Characterizing and Utilizing Microblogs during Emergency Scenarios

A thesis submitted for

B. Tech. Project

in

Computer Science and Engineering

by

Ashish Sharma (13CS30043)

under the guidance of

Prof. Niloy Ganguly

and

mentorship of

Koustav Rudra

Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

May 2017

Certificate

This is to certify that the thesis titled **Characterizing and Utilizing Microblogs during Emergency Scenarios** submitted by **Ashish Sharma (13CS30043)** to the Department of Computer Science and Engineering is a bona fide record of work carried out by him under my supervision and guidance. The thesis has fulfilled all the requirements as per the regulations of the Institute and, in my opinion, has reached the standard needed for submission.

Niloy Ganguly
Professor
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
May 2017

Abstract

Recent emergence of social media platforms such as Twitter provide convenient ways and fast access to disseminate and consume information to/from a wider audience. As a result, data in huge volume and at rapid rates is posted on such platforms specially during emergency scenarios like an epidemic or a disaster, which if utilized effectively can be helpful to the common users as well as professionals. In this project, we first develop models for classifying and summarizing microblogs during epidemic. We observe that vulnerable users look for known symptoms & preventive measures and affected users look for treatment strategies. On the other hand, health organizations try to get situational updates to assess the severity of the outbreak, known affected cases, and other details. Thus, we classify microblogs into one of these categories using Lexical and Latent features and extract information from the individual categories depending on the target users. Similar models can be applied during disaster focusing more on updates. However, we observe that during disaster, some posts target people belonging to a certain community/race blaming them for the crisis. We call these tweets as *Communal Tweets*. Detection of these tweets and a way to counter them is necessary as such tweets can disrupt the spirit of unity among people. Thus, we first identify these tweets, and then find their characteristics. We show that communal tweets gain more popularity than the non-communal tweets, and also at a much faster rate. We find that communal tweets are posted not only by common users, but also by many popular users (having large no. of followers), most of whom are related to media and politics. We also find that common masses try to provoke popular users by mentioning them in their tweets. As a result, communal tweets get much higher exposure than non-communal tweets. Considering the potentially adverse effects of communal tweets during disasters, we also indicate a way to counter such tweets, by utilizing anti-communal tweets posted by some users during such events.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, Social Networking platforms such as Twitter, Facebook, online forums and message boards have become important sources of real-time information, especially during health epidemic, disasters and other such wide-ranging crisis. The information during such scenarios is huge in volume and ranges from death estimates, relief strategies to sentiments and opinions of the people. The rate at which this information is posted is very high and this encourages researchers to propose methodologies for efficiently utilizing it so that this diverse data can be effectively used by common users and professionals alike. An ideal support system aimed for crisis should first classify the data, then extract relevant information from the individual classes, present the extracted information in a user-friendly format and finally, filter out the posts creating a feeling of negativity and hatred in users. Such a system can be helpful in providing instant updates, instructions, etc. to people and could help the officials in decision-making.

To start with, let us first describe the data posted on these platforms during epidemic and during disaster.

1.1 Social Media Posts during Epidemic

Information posted on these platforms during an epidemic is multi-dimensional and provides rapid access to diverse and useful insights helpful to understand the various aspects of the disease outbreak. This includes the following major type of posts - (i) Disease Signs and

Symptoms, (ii) Prevention Techniques, (iii) Treatment Mechanisms, (iv) Medium of Transmission, (v) Death Reports, and (vi) Others (User Sentiments, Reactions, etc.). Furthermore, the information to be extracted from posts belonging to these categories can depend upon the specific information needs of the target users(health organizations or affected/vulnerable communities). Based on the information need, the target users can be classified as follows — (i) **Pro-disease Users**: The uninfected users primarily looking for signs or symptoms for understanding the traits of the disease and necessary preventive measures which can be taken, (ii) **Post-disease Users**: Users diagnosed with the disease and looking for effective methods of treatment, (iii) **Monitoring community**: Government and Health Organizations figuring out general situational updates like the number of infected people, number of deaths and the way they can carry out relief operations.

1.2 Social Media Posts during Disaster

As observed by Rudra et al [54], Social Media posts during a disaster event can be classified into 2 broad categories: *Situational Posts* and *Non-Situational Posts*. Situational posts are the one which provide you status updates like number of casualties, the current situation in various regions affected by disaster and information regarding relief operations going on. Non-Situational posts, on the other hand, comprise of posts reflecting sentiments of the people sympathizing with the victims; posts praising / criticizing the relief operations; posts analyzing the event and about how similar incidents can be prevented in future and posts related to charities being organized. Also fall in this category are the posts which target a particular religious or racial community, which we call *Communal Posts*. These posts are targeted towards certain religious or racial communities, most of the time blaming them for the calamity and are potentially harmful & dangerous. These type of posts create a sense of fear among the people as there is a possibility of online threats getting materialized in the “real world” [3]. It has been observed earlier that such posts are often posted during man-made disasters, such as terrorist attacks. For instance, [9] observed that during Woolwich attack, the UK masses targeted a certain religious community to which the attackers belonged. However, we surprisingly observe that communal posts are also posted during natural disasters like floods and earthquakes, at least in certain geographical regions such as the Indian subcontinent.

In this project, we use posts coming from the popular microblogging site - Twitter. We first focus our attention on tweets posted during 2 epidemic. We classify the epidemic tweets into 6 broad categories using low-level lexical and latent features. Using the classified tweets, we extract relevant targeted information based on information need.

We next switch to tweets posted during 5 disaster events. The classification of disaster posts into Situational and Non-Situational Tweets and extracting information from situational tweets is an area well-studied by many researchers. However, to the best of our knowledge, no one has tried in-depth characterization of Non-situational tweets. Thus, we focus our attention on Communal Tweets posted during disaster. We develop a classifier for automatically separating out communal tweets from non-communal ones. Keeping in mind the limitations of previous works [9] [22], we develop an event-independent communal tweet classifier which can be trained on past events and used to filter out communal tweets during future events. The classifier relies on a proposed set of low-level lexical and syntactic features extracted from the contents of tweets.

Using the identified communal tweets in our dataset, we check if these tweets are popular enough to cause a real threat to the masses. We then shift our attention to the communal users who post them. We find that, alarmingly, communal tweets are posted not only by common, random users but also by some very popular users who are influential. Interestingly, most of these popular users who post communal tweets, belong to either media houses or have interest in politics. We also indicate a potential way of countering the spread of such communal content. We observe that a small number of users post anti-communal content which aim to maintain peace and harmony. However, such anti-communal tweets usually receives lot less exposure than communal content. Promoting anti-communal tweets can be a promising way to counter communal tweets posted during disasters and we propose a method to identify them.

Chapter 2

Related Work

2.1 Information Extraction for Healthcare

Fox et al [25] reported in their study that 80% of Internet users look online for getting information on various health-related topics such as disease symptoms, diagnosis or treatment. Further, it was shown by Elkin [23] that 34% of health searchers use social media resources to find health-related topics through Web 2.0 sites. The use of these resources becomes more prevalent during an epidemic for which biomedical literature, clinical notes and other information sources are absent. The popularity of social media in medical and health domain has gained attention from various researchers in the past. We here present the type of researches which have been conducted for utilizing medical social media data in order to extract meaningful information and show their differences with the traditional systems used for clinical notes.

2.1.1 Mining information from Clinical Notes

Various methodologies have been proposed for mining health and medical information from clinical notes. Most of these works have focused on extracting a broad class of medical conditions (eg, diseases, injuries, and medical symptoms) and responses (eg, diagnoses, procedures, and drugs), with the goal of developing applications that improve patient care [31] [26] [27] [35]. The 2010 i2b2/VA challenge [61] presented the task of extracting medical concepts, tests and treatments from a given dataset. Out of the top 10 submissions, 7 used Conditional Random Fields (CRF) or a similar sequence classifier. A number of other submissions proposed rule-based methods for performing this extraction task. Roberts et al [52] later built a flexible framework for identifying medical concepts in clinical text, and classifying assertions, which indicate the existence, absence, or uncertainty of a medical problem. The framework was evaluated on the 2010 i2b2/VA challenge data. Goodwin et al [29] recently utilized this framework for building a clinical question-answering system. It used a probabilistic knowledge graph, generated from electronic medical records (EMRs), in order to carry out answer inference.

2.1.2 Problems with methods proposed for clinical notes

Most of the methods proposed for extracting information from clinical text map clinical documents to concepts of medical terminologies and ontologies (eg. UMLS [6], SNOMED CT [59]) using earlier proposed systems (eg. MetaMap [1], cTakes [55]). For a given text, these systems provide extracted concepts of clinical terminologies that can be used to describe the content of a document in a standardised way. However, tools like MetaMap were designed specifically to process clinical documents and are thus, specialised to the linguistic characteristics of these type of documents [18]. The user-generated medical text from social media differs significantly from professionally written clinical notes. Recent studies have shown that directly applying Metamap on Social Media Data is challenging and leads to low quality word labels [60]. There have also been works which propose methods for identifying the kind of failures MetaMap experiences when applied on social media data. In a recent study, [45] characterized failures of MetaMap into Boundary Failures, Missed Term Failures and Word Sense ambiguity failures.

2.1.3 Mining information from Medical Social Media Data

Scanfield et al [56] used Q-Methodology [65] to determine the main categories of content contained in Twitter users' status updates mentioning antibiotics. Lu et al [43] built a framework based on clustering analysis technique to explore interesting health-related topics from online health community. It utilized sentiment based features extracted from SentiWordNet [24] and domain specific features from MetaMap. Denecke et al [20] performed a comprehensive content analysis of different health-related Web resources. It also classified medical weblogs according to their information type using features extracted from MetaMap. A framework based on Latent Dirichlet Allocation (LDA) to analyze discussion threads in a health community was proposed by Yang et al [68]. They first extracted medical concepts, used a modified LDA to cluster documents and finally performed sentiment analysis for each conditional topic. Recently large scale researches have been done in exploring how microblogs can be used to extract symptoms related to disease [46], mental health [32] and so on.

Most of the above works fail to address the challenges medical social media data presents. In addition to this, all of the above researches deal with static datasets and use them for a defined task. None of them deal with extracting information during a new disease outbreak. Information Extraction from the social media data of a disease outbreak is challenging as it consists of new preventive and treatment measures and lacks any standard lexicon of medical concepts of the disease. The prior works on disease outbreaks have mainly focused on modelling their detection and spread [51] [17] [41]. To the best of our knowledge, there is no work which has extended outbreak detection to gather useful information related to the outbreak. The research has been limited to classifying the tweets based on content (informative, opinion, experience) & sentiment and using this classification to predict outbreak [15] [16].

2.2 Identification and Characterization of Communal Tweets

Online forums are increasingly being used by the masses to post hate speeches and offensive content. Hence, there have been lot of effort in recent years for automatic identification of such offensive content [9], [12], [58], [28], [40]. We briefly discuss such studies, and point out how the second part of this project is different from the prior works.

Several studies have attempted to identify online content which are potentially hate speeches or offensive in nature. For instance, Greevy et al. [30] classified racist content in webpages using a supervised Bag-Of-Words (BOW) model. This approach was later improved in [47], where some context features were incorporated along with predictive words to improve classification accuracy. Dinakar et al. [21] identified cyberbullying, using features like profane words, parts-of-speech tags, words with negative connotations, and so on. Similarly, Chen et al. [14] used profanities, obscenities, and pejorative terms as features with appropriate weightage to identify offensive content in Youtube comments. Wang et al. [64] investigated the use of curse words in the context of Twitter and proposed a method to identify them. Mahmud et al. [44] identified insulting syntactic constructs, relationship between terms to detect online flaming behaviour. More recently, Burnap et al. [8] [9] [10] attempted to detect hate speech posted during a disaster event (the Woolwich attack).

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. People target specific characteristics of users like gender, race, sex, nationality, religion, ethnicity, and so on. Prior studies show that most prevalent hate speech is targeted towards certain races, while religion-induced hate speech is very sparse [58]. 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 for a short period of time such category of tweets are tweeted in huge number. As an example, classifier proposed by [58] can hardly capture communal tweets. Consequently, in recent times, researchers focus on more granular levels of hate speech detection in Twitter. For example, Chaudhry [12] tried to track racism in Twitter and Burnap et al. [9] detected religious hate speeches posted during Woolwich attack.

Second, most of the prior studies on hate speech have focused on content posted in blogs or webpages [22] [28]. On the contrary, this study focuses on Twitter, and it has been widely demonstrated that standard Natural Language Processing-based methodologies, that have been developed for formally-written text, do not work well for short, informal tweets. Hence new methodologies are necessary to deal with noisy content posted on Twitter.

Burnap et al. detected hate speech (religious, racial) posted during the Woolwich attack using a bag-ofwords model, where n-grams containing specific hate terms and some dependencies like ‘det’ (determiner) and ‘amod’ (adjectival modifier) are considered as features. However, the bag-of-words 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. The vocabularies corresponding to 2 separate events are completely different. As a result, a bag-of-words based classifier is unlikely to perform well if trained on one of these events and used on the other. On the other hand, using low-level lexical and syntactic features (instead of specific terms) can make the classifiers performance largely independent of specific disaster events considered for training [54]. This motivated us to propose an event-independent classifier for identifying communal tweets. Finally, almost all prior works have focused on identifying offensive content and hate speech, and there has been very few efforts towards characterizing the users who post such contents. To the best of our knowledge, the recent study by Silva et al [58] is the only one which attempted to identify the sources and targets of such hate speeches. However, there has not been any detailed effort in characterizing users who post such content and extracting features from it. In this project, we take the first step in this direction by characterizing the users posting communal tweets based on their popularity, interests, social interactions and behaviour.

Chapter 3

Classifying Microblogs during Epidemic

Let us first look into how the tweets coming from Microblogs during an epidemic can be classified into different categories. We follow a supervised machine learning approach. We first create a gold standard dataset through manual annotation, extract lexical and latent features from the tweets and train a Support Vector Machine on the gold standard dataset.

3.1 Dataset

We used tweets from 2 recent epidemics - *Ebola* and *MERS*. The tweets were extracted from Twitter using AIDR platform [36]. The details of the 2 dataset are as follows:

1. **Ebola**: Ebola, previously known as Ebola hemorrhagic fever, is a rare and deadly disease caused by infection with a virus of the family *Filoviridae*, genus *Ebolavirus* [1]. The recent West African Ebola virus epidemic from 2014–2016 was the most widespread outbreak of Ebola virus disease (EVD) in history—causing major loss of life and socioeconomic disruption in the region [2]. We take tweets posted during this outbreak of Ebola. The tweets posted between between August 6th, 2014 and January

[1] https://www.cdc.gov/vhf/ebola/about.html

[2] https://en.wikipedia.org/wiki/West_African_Ebola_virus_epidemic

19th, 2015 were collected using different keywords (e.g., #Ebola). The dataset finally consists of 5.08 million tweets.

2. **MERS**: Middle East respiratory syndrome (MERS), also known as camel flu, is a viral respiratory infection caused by the MERS-coronavirus (MERS-CoV). MERS was implicated in an outbreak in April 2014 in Saudi Arabia [3]. The tweets posted between April 27th, 2014 and July 16th, 2014 were collected using different keywords (e.g., #MERS). This dataset consists of 0.215 million tweets.

3.2 Term Dictionaries

In addition to the dataset, we also created the following dictionaries. We will see their use later.

1. **Symptom Dictionary (SymD)**: A dictionary of words containing the candidate symptoms of a particular disease. For creating this dictionary, we extract symptoms from various credible online sources such as Wikipedia[4], MedicineNet[5], Healthline[6], etc. These web sources have a list of symptoms observed in a large number of diseases. Combining the lists from these sources, a symptom dictionary containing approximately 770 symptoms was created.

2. **Prevention Dictionary (PreD)**: A dictionary containing terms and phrases associated with different preventive measures for diseases. An initial dictionary was created using general preventive measures which can be taken for remaining disease-free as provided by Wikipedia [7]. The dictionary was then extended by taking disease-specific preventive measures into account. The disease-specific preventive measures were obtained from Centers for Disease Control and Prevention (CDC) [8] and a Live

[3] https://en.wikipedia.org/wiki/Middle_East_respiratory_syndrome

[4] https://en.wikipedia.org/wiki/List_of_medical_symptoms

[5] http://www.medicinenet.com/symptoms_and_signs/alpha_a.htm

[6] http://www.healthline.com/directory/symptoms

[7] https://en.wikipedia.org/wiki/Preventive_healthcare

[8] https://www.cdc.gov/

Science Blog [9] covering a large number of diseases and their preventive strategies. The final dictionary contained around 300 words and phrases.

3. **Treatment Dictionary (TreD)**: A dictionary containing terms and phrases associated with different treatment mechanisms and useful drugs for medications. An initial dictionary was created containing the type of treatment mechanisms as provided by Wikipedia [10]. The dictionary was then extended by taking disease-specific treatment mechanisms into account. The disease-specific mechanisms were obtained from Centers for Disease Control and Prevention (CDC) [11] and a Live Science Blog [12] covering a large number of diseases and treatments & mechanisms associated with them. The final dictionary contained around 525 words and phrases.

4. **Transmission Dictionary (TraD)**: A dictionary containing the action verbs associated with transmission like 'transmit', 'spread', etc. and transmission mediums like 'air', 'water', etc. The dictionary is small and contains around 25 such terms.

5. **Report Dictionary (ReD)**: A dictionary containing terms associated with the death reports such as 'died', 'killed', etc. and major cities and countries of the world as obtained from Wikipedia [13]. The dictionary consists of around 300 terms.

3.3 Type of Tweets posted during epidemic

Tweets posted during epidemic can be categorized as follows:

1. **Information-Extensive Tweets**: Tweets containing health and disease-related information which can be useful for various users. These type of tweets can be further divided into following 5 classes:

   (a) **Symptoms** - Tweets containing symptoms of the disease as observed by the infected users or health professionals.

[9] http://www.livescience.com/36519-diseases-conditions-symptoms-treatments.html

[10] https://en.wikipedia.org/wiki/Therapy

[11] https://www.cdc.gov/

[12] http://www.livescience.com/36519-diseases-conditions-symptoms-treatments.html

[13] https://en.wikipedia.org/wiki/List_of_towns_and_cities_with_100,000_or_more_inhabitants

(b) **Prevention** - Tweets containing measures/actions to be taken for disease prevention.

(c) **Treatment** - Tweets containing ways the disease is being treated and medications being taken.

(d) **Transmission** - Tweets regarding the ways in which the disease can get transmitted from human-to-human and can spread in different regions of the world.

(e) **Death Reports** - Reports containing statistics of the number of people who have lost their life due to the disease. Such tweets also contain the location for which the data is being reported.

2. **Uninformative Tweets**: Tweets which do not contribute to disease awareness, most of the time containing sentiment/opinion of common people.

3.4 Classification of Tweets

We next describe the model for classifying tweets in our dataset in one of the above 6 categories.

3.4.1 Establishing Gold Standard

We start by creating a gold standard which would be used for training the model. Initially, we pick a small sample of tweets out of the total tweets in each disease corpus and manually assign them class labels. However, the number of tweets in uninformative class are much higher than those in each of the 5 informative categories. In order to ensure that this class imbalance does not affect our classification results, we make the number of tweets in the 6 classes similar by repeatedly sampling tweets from our dataset until significant number of tweets belonging to each of the 6 categories are found. We discard the large number of extra tweets from Uninformative category for tackling class imbalance. Table 3.1 shows the number of tweets in the gold standard created.

Table 3.1: Gold Standard - Number of tweets in different classes

	Symptom	**Prevent-**	**Treatm-**	**Transmi-**	**Death**	**Un- -**
		-ion	**-ent**	**-ssion**	**Reports**	**informative**
Ebola	52	69	65	59	51	56
MERS	105	70	77	74	68	84
Total	157	139	142	133	119	140

3.4.2 Feature Extraction

After creating the training dataset, we extract features corresponding to each tweet on which a classifier could work for assigning the category to which the tweet belongs. We aim to build a classifier which can be trained over tweets posted during few past epidemics and can be later used directly over tweets posted for new epidemic. Hence, we take the approach of extracting a set of lexical and latent features for the classification task, which is known to make the classifier's performance largely independent of specific events considered for training [54]. The features we used for classification are described as below:

Lexical Features

An epidemic-independent classification of tweets requires lexical resources which provide domain knowledge and terms associated with it. In the medical and health domain, multiple standardized vocabularies and ontologies are available [19]. A widely used medical knowledge resource is Unified Medical Language System (UMLS) [6]. It integrates over 100 biomedical terminology resources and provides a mapping structure between them. It comprises over 3 million concepts, each of which is assigned to atleast one of the 134 semantic type.

A number of systems have been developed in the past for mapping text to UMLS concepts. One such popular and widely available tool is MetaMap [1] provided by the National Library of Medicine (NLM). MetaMap uses a knowledge-intensive approach based on symbolic, natural-language processing and computational-linguistic techniques.

We employ MetaMap and UMLS for extracting our Lexical features. Given a tweet, we first clean it by removing mentions, urls and redundant characters. We pass it to MetaMap as input which returns the subset of tokens in the tweet which are present as concepts in UMLS Metathesaurus along with their corresponding semantic types. These semantic types are utilized for finding the following features:

1. **Presence of Sign/Symptoms:** We check if a concept related to Signs or Symptoms of disease is present in the tweet. Presence of such a term forms an important feature for correctly classifying a tweet originally belonging to Symptom class. The semantic types which indicate the presence of such term are *Sign or Symptom; Physiologic Function; Laboratory or Test result* and *Injury or Poisoning*.

2. **Presence of Procedures:** We check if a concept related to Procedures employed in medical domain is present in the tweet. Presence of such a term forms an important feature for correctly classifying a tweet originally belonging to Prevention class. The semantic types which indicate the presence of such term are *Therapeutic or Preventive Procedure; Laboratory Procedure* and *Diagnostic Procedure*.

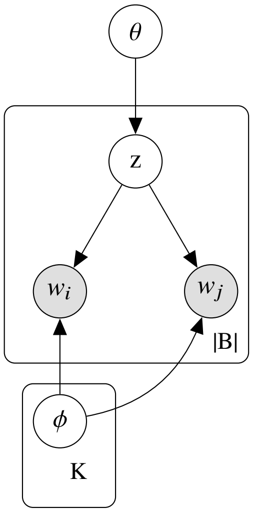
3. **Presence of Anatomy:** We check if a concept related to anatomy is present in the tweet. The semantic types which indicate the presence of such term are Body System; Body Substance; Body Space or Junction; Body Part, Organ, or Organ Component; Body Location or Region; Anatomical Structure.

4. **Presence of Drugs:** We check if a concept related to drugs is present in the tweet. The semantic types which indicate the presence of such term are Pharmacologic Substance and Clinical Drug. Procedures, Anatomy and Drugs can be found in tweets of Prevention class. However, such terms are also common in treatment related tweets. But, Anatomy and Drugs are used lot more in treatment related tweets. Utilizing, these 3 features along with latent features forms an important distinguishing factor of these 2 classes.

5. **Presence of Transmission Terms:** We check if a word in a tweet is present in our Transmission Term Dictionary, TraD.

6. **Presence of Report-related Terms:** We check if a word in a tweet is present in our Report Term Dictionary, ReD.

Figure 3.1: Bi-Term LDA Graphical Representation



Latent Features

As stated earlier, MetaMap has been found to not work well on Social Media Data and thus, only using features extracted using MetaMap is not likely to give us accurate classification. Thus, for making our classification model better, we extend the feature set as follows. Each tweet can be considered as a distribution over a certain topics and this distribution can be helpful in determining the class to which a tweet belongs. Thus, we try to make use of hidden topics present in a tweet. For extracting these hidden topics, we make use of Latent Dirichlet Allocation (LDA) [5]. A distribution over a set of latent topics can be inferred for each tweet using LDA which can be then used as a feature set for classification along with the lexical features.

However, as noted by Hong et al [34], conventional latent dirichlet allocation implicitly capture document-word co-occurrence patterns for learning the latent topics of the document. Document in our case is a tweet which has a 140-character limitation and thus, has a small number of tokens. Thus, directly applying the conventional LDA (or other conventional topic models) on tweets will suffer from the severe *data sparsity* problem (i.e. the sparse tweet-word co-occurrence patterns). More specifically,

1. The occurrences of words in tweets play less discriminative role compared to long documents where the model has enough word counts to know how words are distributed

and how they are related.

2. The context in case of tweets is also small as compared to long documents and this makes it more difficult for LDA to identify the topics of ambiguous words.

This data-sparsity problem makes convergence of statistical inference algorithms like Gibbs Sampling difficult and thus, the document-topic and topic-word distributions are not fixed for a given set of tweets. To overcome this, Yan et al [67] proposed a modified approach towards topic inference which they call *Bi-Term LDA*.

The graphical representation of Bi-Term LDA is as shown in Figure 3.1. A *bi-term* is an unordered word-pair co-occurred in a short context of the tweet. Each topic is represented as distribution over bi-terms, thus, containing groups of co-related words. As word-word co-occurrences are much higher that tweet-word co-occurrences, it is able to resolve the data sparsity problem.

Guided Bi-Term LDA: In order to ensure that each topic is distinct and captures a distinct class as needed for classification, we guide the Bi-Term LDA as follows. We take the number of topics as 6, each depicting one of the 6 classes. For each topic, we provide a set of seed words. We use the 5 Dictionaries defined earlier(SymD, PreD, TreD, TraD and ReD) for seeding the first 5 topics. The last topic is left unseeded. We define two topic to bi-term distributions. ϕ^r is the regular latent distribution. On the other hand, ϕ^s is an observed distribution seeded using the 5 dictionaries and thus, is known aprior. We use a Bernoulli switch x_i to choose whether to sample from ϕ^r or ϕ^s. The generative model of *Guided Bi-Term LDA* is as described below:

1. For each topic k=1....T:

   (a) Choose regular topic $\phi_k^r \sim \mathrm{Dir}(\beta^r)$.

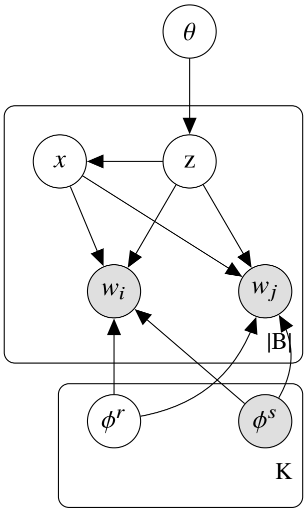
   (b) Choose seed topic $\phi_k^s \sim \mathrm{Dir}(\beta^s)$.

   (c) Choose $\pi_k \sim \mathrm{Bin}(1, 1)$.

2. Draw a topic distribution $\theta \sim \mathrm{Dir}(\alpha)$ for the whole collection

3. For each biterm b in the biterm set B

Figure 3.2: Guided Bi-Term LDA Graphical Representation



(a) Draw a topic assignment $z \sim \text{Mult}(\theta)$.

(b) Select an indicator $x_b \sim \text{Bern}(\pi_{z_b})$.

(c) if x_b is 0

   i. Draw two words: $w_i, w_j \sim \text{Mult}(\phi^r_k)$. // Choose from regular topic

(d) if x_b is 1

   i. Draw two words: $w_i, w_j \sim \text{Mult}(\phi^s_k)$. // Choose from seed topic

The same can be graphically represented as in Figure 3.2. The concept of Guided-LDA was first introduced by Jagarlamudi et al [37]. However, it's application in the case of Bi-Term LDA is novel. The features(distribution of tweets over 6 topics) extracted using Guided Bi-Term LDA are more effective in the classification task than a conventional LDA as we observe in results.

3.4.3 Evaluation

A Support Vector Machine (SVM) classifier with RBF kernel and gamma = 0.5 is trained to classify the tweets into the 6 classes using Lexical and Latent features as described above. We choose a simple Bag-of-Words (BOW) classification as baseline. Table 3.2 shows the accuracy of our model under 2 different scenarios: **(i) In-domain classification:** Here,

Table 3.2: Classification accuracies for 2 epidemics, using (i) Bag-Of-Words Model (BUR), (ii) Lexical Features only (LEX) and (iii) Lexical+Latent Features (LEX+LDA)

Train Set	**Test Set**					
	Ebola			**MERS**		
	BOW	LEX	LEX+LDA	BOW	LEX	LEX+LDA
Ebola	50.13 %	56.73 %	**75.02 %**	30.75 %	54.23 %	**66.91 %**
MERS	32.67 %	59.16 %	**71.20 %**	53.45 %	60.60 %	**76.53 %**

Table 3.3: Number of tweets in different classes as per the classification model

	Symptom	**Prevent- -ion**	**Treatm- -ent**	**Transmi- -ssion**	**Death Reports**	**Un- - informative**
Ebola	4390	10656	2305	2831	5751	54887
MERS	8293	9787	3494	13237	16926	126955
Total	12683	20443	5799	16068	22677	181842

the classifier is trained and tested with the tweets of same epidemic using 10-fold cross validation. The results are shown in the diagonal entries of Table 3.2. **(ii) Cross-domain classification:** Here, the classifier is trained with tweets of one epidemic, and tested on another epidemic. The results are shown in the non-diagonal entries of the table.

As can be observed from the table, our classification model outperforms BOW specially in cross-domain scenario. Also, using Latent features coming from LDA improves the accuracy significantly.

3.4.4 Classification Statistics

The number of tweets in each category after classification are shown in Table 3.3

Chapter 4

Extracting and Summarizing Targeted Information

After building the classification model, we run it on the tweets of our dataset. Table **??** shows the distribution of tweets in our dataset over 6 different categories.

We next extract and summarize information from the first 5 information-extensive categories to assist different stakeholders. The information can be extracted depending upon the target users during epidemic:

1. **Pro-disease Users:** Information coming from 'symptom', 'prevention', and 'transmission' classes helps assist pro-disease users and primary health care service. Pro-disease users are vulnerable to disease and precautionary steps are extremely helpful to restrict further spreading of the disease. For example, if people are aware of possible transmission mediums (human-to-human, animal-to-human etc) of the disease then they can avoid getting infected by taking relevant precautionary measures.

2. **Post-disease users:** Post-disease users are the ones who have already been diagnosed with the disease and are looking for treatment related information like hospital, drugs, medicines etc. These kind of information helps in secondary health care services where treatment of patients is going on.

3. **Monitoring community:** Government, health related organizations (WHO, CDC)

figuring out general situational updates like the number of infected people, number of deaths, etc. in various regions can benefit from summarization of death reports. Such a summary could make their planning of relief operations better.

4.1 Extracting Disease Signs and Symptoms

We take the tweets classified into Symptoms categories and extract symptoms of the disease from them. Given a tweet t, we check if it contains a word from our symptom dictionary, *SymD*. If a symptom s is found in t, then there can be 2 possibilities:

1. **Positive Symptom**: The user who posted tweet t might be conveying that symptom s *would be observed* if a person is infected with the disease.

   Eg. *Symptoms of mers include fever and shortness of breath*

2. **Negative Symptom**: The user who posted tweet t might be conveying that symptom *s would not be observed* even if a person is infected with the disease.

   Eg. *'#Ebola symptoms are different than upper respiratory tract pathogens, no cough, nasal congestion' Dr. Wilson*

We distinguish between the above 2 cases by using the terms having dependencies with the symptom term. We check if symptom s has a dependency with any strongly negative term in the tweet t. Symptom s is *negative symptom* (not a symptom) of the disease if s has dependency with atleast one strongly negative term in t. If there is no such dependency in t, then symptom s is *positive symptom* of the disease. For identifying strongly negative terms, we use Christopher Potts' sentiment tutorial [1].

The same symptom s might occur in multiple tweets. If s is classified as a symptom in one tweet and not a symptom in another tweet, then s is considered as *ambiguous* and is consequently, removed from both the classes.

[1] http://sentiment.christopherpotts.net/lingstruc.html

Table 4.1: Precision and Recall of Symptom Extraction Algorithm

Epidemic	Precision@k	Recall
Ebola	0.80	0.842
MERS	0.80	0.75

4.1.1 Ranking the Symptoms

After identifying the symptoms of a disease, we rank them on the basis of the number of tweets in which the symptom was mentioned. The ranking of symptoms helps us in knowing the symptoms which are being observed more for the disease.

4.1.2 Evaluation

We extract the actual symptoms which were observed for a disease as provided by organizations like Centers for Disease Control and Prevention[2] and some other online sources[3]. We compare the output of our algorithm with the actual symptoms to compute the precision and recall scores. Table 4.1 shows the precision@k and recall of the proposed symptom extraction algorithm for the 2 epidemics as considered in our dataset. Here, k is a hyperparameter which is chosen based on the count of actual symptoms the disease has. We take $k = 20$ for Ebola and $k = 5$ for MERS.

Please note that the sources containing actual symptoms are accessible to us as we are conducting the experiment after a significant amount of time from the epidemics. Such sources are not likely to be available at the time the disease outbreak.

4.2 Extracting Preventive Measures

We next take the tweets belonging to Prevention category and extract Preventive Measures from them. For this, we first extract candidate keyphrases from each tweet. We then rank based on their potential of being a genuine Preventive measure and finally output the top k ranked phrases.

[2] https://www.cdc.gov/vhf/ebola/symptoms/index.html

[3] http://chealth.canoe.com/condition/getcondition/mers

4.2.1 Keyphrase Extraction

We extract candidate keyphrases using verbs and their dependencies as present in the tweet's dependency tree [39]. A dependency tree indicates the relation among different words present in a tweet. All the verbs and the words dependent on them are taken in-order to form a candidate keyphrase.

4.2.2 Ranking the Keyphrases

Next, we rank these keyphrases in order to extract the relevant preventive measures out of the candidates. We provide each of the candidate keyphrase a *relevance score* (or a Information Score). The relevance score is computed as follows:

We make use of normalized *Phrasal Overlap* [48] between a candidate keyphrase and all the phrases in our Prevention Dictionary, PreD. 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:

$$phrasal\_overlap(c1, c2) = tanh\left(\frac{\sum_{i=1}^{n} m(i) * i^2}{|c1| + |c1|}\right)$$

where m(i) is the number of i-gram phrases which match in phrases $c1$ and $c2$.

Relevance Score of a candidate kephrase c is defined as

$$\max_{p \in PreD}(phrasal\_overlap(c, p)) * \log(freq(c))$$

We $freq(c)$ denotes the number of times c is chosen as a candidate keyphrase. This score denotes probability of a particular keyphrase being relevant to any preventive measure. Based on this score, we rank the keyphrases and pick top-k. The intuition is that the prevention measures for the new disease will have similar terms and phrases as the prevention measures of known diseases. We set k = 10 and n = 3 for our purposes. Similar to Symptoms, we filter out negative and ambigous cases using negated context.

Table 4.2: Precision and Recall of Prevention Extraction Algorithm

Epidemic	Precision@k	Recall
Ebola	0.50	0.714
MERS	0.40	0.571

4.2.3 Evaluation

We extract the actual preventive measures for the 2 epidemics as provided by Centers for Disease Control and Prevention [4] [5]. We compare the output of our algorithm with the actual measures. Table 4.2 shows the Precision@k and Recall of our system. The scores are relatively low. However, we found that the preventive measures which are not present in the actual list correspond to general methods which people propose which can be useful for the Pro-Disease uers.

4.3 Extracting Transmission Medium

During epidemics, vulnerable users looking for information about possible disease transmission mediums so that precautionary steps can be taken. Common users and health organizations post tweets regarding possible transmission possibilities of a disease for public awareness. It is observed that information about transmission mediums is mostly centered around keywords like ‘transmission’, ‘transmit’ etc. To identify informative components centered around such keywords, we explore the dependency relation among the words in a tweet using a dependency tree [22]. As an example, dependency tree for the tweet *’Ebola virus could be transmitted via infectious aerosol’* contains the following two dependency relations centered around keyword ‘transmitted’– (via, transmitted), (aerosol, transmitted). In this case, only aerosol is a proper transmission medium. However, as it is a Noun, it is distinguishable from the other candidate transmission medium.

For extraction of transmission medium, we detect all nouns connected to keywords in the dependency tree within a 2-hop distance. For separating the negated transmission mediums,

[4] https://www.cdc.gov/vhf/ebola/prevention/index.html

[5] https://www.cdc.gov/coronavirus/mers/about/prevention.html

Table 4.3: Evaluation of Transmission Medium Extraction

Epidemic	K	Precision@K	Recall
Ebola	10	0.30	0.30
	20	0.25	0.50
MERS	10	0.30	0.375
	20	0.25	0.625

we check for negated context similar to that in Section 4.1.

Finally, we rank the transmission mediums based on their term frequency i.e. number of tweets in which they occur and remove ambiguous mediums (present in both positive and negative list)

4.3.1 Evaluation

We extract the actual transmission mediums for both the diseases from credible online sources such as WHO [6] and CDC [7]. We compare the output of our algorithm with the actual transmission mediums to compute the precision and recall. Table 4.3 shows the precision@k and recall of the proposed transmission extraction algorithm for the 2 epidemics as considered in our dataset and for 2 different values of k. As can be observed from the table, recall value increases significantly on increasing the value of k. However, the decrease in precision is small.

4.4 Summarizing Death Reports

People suffering from the epidemic need support from government and health agencies. Government and health organizations generally keep track of number of people died or under treatment for better planning their relief operation. To help them, we summarize tweets providing reports and updates related to the epidemic. Primarily, we observe that such information is centered around keywords like ‘died’, ‘killed’, ‘dead’, ’death’ etc.

[6] http://www.who.int/mediacentre/factsheets/fs103/en/

[7] https://www.cdc.gov/coronavirus/mers/

For summarization of death reports, we use an Integer Linear Programming (ILP) based technique as proposed by Rudra et al [54] in the context of situational tweets posted during disaster. The summarization is achieved by optimizing the following ILP objective function:

$$max(\sum_{i=1}^{n} x_i + \sum_{j=1}^{m} Score(j).y_j)$$

subject to constraints

$$\sum_{i=1}^{n} x_i.Length(i) \leq L$$

$$\sum_{i \in T_j} x_i \geq y_j, j = [1...m]$$

$$\sum_{j \in D_i} y_j \geq |D_i| * x_i, i = [1...n]$$

where L is the desired summary length (in number of words), n is the number of tweets considered for summarization, m is the number of distinct content words included in the n tweets, i is the index for tweets, j index for report-related terms, x_i is a boolean indicator variable for tweet i (1 if tweet i should be included in summary, 0 otherwise), y_j is a boolean indicator variable for death-related term j, $Length(i)$ is the number of words present in tweet i, $Score(j)$ is the tf-idf score of report-related term j, T_j is the set of tweets where content word j is present and D_i is the set of death-related terms present in tweet i.

The objective function considers both the number of tweets included in the summary (through the x_i variables) as well as the number of important report-related terms (through the y_j variables) included. The first constraint ensures that the total number of words contained in the tweets that get included in the summary is at most the desired length L (user-specified) while the second constraint ensures that if the death-related term j is selected to be included in the summary, i.e., if $y_j = 1$, then at least one tweet in which this report-related term is present is selected. Similarly, the last constraint in ensures that if a particular tweet is selected to be included in the summary, then the deaath-realted terms in

Table 4.4: ROUGE Metric for Summaries of Death-Reports

Epidemic	F-Score	Recall
Ebola	0.7037	0.6930
MERS	0.7462	0.6689

that tweet are also selected.

Rudra et al [54] considered all nouns and verbs as content words. However, in our case, all such keywords present in a tweet do not contribute to death-related updates. Based on our observation, we only take numerals(e.g., number of casualties) and terms in our Report Dictionary (ReD) (major locations and death-related verbs).

4.4.1 Evaluation

We use GUROBI Optimizer [8] to solve the ILP. After solving this ILP, the set of tweets i such that $x_i = 1$, represent the summary. The summary generated is compared with a manually generated summary of the death-reports. We use the standard ROUGE [42] metric (ROUGE-1 F-score and recall) for evaluating the quality of the summaries generated. Table 4.4 gives the ROUGE-1 F-scores and recall values for both the 2 epidemics.

4.5 Summarizing Treatment Information

Treatment Mechanisms and information related to it can be very useful for Post-Disease Users who have been diagnosed with the disease. However, as reported by Centers for Disease Control and Prevention, both the epidemics chosen by us had no vaccine or medicine for treatment and the treatment procedures are also very low in number. Similar phenomenon is observed in tweets who rarely mention any medication or treatment mechanisms. Thus, directly extracting the most relevant treatment keyphrases, as in Preventive measures case, won't work.

Instead, tweets belonging to this category contain names of hospitals where people are being treated and treatment-related updates. Thus, instead of extracting phrases, we

[8] http://www.gurobi.com

Table 4.5: ROUGE Metric for Summaries of Treatment Related Tweets

Epidemic	**F-Score**	**Recall**
Ebola	0.6210	0.5363
MERS	0.7802	0.6666

summarize the tweets, using a similar methodology as used for death-reports.. We follow same ILP framework proposed in previous section but instead of optimizing death-report related terms, we optimize the coverage of treatment and their update related terms. We take care of the fact that some future outbreak might have proper medication by using the out Treatment Phrase Dictionary, TreD, along with some update related terms like numerals as the content words.

4.5.1 Evaluation

We prepare a ground truth summary of 200 words for each of the diseases and use the standard ROUGE metric (ROUGE-1 recall and F-score) for evaluating the quality of the summaries generated. Table 4.5 gives the ROUGE-1 F-scores and recall values.

Chapter 5

Identifying Communal Microblogs during Diaster

In this section, we identify communal tweets posted during disaster events which are potentially harmful and dangerous.

5.1 Dataset

We consider tweets posted during the following five disaster events – (i) **NEQuake** - a devastating earthquake in Nepal, (ii) **KFlood** - floods in the state of Kashmir in India, (iii) **GShoot** - a terrorist attack in Gurudaspur, India, (iv) **PAttack** - coordinated terrorist attacks in Paris, and (v) **CShoot** - terrorist attack at the Inland Regional Center in San Bernardino, California. Note that the first two events are natural disasters, while the last three events are man-made disasters. Additionally, events occurring in different geographical regions have been selected so that the study would not be biased to any particular demographics. Relevant tweets posted during each event were crawled through the Twitter API [1] using keyword matching. Only English tweets were considered based on the language identified by Twitter. From these tweets, situational tweets are removed using the classifier of [54]. Table 5.1 states the number of tweets collected for the events, and the number of distinct

[1] "REST API Resources, Twitter Developers," 2015, https://dev.twitter.com/docs/api

Table 5.1: Statistics of data collected

Disaster Event	No. of Tweets	No. of 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

users who posted them.

5.2 Identifying Communal Tweets

We identify communal tweets from other non-situational tweets by developing a supervised classifier. We aim for an event-independent classifier and use the following features for classification:

1. **Presence of communal slang phrases:** In order to develop the classifier, a lexicon of religious terms and antagonistic hate terms about religion and related nationality is needed. For this, the terms in a standard lexicon of religious terms [2] are considered. The terms in the lexicon were marked as hate-terms or normal religious term. Further, we collected all the hate terms related to religion and nationality from a repository of terms frequently used in hate speeches [3].

2. **Presence of religious/racial negated or hate terms:** The presence of any strongly negative term or slang term in the vicinity of neutral religious terms like ‘Muslim’ or ‘Christian’ is detected. A subjectivity lexicon developed in [62] is used to identify strongly negative terms, and we obtain a standard list of slang terms from online sources [4]. Then, it is checked whether such terms appear within a left and right word window each of size three with respect to a religious term. Thus, presence of phrases like ‘bastard missionaries’, ‘islamic scoundrels’, ‘jesus fucktards’ are identified.

[2] http://www.translationdirectory.com/glossaries/

[3] www.hatebase.org

[4] www.noswearing.com

Table 5.2: Classification accuracies for communal tweets, using (i) Burnap Model (BUR), (ii) Proposed features (PRO).

Train Set	**Test Set**					
	NEQuake		**KFlood**		**GShoot**	
	BUR	PRO	BUR	PRO	BUR	PRO
NEQuake	79.98 %	**87.73 %**	54.17 %	**73.96 %**	60.17 %	**73.85 %**
KFlood	60.40 %	**86.54 %**	**76.55 %**	76.02 %	61.81 %	**69.81 %**
GShoot	59.64 %	**80.96 %**	53.12 %	**70.31 %**	**80.09 %**	76.51 %

3. **Presence of communal hashtags:** Some specific hashtags are explicitly used across various events to curse certain religious communities, such as, '#SoulVultures', '#evangelicalvultures', '#WeAreThanklessMuslims', '#TweetlikeSecularJamat'. Presence of such hashtags in tweets is identified.

A Support Vector Machine (SVM) classifier with RBF kernel and gamma = 0.5 is trained to classify the tweets into two classes based on the features described above. Table 7.1 shows the comparison of the proposed features (PRO) with the features in the hate-speech model developed by Burnap et al. [9] (BUR), using the same classifier, under two different scenarios. **(i) In-domain classification:** Here the classifier is trained and tested with the tweets of same event using 10-fold cross validation. **(ii) Cross-domain classification**: Here the classifier is trained with tweets of one event, and tested on another event. The accuracies are almost same for in-domain classification but the proposed classification model performs much better than the BUR model in cross-domain classification, since it is independent of the vocabulary of specific events.

Chapter 6

Characterizing the Communal Microblogs

6.1 Popularity of communal tweets

Communal Tweets would be a real threat to the people if they become hugely popular among the masses. With this motivation, we investigate whether communal tweets receive large exposure among the population. Also, we need to know how quickly should we act against communal tweets. Thus, it is important to study the rate at which these tweets gain popularity.

A standard metric for estimating the popularity of a tweet is it's *Retweet Count* [33][69][13]. We express the popularity of a tweet in terms of no. the number of retweets the tweet has. However, the tweets in our dataset were crawled within a short span of time from the corresponding disaster events. Thus, most of the tweets in our dataset were posted only few minutes before they were crawled. As a result, We missed out the retweets to these tweets which occurred after crawling. So, there is a need to predict the ultimate number of retweets a tweet has received, instead of directly using the retweets from our dataset. In addition to this, there is a need to predict the number of retweets at different time instances from the time of posting, in order to infer the rate at which communal tweets gain popularity.

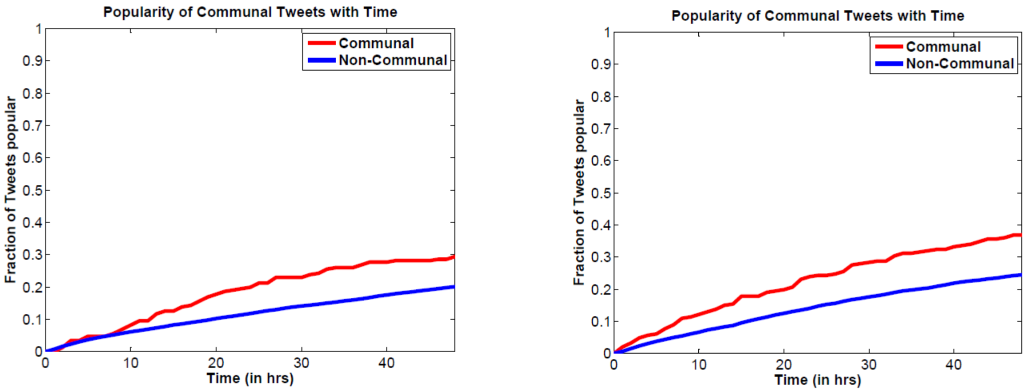
For making these 2 predictions, we use the method described in SEISMIC [70]. In brief, SEISMIC builds on the theory of self-exciting point processes to develop a statistical model using human reaction times (i.e how long it takes for a person to retweet a post) and post infectiousness (i.e. how likely the post w is to be retweeted at time t). Using reaction times and post infectiousness, SEISMIC makes reasonably accurate predictions of the retweet count of a tweet after a time t (which is an input to the model) from the time of posting. Using this model, we find $RetweetCount(tweet, t)$, which denotes the predicted retweet count of a $tweet$ after time t relative to the time it was posted on twitter. For making predictions, we define a function $IsPopular(tweet, t)$ as follows:

$$IsPopular(tweet, t) = \begin{cases} 1 & \text{if } RetweetCount(tweet, t) >= pop\_threshold \\ 0 & \text{otherwise} \end{cases}$$

A tweet is said to become popular after time t if $IsPopular(tweet, t) = 1$. For this experiment, we take the value of $pop\_threshold = 500$. Figure 6.1 shows the variation of Fraction of popular of tweets, i.e., $\frac{\sum_{tweet} IsPopular(tweet,t)}{totalno.oftweets}$, between $t = 0$ and $t = 48hrs$, for communal and non-communal tweets of two events NEQuake and GShoot. Plots from other disaster events also follow a similar trend. Based on these plots, we draw the following 2 conclusions:

1. For the initial 2 hours from the time of posting the rate of increase in Fraction of Communal Tweets for communal tweets is higher than that of non-communal tweets. Thus, *Communal Tweets gain popularity at a significantly faster rate than non-communal tweets*.

2. After a particular instant of time, Fraction of Communal Tweets becomes approximately constant for both communal as well as non-communal tweets. This constant value denotes the *ultimate fraction of tweets* in the communal and non-communal category which became popular. We observe that this value of significantly higher in the case of communal tweets than non-communal tweets for all the disaster events. Thus, in general, *communal tweets are retweeted more than non-communal tweets*.



(a) NEQuake

(b) GShoot

Figure 6.1: Fraction of Tweets popular at different relative time instances

Note that for this analysis, we only considered original tweets, i.e., tweets which are not retweets themselves. Out of original tweets, we filter out the tweets which have 0 retweets as per our dataset.

6.2 Characterization of Communal Users

We next characterize **Communal users**, i.e., users who are involved in communal tweets during Disaster events. For this characterization, we first divide the users into 3 classes, namely, originators, propagators and mentions. These 3 classes are defined as follows:

(i) Initiators: Users who initiate a tweet.

(ii) Propagators: Users who retweet the communal tweets posted by initiators/other propagators or copy the content of initiator/other propagators and post their own tweet with minor changes.

(iii) Mentioning & Mentioned Users: Mentioning Users are those who mention other users in their tweets and Mentioned Users are those who get mentioned in the tweets of mentioning users.

6.2.1 Construction of Originator and Propagator Set

For dividing users into originators 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 would be classified as originator. 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 [7]. 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 [4]. 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 achieve this classification as follows. We first compute normalized *Phrasal Overlap Measure* [48] between all pair of tweets in 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:

$$phrasal\_overlap\_norm(t1, t2) = tanh\left(\frac{\sum_{i=1}^{n} m(i) * i^2}{|t1| + |t2|}\right)$$

where m(i) is the number of i-gram phrases which match in tweets t1 and t2. We then cluster together the tweets t1 and t2 having $phrasal\_overlap\_norm(t1, t2) >= similarity\_threshold$ 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 smallest timestamp). Phrasal overlap between 2 clusters is defined as the overlap between the representatives of those clusters.

The algorithm is described in Algorithm **??**. It returns a (tweet, retweet set) pairs. The users corresponding to tweet become originators and those corresponding to retweet become propagators.Please note that we remove "RT @user" from the tweet, if present, before finding the overlap similarity.

For our purposes, we take the value of n as 3 and similarity_threshold as 0.8.

Construction of Mentioning and Mentioned Set: Construction of mentioning and mentioned sets, however, is less challenging due to lesser possible variations. Presence of @v without RT in a tweet x of user u signifies that user v was mentioned in the tweet of user u. Thus, user u would be classified into Mentioning Users and user v would be classified into Mentioned users.

6.2.2 Common and Popular Users

We use the follower count of users to classify a user as **common** or **popular**. Specifically, a user is called popular if his/her $followercount > popu\_threshold$ and a user is called common if his/her $followercount <= comu\_threshold$. For our experiments, we use the value of comu_threshold as 5,000 and popu_threhold as 10,000.

6.2.3 Questions of Interest

Based on the above mentioned classification of users, we aim to answer the following questions:

Originators and Propagators:

1. Which type of users among popular and common are more involved in communal tweets?
2. Which among communal originators and communal originators are more popular?
3. How many users are common among communal originators and propagators?
4. What are the topical interests of these users?
5. Are the communal users getting outraged suddenly due to disaster?

Mentioning and Mentioned Users:

1. Do common masses try to provoke popular users to make their communal tweets popular?

Algorithm 1 Hierarchical Clustering Algorithm for finding tweet-retweet pairs

```
function FindRetweets($Tweets, Timestamps, sim\_threshold$)
    $T = [][]$                                              ▷ Distance Matrix
    $N = Tweets.length()$
    for $i = 1$ to $N$ do
        for $j = 1$ to $N$ do
            $T[i][j].sim = phrasal\_overlap\_norm(Tweet[i], Tweet[j])$
            $T[i][j].index = j$
        end for
        $I[i] = 1$                                          ▷ Keeps track of active clusters
        $P[i] = priority\_queue(T[i])$                      ▷ Sorted on sim
    end for
    $A = []$                                                ▷ List of Tweet-Retweet Pairs
    for $k = 1$ to $N - 1$ do
        $k_1$ =$\arg\max_{k:I[k]=1}$ P[k].MAX().sim
        $k_2 = P[k_1].MAX().index$
        $curr\_sim = P[k_1].MAX().index$
        if $curr\_sim < sim\_threshold$ then
            $break$                                         ▷ Cut the Dendogram on Similarity Threshold
        end if
        if $Timestamps[k_1] > Timestamps[k_2]$ then
            $swap(k_1, k_2)$                                ▷ Earliest tweet as cluster representative
        end if
        $A.append(< k_1, k_2 >)$
        $I[k_2] = 0$
        $P[k_1] = []$
        for $i$ $with$ $I[i] = 1$ $and$ $i \neq k_1$ do     ▷ Update P and T
            $P[i].DELETE(T[i][k_1])$
            $P[i].DELETE(T[i][k_2])$
            $T[i][k_1].sim = phrasal\_overlap\_norm(Tweet[i], Tweet[k_1])$
            $P[i].INSERT(T[i][k_1])$
            $T[k_1][i].sim = phrasal\_overlap\_norm(Tweet[i], Tweet[k_1])$
            $P[k_1].INSERT(T[k_1][i])$
        end for
    end for
    return A
end function
```

6.2.4 Popularity of Originators and Propagators

We check the distributions of originators and propagators of communal and non-communal tweets. Fig 6.2 shows the popularity distribution of communal and non-communal originators. From the cdf, it is evident that communal originators are in general more popular than non-communal originators. Also, out of the total common users who originated a tweet **6.50%** were communal users. On the other, out of the total popular users who had posted a tweet **9.97%** were found to be communal. Thus, more percentage of popular originators are communal than common ones.

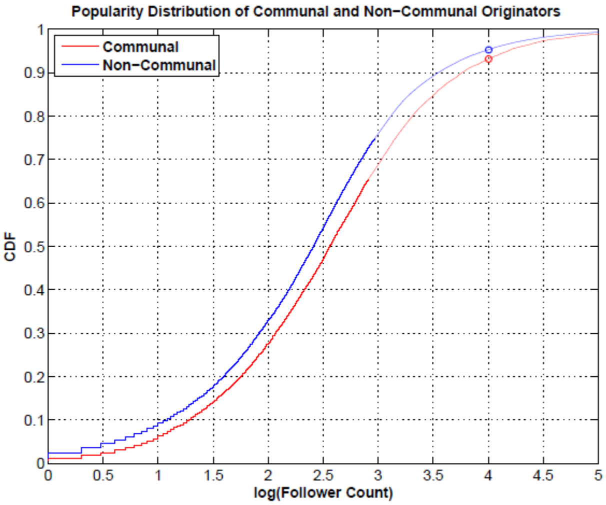


Figure 6.2: Popularity Distribution of Originators

Next, we check the distribution of propagators of communal and non-communal users. Fig 6.3 shows the probability distribution of communal and non-communal propagators. Interestingly, it can be seen from the cdf that there are more fraction of communal propagators having low follower count than non-communal propagators. However, approximately same fraction of communal propagators have high follower count as the non-communal propagators. Also, out of the total common users who propagated a tweet **8.81%** were communal users. On the other, out of the total popular propagators **9.30%** were communal users.

We also compare the popularity distributions of communal originators and propagators. Fig 6.4 shows the cdf of follower count for communal originators and communal propagators.

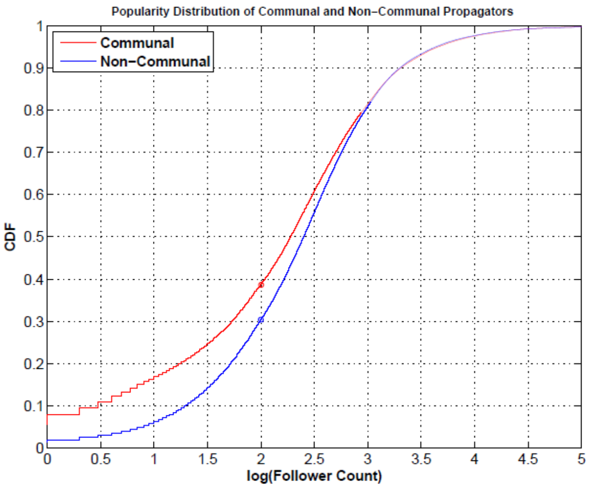


Figure 6.3: Popularity Distribution of Propagators

We see from the plot that propagators are less popular than originators of communal tweets during disaster.

From the above observations, it is evident that popular users are equally involved in posting & propagating of communal tweets during disasters as are common users. As popular users have large no. of followers, they may be influential and might be a source through which these tweets will spread disrupting communal harmony during disaster events.

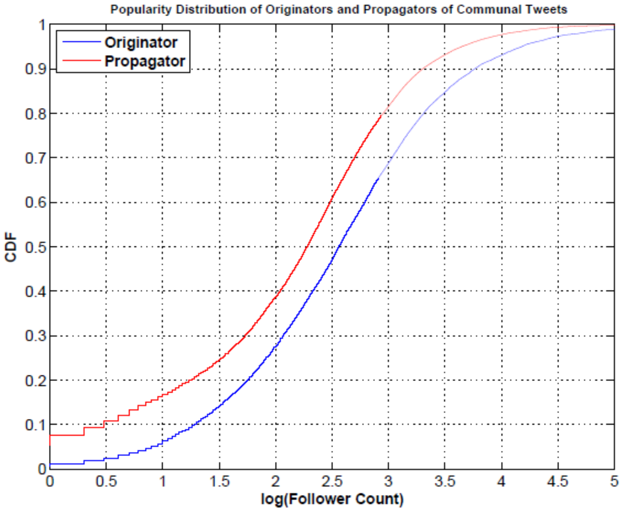


Figure 6.4: Communal Originators vs Propagators

6.2.5 Overlap between Originators and Propagators

We next find out if originators of communal tweets during different disaster events also work as propagators during the same event or vice-versa. For answering this, we compute the *Overlap Similarity* (or *Szymkiewicz-Simpson Similarity*) score between the set of originators and the set of propagators during each event, and averaged the score obtained from the five different events. *Overlap Similarity* between 2 sets X and Y is given by

$$overlap(X,Y) = \frac{|X \cap Y|}{min(|X|,|Y|)}$$

We used Overlap Similarity as it is a good measure of determining if one of the set is completely consumed by other which is the question we are trying to address. We obtain a similarity score of **0.1955** which is low . Thus, most of the originators of communal tweets are mainly interested in posting their own opinion and thoughts, rather than retweeting contents posted by others.

6.2.6 What are the topics of interest of communal users?

We find the topics which are of interest to communal users and the topics related to them. For this analysis, We consider the following 7 topics: (i) Media Houses & Journalists (News), (ii) Politics, (iii) Movies & Entertainment, (iv) Religion, (v) Sports, (vi) Writers/Authors, and (vii) Business.

First, we collected certain keywords which characterize these topics through various online sources [1]. We manually checked these keywords and filtered out the incorrect ones. We found the topical interests of common and popular users in different ways. For Common Users, We used the Bio specified by them on twitter. Their bio was first tokenized and the tokens were normalized. These tokens were then searched in each of the keyword list one by one once without lemmatization and then after lemmatization. If a token matched any of the keywords in 7 topics, the user was marked as interested in that particular topic. If the tokens of a user were not present in any of the keyword lists, the user was marked as

[1] http://bit.ly/2fGlW4c; http://bit.ly/2foknXA; http://bbc.in/2f4WhQV; http://bit.ly/2ebyYpf; http://bit.ly/2fsr6Mx; http://bit.ly/2fbZ6xn; http://bit.ly/2fsryKo; http://bit.ly/2fqsSiM; http://bit.ly/2ecpbPZ; http://bit.ly/2ecrUss; http://bit.ly/2fd7yMR; http://bit.ly/2f68Njk

"Others" and was not considered for our analysis.

For Popular users, instead of searching their Bio, we use *Who is Who* Service [2]. This service can accurately and comprehensively infer attributes(topic/ classes) of interest of a vast majority of Twitter's popular users (with ranking metrics as number of followers) [57]. On querying the screen name of a user, it returns a weighted list of attributes(tags) which characterize a person. The communal users in our dataset were queried on it's webpage and the weighted list of tags were scrapped using *Selenium*[3]. Out of these tags, top 3 were chosen (on the basis of the weights returned) to determine the set of major topic of interest of a user. These tags were then searched in each of the keyword lists of 7 topics defined earlier (both without lemmatization and with lemmatization). If different tags were found in multiple lists, the user was marked in all those lists and was given a normalized weight corresponding to the weight returned by the API for those tags. Similar to common users, if none of the top 3 tags were found in any of the lists, the user was marked as "Others" and was not considered for our analysis.

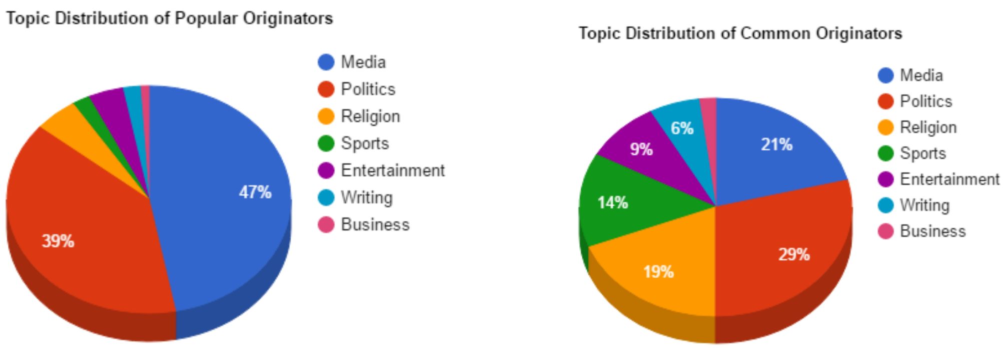
Fig 6.5 shows the distribution of topical interests of popular and common originators. A similar phenomena is also observed for propagators. It is clear that most of popular originators are related to media houses and politics. However, a significant fraction of common users are also interested in religion and sports, along with news, media and politics. It would be interesting to study the users corresponding to these major topics separately in future.

6.2.7 Are the users getting outraged suddenly?

Previous studies argue that a significant rise is observed in communal hate online following 'trigger' events like disaster [9] [2] [66]. According to them, these trigger events work as activators to wake up the old feelings of hatred and negative sentiments towards 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 the distribution of communal users who have a general tendency to post communal tweets

[2] http://twitter-app.mpi-sws.org/who-is-who/

[3] http://www.seleniumhq.org/projects/webdriver/



(a) Popular Users

(b) Common Users

Figure 6.5: Distribution of Topical Interests of Popular and Common Originators

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 dataset 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 scrap all the tweets posted by him/her on $\forall d' \in TimeWindow(u, d)$. In order to do this, we use Twitter's Advanced Search[4] utility which can retrieve tweets posted by a user, given his/her screen name and a particular $TimeWindow(u, d)$. We ran our communal tweet detection algorithm on the scrapped tweets and marked the retrieved tweets as communal and non-communal. Based on the classification, we define a vector v_u for each user u as following:

$$v_u[0] = 1$$

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

where $i \in [-15, 15]$. Next, for each user u, we find his/her **regularity score**, r_u =

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

$\sum_{i=-15}^{15} v_u[i]$. r_u defines the number of days user u posted a communal tweet in his/her $TimeWindow(u,d)$. Fig 6.6 shows the cdf of regularity score for GShoot and NEQuake. We draw following 2 observations from the plot:

1. Most of the users (80-90%) have regularity score either 1 or 2

2. There are a small fraction of users (5-10%) having large values of regularity score (>5).

Thus, a large fraction of users only get outraged at the time of disaster and do not express their hatred towards people of a particular religion or race otherwise. However, there are a few **Regular Communal Users** who repeatedly post communal tweets irrespective of any trigger event. There is a need to mark out such users and take strict actions against them.

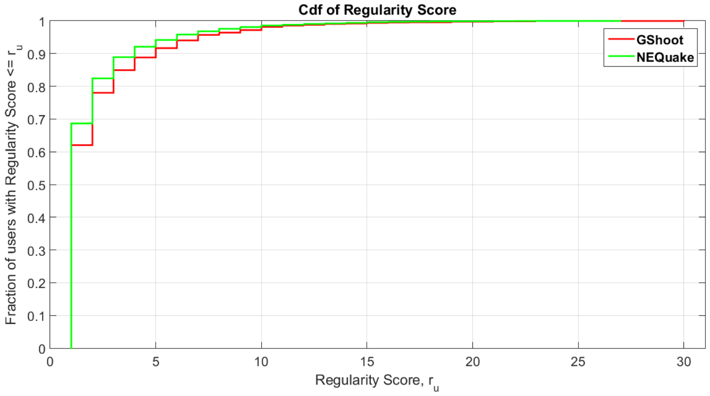


Figure 6.6: Cdf of Regularity Score of Communal Users

6.2.8 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 [53]. Thus, mentioning these influentials in the network helps in achieving a large-scale chain-reaction of influence driven by word-of-mouth [38][11]. 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 [49][63]. Thus, common

Table 6.1: Mentioning Stats for communal and non-communal tweets. C deontes common user while P denotes popular user. u1->u2 means user u1 mentioned user u2 in his/her tweet.

Type of Disasters	**Communal Tweets**		**Non-Communal Tweets**	
	C->P	P->P	C->P	P->P
All Disasters Combined	**70.67 %**	2.99 %	68.23 %	3.39 %
Man-Made Disasters	**74.04 %**	2.81 %	66.01 %	4.41 %
Natural Disasters	51.18 %	4.04 %	**69.15 %**	3.74 %

users, i.e. users with small number of followers on twitter, often mention popular users in their tweets, so as to increase the reachability and effectiveness of tweets.

In this section, we find out if mentioning popular users in order to achieve higher influence becomes more prevalent in the case of communal tweets than non-communal tweets. Table 6.1 shows the percentage of times a common/popular user mentioned a popular user out of the total mentioning instances in communal and non-communal tweets respectively.

We find that for our overall corpus containing tweets from 5 disaster events, the percentage of common/popular user mentioning a popular user is approximately same in both, communal as well as non-communal tweets. We then study this result separately for natural and man-made disaster. In the case of man-made disasters, we find that percentage of cases in which a common user mentioned a popular user is significantly large in the case of communal tweets (**74.04 %**) as compared to non-communal tweets (**66.01 %**). However, to our surprise, in the case of natural disasters, the value of this percentage is much lower in the case of communal tweets (**51.18 %**) as compared to the percentage for non-communal tweets (**69.15 %**). Next, we verify this result by taking more stronger assumptions.

Verification of Mentioning Results: As discussed, we found that the percentage of cases in which a common user mentions a popular user is larger for communal tweets than

Table 6.2: Mentioning Stats for communal and non-communal tweets using Klout Score

Type of Disasters	**Communal Tweets**		**Non-Communal Tweets**	
	C->P	P->P	C->P	P->P
All Disasters Combined	71.82 %	5.05 %	71.14 %	3.97 %
Man-Made Disasters	75.93 %	4.84 %	71.16 %	5.16 %
Natural Disasters	51.49 %	6.05 %	71.13 %	3.41 %

non-communal tweets in the case of man-made disasters and smaller for communal tweets than non-communal tweets in the case of natural disasters. While computing these results, we had used the number of followers of a user as his/her measure of influence and popularity. However, [11] showed that users who have large number of followers are not necessarily influential in terms of spawning retweets or mentions. Thus, directly using the number of users as popularity measure could have resulted in biased results. In this section, we verify the difference in the cases of natural and man-made disasters by using a stronger measure for user influence and classify him/her as popular or common based on this measure.

Klout Score [50] is a very popular and commonly used online rating service for twitter users. It is an influence scoring system which measures the user's overall online influence with a score ranging from 1 to 100, with 100 being the highest amount of possible influence. It incorporates information for the user from multiple networks and communities and aggregates over 3600 features that capture signals of influential interactions across multiple dimensions for each user. Thus, it provides a much better measure for a users' popularity. We classify a user as popular user if he/she has a Klout Score >60 and classify a user as common user if he/she has a Klout Score <= 60. We re-compute the mentioning stats with using this categorization of users. Table 6.2 shows the percentage of times a common/popular user mentioned a popular user out of the total mentioning instances in communal and non-communal tweets respectively as per the new definition of popularity. We find that the relative values for communal and non-communal tweets in the case of natural and man-made disaster remain approximately the same as before. Thus, we conclude that, in the case of natural disasters, common communal users have lower tendency to increase the visibility of their tweets as compared to common non-communal users. However, common communal users have a higher tendency to mention popular users in their and thereby, increase its visibility, in the case of man-made disasters.

Chapter 7

Anti-Communal Tweets

A possible way to counter communal tweets is to identify them and report them to twitter. As their deletion might take time, it is important to promote tweets which spread peace and harmony.

We observe that, during a disaster, while many people post communal tweets, there are some users who post anti-communal content, asking people to stop spreading communal posts. Thus, a potential way of countering communal content is to utilize such anti-communal content. However, as opposed to communal tweets, *anti-communal tweets* receive much less exposure and popularity. Thus, in order to utilize them, we a need to detect these tweets from the set of tweets being posted during disaster and persuade users for making these kind of tweets popular. For detecting them, we filter out the communal tweets and divide the non-communal tweets into anti-communal and not-anti-communal categories.

7.1 Establishing Gold Standard

To understand the pattern of anti-communal tweets and define the rules for its detection, we require gold standard annotation for a set of tweets. For each of the five events, we randomly and repeatedly sample tweets (after removing duplicates) from non-communal tweets and tag them accordingly. By this process, a total of 216 tweets were identified as anti-communal across all 5 disasters.

7.2 Features for classification

As mentioned earlier, our main objective is to make our classifier independent of any specific disaster event. Following communal tweet classifier approach, we rely on using a set of lexical and content features for the classification. We choose the following features for our classification model:

1. **Presence of anti-communal hashtags:** We observe that some specific hashtags are explicitly used across various events to post anti-communal tweets and ask users not to post communal content. Some example Hashtags are:

   "#RespectAllReligion", "#MuslimsAreNotTerrorist", "#ThisisNotIslam", "#NothingToDoWithIslam", "#stopit"

2. **Presence of collocations:** Some collocations are frequently used in anti-communal tweets like ‘nature doesnt discriminate’, ‘has no religion’, ‘terrorism defies religion’ etc. The collocations and hashtags are disjoint for natural and man-made disasters but are common across multiple man-made and multiple natual disasters.

3. **Mentioning multiple religious terms:** The aim of anti-communal tweets is to ask people to treat all religions equally. Thus, either they do not mention religious terms explicity or they mention multiple religions so as to create a sense of unity. As an example: *’WTF people are trying to save their life & this MORONs Tweeting Hindu christian muslim #earthquake #NepalEarthquake’.*

7.3 Evaluation

A Support Vector Machine (SVM) classifier with RBF kernel and gamma = 0.5 is trained to classify the tweets into two classes based on the features described above. Table 7.1 shows the comparison of the proposed features (PRO) with the features in the hate-speech model developed by Burnap et al. [9] (BUR), using the same classifier, for in-domain and cross-domain classification. Once again, the proposed model performs significantly better than BOW in cross-domain scenario.

Table 7.1: Classification accuracies for Anti-communal tweets, using (i) Burnap Model (BUR), (ii) Proposed features (PRO).

Train Set	**Test Set**					
	NEQuake		**GShoot**		**PAttack**	
	BUR	PRO	BUR	PRO	BUR	PRO
NEQuake	62.33 %	**77.27 %**	50.0 %	**65.23 %**	51.13 %	**73.65 %**
GShoot	46.0 %	**74.03 %**	63.25 %	**70.83** %	52.26 %	**72.17 %**
PAttack	58.0 %	**75.65 %**	53.13 %	**62.49 %**	**77.06 %**	76.86 %

Chapter 8

Conclusion and Future Work

In this project, we explored tweets posted in 2 types of emergency scenarios- *Epidemic* and *Disaster*. We build a classifier for categorizing tweets posted during epidemics in 6 classes. We overcame the failures of MetaMap on Social Media Data by utilizing latent features extracted using a novel *Guided Bi-Term LDA*. After classification, we extracted information from the 5 informative classes which can be targeted towards Pro-Disease, Post-Disease and Monitoring Communities.

We then switched our attention towards Disaster and observed that they contain a special category of tweets called *Communal Tweets* which need attention. We applied an event-independent classifier to identify communal tweets. We then used these communal tweets to perform a detailed study of the nature of communal tweets and characteristics of users who post and propagate them. We found that communal tweets are popular among the masses and gain popularity at a rapid rate. We also found that communal tweets are posted by many popular users who have politics and media as their main topic of interest. Additionally, some common users try to provoke popular users specially in the case of man-made disasters so as to expand exposure of their communal tweets. Also, we found that there are a small fraction of users who have a general tendency of posting communal content irrespective of any 'trigger' event. Next, we discovered that some users do aim to counter them by posting *anti-communal tweets*. We build a classifier for their identification which can be a useful tool for countering communal tweets.

In future, we would like to improve the different classifiers developed in this project by incorporating more features. We would like to make the ranking of symptoms, preventions and transmission better by including user's domain knowledge. We would like to explore uninformative tweets posted during epidemic and would like to characterize their users. Post identification of anti-communal tweets, we would like to facilitate their exposure and look for techniques for persuading users to make such tweets popular.

Bibliography

[1] Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

[2] Imran Awan and Irene Zempi. 'i will blow your face off'—virtual and physical world anti-muslim hate crime. *British Journal of Criminology*, page azv122, 2015.

[3] Imran Awan and Irene Zempi. The affinity between online and offline anti-muslim hate crime: Dynamics and impacts. *Aggression and violent behavior*, 27:1–8, 2016.

[4] Norhidayah Azman, David Millard, and Mark Weal. Patterns of implicit and non-follower retweet propagation: Investigating the role of applications and hashtags. 2011.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[6] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

[7] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10. IEEE, 2010.

[8] Pete Burnap, Omer F Rana, Nick Avis, Matthew Williams, William Housley, Adam Edwards, Jeffrey Morgan, and Luke Sloan. Detecting tension in online communities with computational twitter analysis. *Technological Forecasting and Social Change*, 95:96–108, 2015.

[9] Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242, 2015.

[10] Pete Burnap, Matthew L Williams, Luke Sloan, Omer Rana, William Housley, Adam Edwards, Vincent Knight, Rob Procter, and Alex Voss. Tweeting the terror: modelling the social media reaction to the woolwich terrorist attack. *Social Network Analysis and Mining*, 4(1):1–14, 2014.

[11] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 10(10-17):30, 2010.

[12] Irfan Chaudhry. # hashtagging hate: Using twitter to track racism online. *First Monday*, 20(2), 2015.

[13] Kailong Chen, Tianqi Chen, Guoqing Zheng, Ou Jin, Enpeng Yao, and Yong Yu. Collaborative personalized tweet recommendation. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 661–670. ACM, 2012.

[14] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 71–80. IEEE, 2012.

[15] Cynthia Chew and Gunther Eysenbach. Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PloS one*, 5(11):e14118, 2010.

[16] Xiangfeng Dai and Marwan Bikdash. Hybrid classification for tweets related to infection with influenza. In *SoutheastCon 2015*, pages 1–5. IEEE, 2015.

[17] Ed De Quincey and Patty Kostkova. Early warning and outbreak detection using social networking websites: The potential of twitter. In *International Conference on Electronic Healthcare*, pages 21–24. Springer, 2009.

[18] Kerstin Denecke. Extracting medical concepts from medical social media with clinical nlp tools: a qualitative study. In *Proceedings of the Fourth Workshop on Building and Evaluation Resources for Health and Biomedical Text Processing*, 2014.

[19] Kerstin Denecke. Information extraction from medical social media. In *Health Web Science*, pages 61–73. Springer, 2015.

[20] Kerstin Denecke and Wolfgang Nejdl. How valuable is medical social media data? content analysis of the medical web. *Information Sciences*, 179(12):1870–1880, 2009.

[21] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18, 2012.

[22] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, pages 29–30. ACM, 2015.

[23] Noah Elkin. How america searches: Health and wellness. *Opinion Research Corporation: iCrossing*, pages 1–17, 2008.

[24] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A high-coverage lexical resource for opinion mining. *Evaluation*, pages 1–26, 2007.

[25] Susannah Fox. *The social life of health information, 2011.* Pew Internet & American Life Project Washington, DC, 2011.

[26] Carol Friedman, George Hripcsak, Lyuda Shagina, and Hongfang Liu. Representing information in patient reports using natural language processing and the extensible markup language. *Journal of the American Medical Informatics Association*, 6(1):76–87, 1999.

[27] Carol Friedman, Lyudmila Shagina, Yves Lussier, and George Hripcsak. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, 11(5):392–402, 2004.

[28] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.

[29] Travis R Goodwin and Sanda M Harabagiu. Medical question answering for clinical decision support. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 297–306. ACM, 2016.

[30] Edel Greevy and Alan F Smeaton. Classifying racist texts using a support vector machine. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 468–469. ACM, 2004.

[31] Daniel T Heinze, Mark L Morsch, and John Holbrook. Mining free-text medical records. In *Proceedings of the AMIA Symposium*, page 254. American Medical Informatics Association, 2001.

[32] Christopher M Homan, Naiji Lu, Xin Tu, Megan C Lytle, and Vincent Silenzio. Social structure and depression in trevorspace. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 615–625. ACM, 2014.

[33] Liangjie Hong, Ovidiu Dan, and Brian D Davison. Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web*, pages 57–58. ACM, 2011.

[34] Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88. ACM, 2010.

[35] George Hripcsak, John HM Austin, Philip O Alderson, and Carol Friedman. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports 1. *Radiology*, 224(1):157–163, 2002.

[36] Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. Aidr: Artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 159–162. ACM, 2014.

[37] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213. Association for Computational Linguistics, 2012.

[38] Elihu Katz and Paul Felix Lazarsfeld. *Personal Influence, The part played by people in the flow of mass communications*. Transaction Publishers, 1966.

[39] Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A Smith. A dependency parser for tweets. 2014.

[40] Irene Kwok and Yuzhou Wang. Locate the hate: Detecting tweets against blacks. In *AAAI*, 2013.

[41] Jiwei Li and Claire Cardie. Early stage influenza detection from twitter. *arXiv preprint arXiv:1309.7340*, 2013.

[42] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain, 2004.

[43] Yingjie Lu, Pengzhu Zhang, and Shasha Deng. Exploring health-related topics in online health community using cluster analysis. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on*, pages 802–811. IEEE, 2013.

[44] Altaf Mahmud, Kazi Zubair Ahmed, and Mumit Khan. Detecting flames and insults in text. 2008.

[45] Albert Park, Andrea L Hartzler, Jina Huh, David W McDonald, and Wanda Pratt. Automatically detecting failures in natural language processing tools for online community text. *Journal of medical Internet research*, 17(8):e212–e212, 2014.

[46] Michael J Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. *Icwsm*, 20:265–272, 2011.

[47] Nick Pendar. Toward spotting the pedophile telling victim from predator in text chats. In *ICSC*, volume 7, pages 235–241, 2007.

[48] Simone Paolo Ponzetto and Michael Strube. Knowledge derived from wikipedia for computing semantic relatedness. *J. Artif. Intell. Res.(JAIR)*, 30:181–212, 2007.

[49] Soumajit Pramanik, Maximilien Danisch, Qinna Wang, and Bivas Mitra. An empirical approach towards an efficient "whom to mention?" twitter app. In *Twitter for Research, 1st International & Interdisciplinary Conference*, 2015.

[50] Adithya Rao, Nemanja Spasojevic, Zhisheng Li, and Trevor Dsouza. Klout score: Measuring influence across multiple social networks. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 2282–2289. IEEE, 2015.

[51] Joshua Ritterman, Miles Osborne, and Ewan Klein. Using prediction markets and twitter to predict a swine flu pandemic. In *1st international workshop on mining social media*, volume 9, pages 9–17. ac. uk/miles/papers/swine09. pdf (accessed 26 August 2015), 2009.

[52] Kirk Roberts and Sanda M Harabagiu. A flexible framework for deriving assertions from electronic medical records. *Journal of the American Medical Informatics Association*, 18(5):568–573, 2011.

[53] Everett M Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.

[54] Koustav Rudra, Subham Ghosh, Niloy Ganguly, Pawan Goyal, and Saptarshi Ghosh. Extracting situational information from microblogs during disaster events: a classification-summarization approach. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 583–592. ACM, 2015.

[55] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

[56] Daniel Scanfeld, Vanessa Scanfeld, and Elaine L Larson. Dissemination of health information through social networks: Twitter and antibiotics. *American journal of infection control*, 38(3):182–188, 2010.

[57] Naveen Kumar Sharma, Saptarshi Ghosh, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. Inferring who-is-who in the twitter social network. *ACM SIGCOMM Computer Communication Review*, 42(4):533–538, 2012.

[58] Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. Analyzing the targets of hate in online social media. *arXiv preprint arXiv:1603.07709*, 2016.

[59] Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. Snomed clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association, 2001.

[60] Hongkui Tu, Zongyang Ma, Aixin Sun, and Xiaodong Wang. When metamap meets social media in healthcare: Are the word labels correct? In *Information Retrieval Technology*, pages 356–362. Springer, 2016.

[61] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[62] Svitlana Volkova, Theresa Wilson, and David Yarowsky. Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual twitter streams. In *ACL (2)*, pages 505–510, 2013.

[63] Beidou Wang, Can Wang, Jiajun Bu, Chun Chen, Wei Vivian Zhang, Deng Cai, and Xiaofei He. Whom to mention: expand the diffusion of tweets by@ recommendation on micro-blogging systems. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1331–1340. ACM, 2013.

[64] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. Cursing in english on twitter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 415–425. ACM, 2014.

[65] Thomas Webler, Stentor Danielson, and Seth Tuler. Using q method to reveal social perspectives in environmental research. *Greenfield MA: Social and Environmental Research Institute*, 54:1–45, 2009.

[66] Matthew Williams and Olivia Pearson. Hate crime and bullying in the age of social media. 2016.

[67] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456. ACM, 2013.

[68] Fu-Chen Yang, Anthony JT Lee, and Sz-Chen Kuo. Mining health social media with sentiment analysis. *Journal of medical systems*, 40(11):236, 2016.

[69] Tauhid Zaman, Emily B Fox, Eric T Bradlow, et al. A bayesian approach for predicting the popularity of tweets. *The Annals of Applied Statistics*, 8(3):1583–1611, 2014.

[70] Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1513–1522. ACM, 2015.