred toward people of a particular religion or race otherwise. However, there are a few users who repeatedly post communal tweets irrespective of any trigger event. We define them as *regular communal users*. These phenomena also agrees with what prior works found [37].

1) Overlap Between Core Communal Users and Regular Communal Users: We next find the overlap between core communal users (Section V-B4) and regular communal users using *Szymkiewicz-Simpson similarity* [31]. For Regular communal users with $r_u >= 5$, we find an overlap score of 0.44 and for $r_u >= 10$, we obtain an overlap score of 0.22. These regular communal users play the role of core communal users in posting communal tweets across multiple events.

VI. COUNTERING COMMUNAL TWEETS DURING DISASTER SCENARIO

During a disaster event, when the masses are anxious, communal tweets may propagate venom among different religious communities and thus complicate the relief operations. Since OSM like Twitter work as important sentinels during disasters [22], shutting down online media during disasters is not a reasonable solution. On the other hand, if communal content is allowed to circulate freely and get large exposure, antigovernment agencies can use such communal content for propaganda, causing certain religious communities to panic.[10] Hence, communal tweets posted during disasters need to be countered, so as to minimize their potential adverse effects. In this section, we discuss a potential way of countering the communal tweets.

Utilizing Anticommunal Tweets: During disasters, most of the people post communal tweets. However, it is observed that some users also post *anticommunal* content, asking people not to spread communal venom among society. Some examples of anticommunal tweets posted during different disaster events considered in this paper are shown in Table XVII. We also found that just as some communal hashtags are introduced to target certain religious communities, certain other hashtags are introduced to *support* those religious communities. Table XVIII shows some examples of hashtags of both types.

[10]For instance, after the mass shooting incident in California in November 2015, the American Muslims had to live in fear of demonization of Islam, according to the report by Reuters—https://t.co/GzMonqK9Js.

TABLE XVIII
EXAMPLES OF COMMUNAL AND ANTICOMMUNAL HASHTAGS, WHICH ARE USED TO ATTACK OR SUPPORT CERTAIN RELIGIOUS COMMUNITIES DURING DISASTERS

Event	Anti-communal hashtags	Communal hashtags
NEQuake	#RespectAllReligion, #Intolerance, #stopit	#SoulVultures, #EvangelicalVultures, #EvanJihadis
PAttack	#MuslimsAreNotTerrorist, #ThisisNotIslam, #NothingToDoWithIslam	#KillAllMuslims, #IslamAttacksParis, #RadicalIslam

Thus, a potential way of countering communal content would be to utilize such anticommunal content. For this, first question arises about automatic identification of such anticommunal tweets.

A. Identifying Anticommunal Tweets

In Section IV, we have proposed a rule-based classifier to detect communal tweets from large set of tweets. After separating out communal tweets, we try to capture anticommunal tweets from rest of the tweets.

1) Establishing Gold Standard: To understand the pattern of anticommunal tweets and define the rules for its detection, we require gold standard annotation for a set of tweets. For each event, first, we used the communal tweet classifier (proposed in Section IV) to identify communal tweets. Then, from rest of the tweets, we randomly sampled 2000 tweets (after removing duplicates). These tweets were independently observed by three human volunteers, all of whom are regular users of Twitter, have a good knowledge of English. The volunteers were asked to identify whether a tweet is anticommunal or not.

There was an unanimous agreement for 78% tweets, while we consider the majority decision for the rest. By this process, a total of 196 tweets were identified as anticommunal. We can observe that very less number of anticommunal tweets are posted during such events. In fact, we were able to identify anticommunal tweets only for three events—NEQuake, GShoot, and PAttack. For the other two events, no example of anticommunal tweet was found. Some examples of anticommunal tweets are shown in Table XVII. From the rest of the tweets, we randomly sampled the same number of non-anticommunal tweets to build our training data set.

2) Features for Classification: As mentioned earlier, our main objective is to make our classifier independent of any specific event, i.e., the classifier should be such that it can be directly used over tweets posted over later events without further training. Following communal tweet classifier approach, in this section also, we rely on using a set of lexical and content features for the classification task. We describe the features next.

a) Presence of anticommunal hashtags: While observing the three data sets, the annotators found that some specific hashtags are explicitly used across various events to post anticommunal tweets and ask users not to post communal contents, such as "#RespectAllReligion," "#MuslimsAreNotTerrorist," "#ThisisNotIslam," and "#NothingToDoWithIslam," "#stopit."

TABLE XIX

CLASSIFICATION ACCURACIES (AC), RECALL (R), AND F-SCORES (F) FOR ANTICOMMUNAL TWEETS,USING BOW MODEL

Train set	Test set								
	NEQuake			GShoot			PAttack		
	AC	R	F	AC	R	F	AC	R	F
NEQuake	0.8083	0.8166	0.7990	0.6875	0.8958	0.7413	0.7180	0.7067	0.7148
GShoot	0.56	0.76	0.6333	0.6875	0.9199	0.7382	0.5751	0.8947	0.6780
PAttack	0.6999	0.80	0.7272	0.6041	0.9166	0.6984	0.7596	0.8499	0.7828

TABLE XX

CLASSIFICATION SCORES (PRECISION, RECALL, AND F-SCORE) FOR ANTICOMMUNAL TWEETS AND OVERALL ACCURACY USING RULE-BASED CLASSIFIER WITH PROPOSED FEATURES

Event	Precision	Recall	F-score	Accuracy
NEQuake	0.8461	0.88	0.8627	0.86
GShoot	0.6351	0.9791	0.7704	0.7083
PAttack	0.8012	1	0.8896	0.8759

TABLE XXI

MISCLASSIFIED ANTICOMMUNAL TWEETS POSTED DURING DISASTERS

Could someone on the ground please ask about #gods involvement concerning the #NepalEarthquake ? Just curious.
earthquakes happen because of tectonic plates, they are not a result of lack of jesus. Christians and science, smh
Islam has nothing to do with #GurdaspurAttack. Stop spreading hatred among society

b) Presence of collocations: Some collocations are frequently used in anticommunal tweets across all three data sets, such as "nature doesn't discriminate," "has no religion," "terrorism defies religion," etc.

c) Mentioning multiple religious terms: The aim of anticommunal tweets is to ask people to treat all religions equally. Thus, either they do not mention religious terms explicitly or they mention multiple religions so as to create a sense of unity, e.g., "WTF people are trying to save their life and this MORONs Tweeting *Hindu Christian Muslim* #earthquake #NepalEarthquake."

We make the above-mentioned lexicons publicly available to the research community at `http://www.cnergres.iitkgp.ac.in/disasterCommunal/dataset.html`. In the future, we will try to enrich this lexicon set based on co-occurrence with current lexicons. We follow a simple rule-based classification approach to classify the tweets into two classes based on the features described above. If any of the above-mentioned features is present in a tweet, we mark that tweet as anticommunal; otherwise non-anticommunal.

3) Evaluating Classification Performance: We compare our proposed features (PRO) with the BOW model where we take unigrams as classification features and Naive-Bayes as classifier. Prior research [38] showed that the Naive-Bayes model performs better compared to others when unigrams and bigrams are chosen as features. BOW is a supervised model; hence, requires training. Our proposed method is rule based and can be applied directly to any future event. Table XIX shows the accuracies (AC) of the classifier using the BOW model and Table XX shows recall, F-score of anticommunal tweets, and overall accuracy of our proposed rule-based classifier. We compare the performance of two feature sets with different classification models (rule based and Naive-Bayes based). The BOW model achieves 75% in-domain accuracy (training and testing events are same) but does not perform well in cross-domain setting (training and testing events are different). Our proposed method performs better compared to vocabulary-dependent model.

4) Analyzing Misclassified Tweets: For our proposed method, we have also analyzed different types of errors i.e., how many times an anticommunal tweet is marked as non-anticommunal tweet or vice versa. We achieve precision of 0.76 over three data sets, which indicates around 24% non-anticommunal tweets are marked as anticommunal tweets. On the other hand, Table XX reflects that average recall score is 0.95. 5% of anticommunal tweets are misclassified as non-anticommunal tweets. It is observed that during disaster anticommunal tweets are posted in very low volume compared to other tweets. Hence, objective of the classifier should be high recall so that we can utilize such tweets in maintaining communal harmony during emergency. Table XXI shows some example of misclassified anticommunal tweets. In most of the cases, explicit signal for anticommunal tweets are missing. In the future, we will try to capture such implicit senses and also try to enhance our feature sets.

B. Characterizing Anticommunal Tweets and Its Users

In this section, we study the anticommunal tweets and the users who post them. We apply the classifier described in the previous section, over the data sets; tweets which are identified as anticommunal by our classifier are referred as *anticommunal tweets* and the users who posted them as *anticommunal users*. Specifically, we compare the set of anticommunal tweets and anticommunal users during a particular event with an equal number of randomly sampled communal tweets (as judged by our classifier) and the users who posted them (referred to as communal users) during the same event.

1) Do Anticommunal Tweets Get Similar Exposure as Communal Tweets?: As earlier, we measure the exposure or popularity of a tweet by its retweet count. Fig. 4 shows the distributions of retweet count of communal and anticommunal tweets posted during two of the disaster events. We observe that anticommunal tweets are significantly *less retweeted* compared to communal tweets. We obtained a similar observation across all events.

We next investigate why anticommunal tweets get less popularity compared to communal tweets. Our first intuition was that the users who post communal tweets might be more popular than the ones who post anticommunal tweets. To verify this, we compared the distributions of follower counts of users who post communal tweets and users who post anticommunal tweets during the same event. Fig. 5 shows the comparison for two events (similar results were obtained for all other events). It is clear that both sets of users have very similar follower counts. Thus, variation in user-popularity

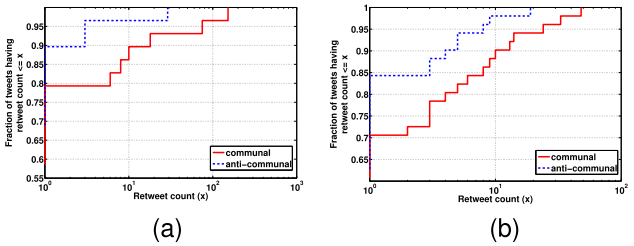

Fig. 4. Comparing the popularity of communal and anticommunal tweets—communal tweets are much more retweeted than anticommunal tweets. (a) NEQuake. (b) GShoot.

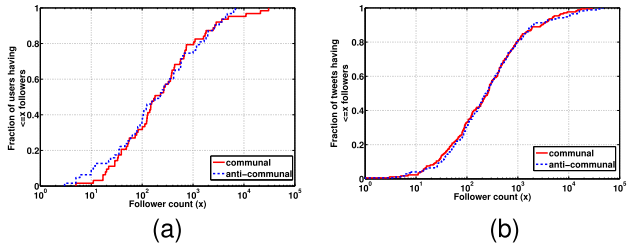

Fig. 5. Comparing the popularity of users who post communal tweets and those who post anticommunal tweets—both types of users have similar follower count distributions. (a) NEQuake. (b) GShoot.

cannot explain why anticommunal tweets get lower exposure than communal tweets.

We find that the number of distinct users who post anticommunal tweets is much lesser than the number of users who post communal tweets. As a result, other users receive much lesser exposure to such tweets. We believe that an effective way of countering communal content would be to automatically identify anticommunal tweets, and to promote such tweets by getting more and more users (preferably popular users) to retweet them. Additionally, proper wording of tweets are also necessary to make them popular. In the future, we will try to promote and increase the popularity of such anticommunal tweets.

VII. DisCom: Communal Tweet Identifier During Disaster

As stated earlier, the focus of the research community has been mostly on the situational information posted in Twitter during a disaster scenario, such as extracting and summarizing situational tweets [22], [39], [40]. There exist online systems to classify situational tweets [39] whereas there is no existing service to identify communal and anticommunal tweets from the large collection of tweets. Based on this identification, a system can filter communal tweets and take necessary actions to promote anticommunal tweets. Hence, we have developed *DisCom* (`http://www.cnergres.iitkgp.ac.in/projects/disaster_communal_identifier/`), a service where one can collect tweets corresponding to a disaster scenario based on keywords and hashtags (e.g., #NepalEarthquake in the case of Nepal earthquake), identify communal and anticommunal tweets and accordingly take necessary actions like filtering or promoting some contents.

To evaluate the quality of our identified communal and anticommunal tweets, we used human feedback since judgment of a tweet as communal or anticommunal is subjective in nature. The evaluators were shown 50 communal and anticommunal tweets (randomly sampled from the whole identified set), and were asked to judge whether a communal or anticommunal tweet is really so or not. 15 human volunteers (institute undergraduate students) individually judged 50 communal and anticommunal tweets identified by our service from the PAttack event. Out of 50 tweets, more than 80% and 70% were judged as proper communal or anticommunal tweet, respectively, by *all* the evaluators.

VIII. Conclusion

To our knowledge, this paper is the first attempt in the direction of characterizing communal tweets posted during the disaster scenario and analyzing the users involved in posting such tweets. We proposed an event-independent classifier that can be used to filter out communal tweets early. We also found that communal tweets are retweeted heavily and posted by many popular users; mostly belong to news media and politics domain. Users involved in initiating and promoting communal contents form a strong social bond among themselves. Additionally, most of the users get angry suddenly due to such kind of events and express their hates to specific religious communities involved in the event. We observe that, during a disaster, some users also post anticommunal content asking people to stop spreading communal posts, and it is necessary to counter the potential adverse effects of communal tweets. We have proposed an event-independent classifier to identify such anticommunal tweets. However, we have found such anticommunal tweets are retweeted much less compared to communal tweets and they are also very few in number compared to communal tweets. Finally, we proposed a real-time system DisCom which can be used directly in the future disaster events to identify communal and anticommunal tweets.

A. Limitations of the Study

Our work has some limitations as follows.

1) We collected only English tweets posted during disaster events using some specific event based keywords. Hence, some domain-specific biases may exist in the data set. Additionally, the features for communal tweet classifier were developed based on the analysis of disaster-specific tweets. Hence, some of the features like "presence of wh-words/intensifiers with religious terms" may not be suitable for any general kind of event. Side by side, tweets posted in other languages may contain different kinds of patterns as compared to English tweets.
2) The users analyzed in this paper are also identified from the data set collected through keyword search. Hence, there might be some bias among these users as well.
3) We observed that number of distinct anticommunal tweets is much less in number. In this paper, we are able to collect around 200 such tweets (from 6000 annotated

tweets) from three data sets. Side by side anticommunal tweets do not follow any specific pattern and it varies across disasters. In this paper, we captured some common collocation phrases, hashtags used for such kind of tweets. However, this number is less due to availability of small number of anticommunal tweets. In the future, we will try to enlarge our proposed set of lexicons.

4) Tweets are known to be informally written and noisy in nature, containing misspellings, abbreviations etc. In the future, we will handle these variations to improve our classifiers.
5) In this paper, we found that some users post more communal tweets after a disaster event, as compared to before the event. Any kind of "trigger events" like disasters may increase the volume of social media activity in general. However, due to lack of data collected before an event, we could not check whether the increase in communal posts is proportional to the overall increase in activity in Twitter after such an event.

B. Future Directions

We believe that our present study has many potential future applications. For instance, the proposed communal tweet classifier can be used as an early warning signal to identify communal tweets, and then celebrities, political personalities can be made aware of the situation and requested to post anticommunal tweets so that such tweets get higher exposure. We need to promote anticommunal content via mentioning popular celebrities, political persons. Our real-time system DisCom can be used by the Government in taking decisions regarding filtering communal content, promoting anticommunal content etc. We plan to pursue some potential directions of countering communal tweets in the future. This paper also raises many intriguing social questions like "interaction between communal and anticommunal users," "demographic biases," etc. We will try to address these questions in the future.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers whose suggestions greatly helped to improve this paper.

REFERENCES

[1] 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.
[2] I. Chaudhry, "#Hashtagging hate: Using Twitter to track racism online," *First Monday*, vol. 20, no. 2, 2015. [Online]. Available: http://firstmonday.org/ojs/index.php/fm/article/view/5450
[3] L. A. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber, "Analyzing the targets of hate in online social media," in *Proc. ICWSM*, Mar. 2016, pp. 687–690.
[4] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, "A lexicon-based approach for hate speech detection," *Int. J. Multimedia Ubiquitous Eng.*, vol. 10, no. 4, pp. 215–230, 2015.
[5] I. Kwok and Y. Wang, "Locate the hate: Detecting tweets against blacks," in *Proc. 27th AAAI Conf. Artif. Intell.*, 2013, pp. 1621–1622.
[6] M. Mondal, L. A. Silva, and F. Benevenuto, "A measurement study of hate speech in social media," in *Proc. ACM HT*, 2017, pp. 85–94.
[7] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embeddings," in *Proc. WWW*, 2015, pp. 29–30.
[8] W. Magdy, K. Darwish, N. Abokhodair, A. Rahimi, and T. Baldwin, "#ISISisNotIslam or #DeportAllMuslims?: Predicting unspoken views," in *Proc. ACM Web Sci.*, 2016, pp. 95–106.
[9] K. Rudra, A. Sharma, N. Ganguly, and S. Ghosh, "Characterizing communal microblogs during disaster events," in *Proc. IEEE/ACM ASONAM*, Aug. 2016, pp. 96–99.
[10] E. Greevy and A. F. Smeaton, "Classifying racist texts using a support vector machine," in *Proc. SIGIR*, 2004, pp. 468–469.
[11] N. Pendar, "Toward spotting the pedophile telling victim from predator in text chats," in *Proc. ICSC*, Sep. 2007, pp. 235–241.
[12] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *Proc. Int. Conf. Social Comput. Privacy, Secur., Risk Trust (PASSAT), (SocialCom)*, Sep. 2012, pp. 71–80.
[13] 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.
[14] P. Burnap *et al.*, "Tweeting the terror: Modelling the social media reaction to the Woolwich terrorist attack," *Social Netw. Anal. Mining*, vol. 4, no. 1, p. 206, 2014.
[15] N. Alsaedi, P. Burnap, and O. Rana, "Can we predict a riot? Disruptive event detection using Twitter," *ACM Trans. Internet Technol.*, vol. 17, no. 2, p. 18, 2017.
[16] P. Burnap and M. L. Williams, "Us and them: Identifying cyber hate on Twitter across multiple protected characteristics," *EPJ Data Sci.*, vol. 5, no. 1, p. 11, 2016.
[17] R. Delgado and J. Stefancic, "Hate speech in cyberspace," *Wake Forest Law Rev.*, vol. 49, p. 319, Jan. 2014.
[18] K. Jaishankar, "Cyber hate: Antisocial networking in the Internet," *Int. J. Cyber Criminol.*, vol. 2, no. 2, pp. 16–20, 2008.
[19] C. Schieb and M. Preuss, "Governing hate speech by means of counterspeech on Facebook," in *Proc. 66th ICA Annu. Conf.*, Fukuoka, Japan, 2016, pp. 1–23.
[20] E. Chandrasekharan, M. Samory, A. Srinivasan, and E. Gilbert, "The bag of communities: Identifying abusive behavior online with preexisting Internet data," in *Proc. ACM CHI*, 2017, pp. 3175–3187.
[21] K. Gimpel *et al.*, "Part-of-speech tagging for Twitter: Annotation, features, and experiments," in *Proc. ACL/HLT*, 2011, pp. 42–47.
[22] K. Rudra, S. Ghosh, N. Ganguly, P. Goyal, and S. Ghosh, "Extracting situational information from microblogs during disaster events: A classification-summarization approach," in *Proc. ACM CIKM*, 2015, pp. 583–592.
[23] *Docs—Twitter Developers*, 2017. [Online]. Available: https://dev.twitter.com/docs/api
[24] S. Volkova, T. Wilson, and D. Yarowsky, "Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual Twitter streams," in *Proc. ACL*, Aug. 2013, pp. 505–510.
[25] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[26] A. Rajadesingan, *Sarcasm Detection on Twitter: A Behavioral Modeling Approach*. Tempe, AZ, USA: Arizona State Univ., 2014.
[27] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter," in *Proc. 43rd Hawaii Int. Conf. Syst. Sci. (HICSS)*, Jan. 2010, pp. 1–10.
[28] N. Azman, D. Millard, and M. Weal, "Patterns of implicit and non-follower retweet propagation: Investigating the role of applications and hashtags," in *Proc. Web Sci.*, 2011, pp. 1–4.
[29] S. P. Ponzetto and M. Strube, "Knowledge derived from wikipedia for computing semantic relatedness," *J. Artif. Intell. Res.*, vol. 30, pp. 181–212, Sep. 2007.
[30] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in Twitter: The million follower fallacy," in *Proc. ICWSM*, 2010, pp. 1–8.
[31] (2017). *Szymkiewicz-Simpson Coefficient*. [Online]. Available: https://en.wikipedia.org/wiki/Overlap_coefficient
[32] N. K. Sharma, S. Ghosh, F. Benevenuto, N. Ganguly, and K. Gummadi, "Inferring who-is-who in the Twitter social network," in *Proc. ACM Workshop Online Social Netw. (WOSN)*, 2012, pp. 55–60.
[33] E. M. Rogers, *Diffusion of Innovations*. New York, NY, USA: Simon and Schuster, 2010.
[34] E. Katz and P. F. Lazarsfeld, *Personal Influence: The Part Played by People in the Flow of Mass Communications*. Piscataway, NJ, USA: Transaction Publishers, 1966.

[35] B. Wang *et al.*, "Whom to mention: Expand the diffusion of tweets by @ recommendation on micro-blogging systems," in *Proc. WWW*, 2013, pp. 1331–1340.
[36] I. Awan and I. Zempi, "'I will blow your face OFF'—VIRTUAL and physical world anti-muslim hate crime," *Brit. J. Criminol.*, vol. 52, no. 2, pp. 362–380, 2017.
[37] M. Williams and O. Pearson. (2016). *Hate Crime and Bullying in the Age of Social Media*. [Online]. Available: http://orca.cf.ac.uk/88865/
[38] S. Verma *et al.*, "Natural language processing to the rescue? Extracting 'situational awareness' tweets during mass emergency," in *Proc. ICWSM*, 2011, pp. 385–392.
[39] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, "AIDR: Artificial intelligence for disaster response," in *Proc. WWW*, 2014, pp. 159–162.
[40] K. Rudra, S. Banerjee, N. Ganguly, P. Goyal, M. Imran, and P. Mitra, "Summarizing situational tweets in crisis scenario," in *Proc. 27th ACM Conf. Hypertext Social Media (HT)*, 2016, pp. 137–147.

Koustav Rudra received the B.E. degree in computer science from the Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India, and the M.Tech degree from IIT Kharagpur, India. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, IIT Kharagpur, Kharagpur, India.

His current research interests include social net-works, information retrieval, and data mining.

Ashish Sharma is currently pursuing the Dual degree with the Department of Computer Science and Engineering, IIT Kharagpur, Kharagpur, India.

His current research interests include social networks and information retrieval.

Niloy Ganguly received the B.Tech. degree from IIT Kharagpur, Kharagpur, India, and the Ph.D. degree from the Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India.

He was a Post-Doctoral Fellow with Technical University, 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, and mobile systems.

Saptarshi Ghosh received the B.E. degree in computer science from the Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India, and the Ph.D. degree from IIT Kharagpur, Kharagpur, India.

He was a Humboldt Post-Doctoral Fellow with the Max Planck Institute for Software Systems, Saarbruecken, Germany. He is currently an Assistant Professor with the Department of Computer Science and Engineering, IIT Kharagpur. His current research interests include social media, complex networks, data mining, and information retrieval.