Social Media Analytics for Crisis Response

by

Shamanth Kumar

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2015 by the
Graduate Supervisory Committee:

Huan Liu, Chair
Hasan Davulcu
Ross Maciejewski
Nitin Agarwal

ARIZONA STATE UNIVERSITY

May 2015

ABSTRACT

Crises or large-scale emergencies such as earthquakes and hurricanes cause massive damage to lives and property. Crisis response is an essential task to mitigate the impact of a crisis. An effective response to a crisis necessitates information gathering and analysis. Traditionally, this process has been restricted to the information collected by first responders on the ground in the affected region or by official agencies such as local governments involved in the response. However, the ubiquity of mobile devices has empowered people to publish information during a crisis through social media, such as the damage reports from a hurricane. Social media has thus emerged as an important channel of information which can be leveraged to improve crisis response. Twitter is a popular medium which has been employed in recent crises. However, it presents new challenges: the data is noisy and uncurated, and it has high volume and high velocity. In this work, I study four key problems in the use of social media for crisis response: effective monitoring and analysis of high volume crisis tweets, detecting crisis events automatically in streaming data, identifying users who can be followed to effectively monitor crisis, and finally understanding user behavior during crisis to detect tweets inside crisis regions. To address these problems I propose two systems which assist disaster responders or analysts to collaboratively collect tweets related to crisis and analyze it using visual analytics to identify interesting regions, topics, and users involved in disaster response. I present a novel approach to detecting crisis events automatically in noisy, high volume Twitter streams. I also investigate and introduce novel methods to tackle information overload through the identification of information leaders in information diffusion who can be followed for efficient crisis monitoring and identification of messages originating from crisis regions using user behavior analysis.

DEDICATION

To my family for their love and support without which this would not have been possible.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Huan Liu, without whose guidance, support, and encouragement the completion of this degree would not have been possible. His constant support has been invaluable in the successful completion of this degree. I would like to thank my committee members, Dr. Hasan Davulcu, Dr. Ross Maciejewski, and Dr. Nitin Agarwal, who have provided excellent guidance and helped me look beyond my current endeavors. I would like to sincerely thank my colleagues at the Data Mining and Machine Learning Lab at ASU for their constructive criticism and helpful suggestions. I would like to thank the volunteers at Humanity Road Inc., particularly Cat Graham and Christine Thompson for the deep insight into the complexities and challenges of disaster response. Last but not least, I would like to thank my family and friends who have supported me with their love and encouragement. I would also like to extend my deepest gratitude to Dr. Rebecca Goolsby from the Office of Naval Research for her continuous support of my research and for many interesting discussions throughout my time at ASU.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

INTRODUCTION

Crisis response is a critical aspect of mitigating its effects. Crisis response efforts include various activities by first responders on the ground, such as distribution of food and medicine etc. Information aggregation and analysis are important components of this task. This is essential to gain situational awareness during and after a crises and to plan crisis response activities. In this thesis I introduce and address challenges to accomplishing these tasks using social media.

In the last decade, social media has played an important role during such events. Social media is providing a new information channel where people can voice their concerns and grievances. This is enabling the masses to become actively involved in the reporting of news compared to traditional media where they are passive observers. During the hurricane Sandy in 2012, Twitter was used to describe and record the damage caused by the biggest storm to hit the Eastern seaboard of U.S.A. (Guskin, Emily and Hitlin, Paul (2012)). Thus, microblogging services, such as Twitter are emerging as an alternative news source. For example, the progress of the attack on Westgate Mall in Kenya in 2013 was published in real-time (Khamadi Were (2013)). This phenomenon has been termed as citizen journalism or participatory journalism (Bowman and Willis (2003)). Thus, social media, in general, has emerged as an important source of information. Consequently, there is growing interest in leveraging social media information for humanitarian assistance and disaster relief (HADR) as part of crisis response.

Microblogging sites such as Twitter have a large and active user population. For example, the number of active users on Twitter has been increasing consistently since

2011. As of 2013, there are more than 215 million active users (Kim (2013)). Thus, microblogging platforms provide an important and fertile environment to conduct research on the usage of social media to aid crisis response. The near-real time nature of these platforms and convenient programming interfaces facilitate data access. Therefore, in this thesis I will use microblogging platforms as an example to illustrate challenges and opportunities of using social media for crisis response.

The content on microblogging platforms is also different from traditional curated media sources, such as news media in the following aspects warranting the development of new methods and processes to tackle the following challenges: a) Twitter users produce and consume information in a very informal manner compared with traditional media (Paris *et al.* (2012)). Mis-spellings, abbreviations, contractions, and slang are rampant in microblog posts; b) microblogging platforms are characterized by very high volume of information, often including large fraction of banal chatter as they tend to be uncurated.

Motivated by the use of microblogging platforms as an information sharing platform and the needs of first responders and analysts in the light of above challenges, I will organize the discussion of research problems towards addressing four interconnected problems. These topics will address different aspects of leveraging social media for crisis response.

- **Monitoring Crisis Events:** To leverage the information on Twitter for disaster response, we must first aggregate crisis information. I will introduce a visual analytics based platform, where first responders may collaborate to monitor active crises and analyze the collected data either in near real-time or post crisis to study its impact. The system TweetTracker, which automates the data collection process and provides a conducive environment for first responders and official agencies to collectively monitor emerging crises is introduced in Chapter 2. Since Twitter data is an example of big data, handling such data requires special care so as not to overwhelm the user with too much information. I address this challenge through a visual analytics based system TweetXplorer, to assist first responders to efficiently navigate large volumes of information in Chapter 2.

- **Identifying Crisis Events in Streams:** Identifying crisis events manually can be challenging given the high volume of data. Active monitoring of crisis events can greatly benefit from automated methods for crisis event detection on Twitter. In Chapter 3, I will outline a method to automatically extract crisis events in noisy and dynamic Twitter streams.

- **Identifying Relevant Users During Crisis:** Microblogging data is typically high volume data. Additionally, this data is noisy due to the global visibility of the medium. Thus, identifying crisis relevant information quickly is a challenge. In Chapter 4, I will discuss the challenges of identifying information leaders and present a method to identify users who can be followed during a crisis for faster monitoring. Twitter has a large number of users who are inactive or new. Typically, we must wait for sufficient information in the form of content or other data to be published before such users can be identified. In Chapter 5, I will introduce a study of the relationship between the limited information available in user profile and future behavior such as popularity. I will present a method to identify users who may become popular in future. Thus, enabling us to identify users to follow during crises and other events in the absence of any historical information and monitor emerging events before they peak.

- **Identifying Relevant Information During Crisis:** Fast access to tweets from crisis regions which may contain information pertinent to the crisis is essential to tackle information overload. A challenge here is that a very small fraction of the users typically provide location information. In Chapter 6, I will present a behavior analytics based technique to identify tweets inside crisis regions, so the first responders can use this information to prioritize the information consumption.

Chapter 2

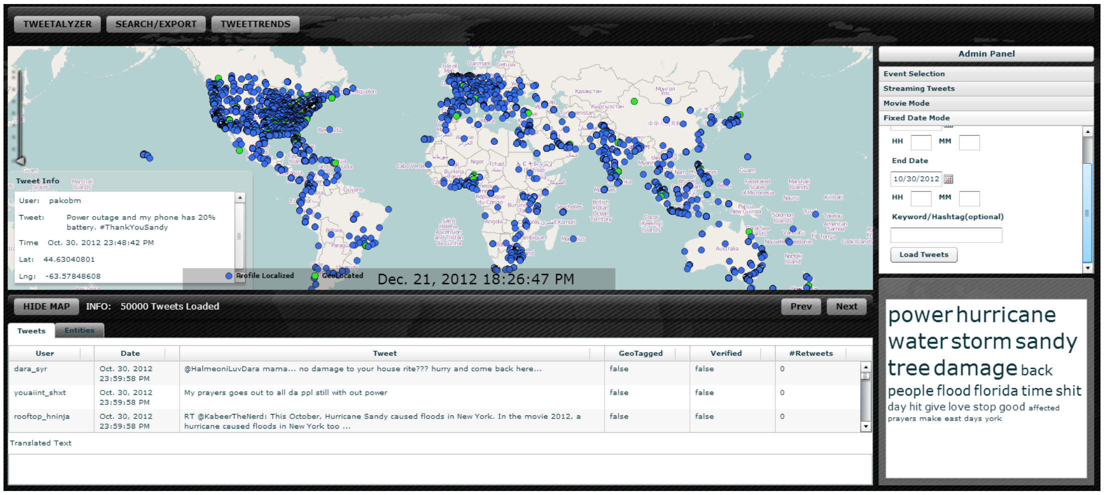
MONITORING & ANALYZING CRISIS EVENTS THROUGH VISUAL ANALYTICS

2.1 Introduction

The first step in responding to a crisis is aggregating information on the ground. Until recently, methods to facilitate the automated collection of social media data during crisis were very limited. An important challenge of working with Twitter data is the tremendously high volume of tweets generated. Typically, this is in the order of millions and therefore it is not tractable to manually collect or analyze them. Few existing systems are designed to facilitate the collection and analysis of data for emerging crises. To address this problem, I introduce two visual analytics based platforms: the Twitter data collection and analysis platform TweetTracker (Kumar *et al.* (2011a)), and the data exploratory platform TweetXplorer (Morstatter *et al.* (2013)). TweetTracker is one of the first systems to fill the void between the needs of first responders and tweets generated during a crisis.

2.2 TweetTracker: Monitoring Tweets Through Visual Analytics

TweetTracker is a tweet monitoring system designed to aid first responders in analyzing disaster-related tweets. The system focuses on collaborative data collection from Twitter using a visual interface. Its primary function is to make Twitter data, conveniently and instantly accessible to the analysts for information gathering and situational awareness during a crisis and for intelligent decision making for disaster relief efforts. To support the monitoring of tweets, TweetTracker has an advanced



Tweet Info

User: pakobm

Tweet: Power outage and my phone has 20% battery. #ThankYouSandy

Time Oct. 30, 2012 23:48:42 PM

Lat: 44.63040801

Lng: -63.57848608

Dec. 21, 2012 18:26:47 PM

INFO: 50000 Tweets Loaded

User	Date	Tweet	GeoTagged	Verified	#Retweets
dara_syr	Oct. 30, 2012 23:59:58 PM	@HalmeoniLuvDara mama... no damage to your house rite??? hurry and come back here...	false	false	0
youaiint_shxt	Oct. 30, 2012 23:59:58 PM	My prayers goes out to all da ppl still with out power	false	false	0
rooftop_hninja	Oct. 30, 2012 23:59:58 PM	RT @KabeerTheNerd: This October, Hurricane Sandy caused floods in New York. In the movie 2012, a hurricane caused floods in New York too ...	false	false	0

Translated Text

HH □ MM □

End Date

10/30/2012

HH □ MM □

Keyword/Hashtag(optional)

□

Figure 2.1: TweetTracker: The central pane is focused on geographic region (blue dots represent tweets whose location has been determined from the user's profile and green dots are geotagged tweets), the bottom-left panel lists individual tweets with characteristics, such as the verified status of the user. Non-English tweets can be translated on demand when analyzing foreign language tweets, such as in the case of the Arab Spring. The panel on the bottom-right of the figure summarizes the text in the tweet in the form of popular words.

job-monitoring interface which enables first responders and analysts to collaboratively identify parameters which describe the event, such as a hurricane. These parameters include keywords, hashtags, geographic bounding boxes, and usernames, which are determined based on the user's firsthand knowledge of the event and the region. To facilitate collaborative collection of tweets, users may share and edit the collection parameters for the events in the system. This strategy mimics the collection of information in the real-world, where information is shared across different agencies and enables better coverage of tweets for an event from different perspectives.

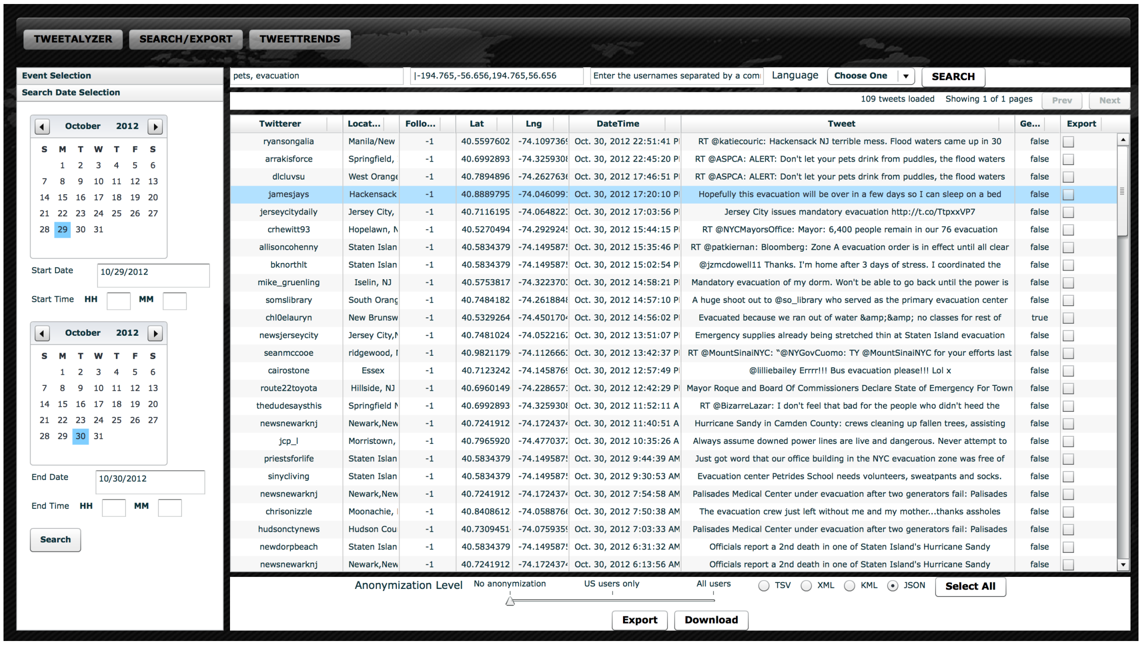
As the volume of collected information is typically too large to peruse manually, the system provides visual analytics to aid the user through aggregated views.

Namely, TweetTracker supports the following views: temporal information that can be used to identify times of peak activity; geospatial information used to identify the location of people affected by the disaster; and content-related views to identify the hot topics on Twitter. An example of these views is shown in Figure 2.1. The system automatically detects a user's location when explicit location information is unavailable and also provides aggregate statistics on frequent hashtags, users, and resources discussed in the tweets.

In addition to the data collection and visualization features offered by TweetTracker, the system also offers mature search functionality. Shown in Figure 2.2, the system allows a user to filter along different perspectives to retrieve information relevant to their interests. Search options include text, geographical regions, language of the tweet, and specific users. Additionally, TweetTracker also supports exporting of information in popular export formats such as TSV, KML, and JSON.

2.3 TweetXplorer: Analyzing Tweets Through Visual Analytics

Due to the large volume of data typically generated during a crisis, it is impractical to analyze it manually. Here, visual analytics can be a valuable tool to facilitate deeper analysis of the data. In TweetXplorer, I build upon the data visualization capabilities of TweetTracker to provide exploratory tools to analysts investigating tweets during a crisis, to answer the following questions: *who* is important in the discussions, *where* are the relevant tweets originating from, and *when* do different trends propagate. By providing deeper knowledge about the Twitter data, the user can better understand the situation on the ground. An overview of TweetXplorer is presented in Figure 2.3. Below, I will use the tweets collected during Hurricane Sandy to explain the different components of the system. Later through a case study, I will illustrate a typical use-case of an analyst monitoring a crisis.



Start Date 10/29/2012

Start Time HH MM

End Date 10/30/2012

End Time HH MM

pets, evacuation

|-194.765,-56.656,194.765,56.656

Enter the usernames separated by a com

109 tweets loaded   Showing 1 of 1 pages

Twitterer	Locat...	Follo...	Lat	Lng	DateTime	Tweet	Ge...	Export
ryansongalia	Manila/New	-1	40.5597602	-74.109736	Oct. 30, 2012 22:51:41 P	RT @katiecouric: Hackensack NJ terrible mess. Flood waters came up in 30	false	□
arrakisforce	Springfield,	-1	40.6992893	-74.325930	Oct. 30, 2012 22:45:20 P	RT @ASPCA: ALERT: Don't let your pets drink from puddles, the flood waters	false	□
dlcluvsu	West Orange	-1	40.7894896	-74.262763	Oct. 30, 2012 17:46:51 P	RT @ASPCA: ALERT: Don't let your pets drink from puddles, the flood waters	false	□
jamesjays	Hackensack	-1	40.8889795	-74.046099	Oct. 30, 2012 17:20:10 P	Hopefully this evacuation will be over in a few days so I can sleep on a bed	false	□
jerseycitydaily	Jersey City,	-1	40.7116195	-74.064822	Oct. 30, 2012 17:03:56 P	Jersey City issues mandatory evacuation http://t.co/TtpxxVP7	false	□
crhewitt93	Hopelawn, N	-1	40.5270494	-74.292924	Oct. 30, 2012 15:44:15 P	RT @NYCMayorsOffice: Mayor: 6,400 people remain in our 76 evacuation	false	□
allisoncohenny	Staten Islan	-1	40.5834379	-74.149587	Oct. 30, 2012 15:35:46 P	RT @patkiernan: Bloomberg: Zone A evacuation order is in effect until all clear	false	□
bknorthlt	Staten Islan	-1	40.5834379	-74.149587	Oct. 30, 2012 15:02:54 P	@jzmcdowell11 Thanks. I'm home after 3 days of stress. I coordinated the	false	□
mike_gruenling	Iselin, NJ	-1	40.5753817	-74.322370	Oct. 30, 2012 14:58:21 P	Mandatory evacuation of my dorm. Won't be able to go back until the power is	false	□
somslibrary	South Orang	-1	40.7484182	-74.261884	Oct. 30, 2012 14:57:10 P	A huge shoot out to @so_library who served as the primary evacuation center	false	□
chl0elauryn	New Brunsw	-1	40.5329264	-74.450170	Oct. 30, 2012 14:56:02 P	Evacuated because we ran out of water &amp;&amp; no classes for rest of	true	□
newsjerseycity	Jersey City,N	-1	40.7481024	-74.052216	Oct. 30, 2012 13:51:07 P	Emergency supplies already being stretched thin at Staten Island evacuation	false	□
seanmccooe	ridgewood, N	-1	40.9821179	-74.112666	Oct. 30, 2012 13:42:37 P	RT @MountSinaiNYC: "@NYGovCuomo: TY @MountSinaiNYC for your efforts last	false	□
cairostone	Essex	-1	40.7123242	-74.145876	Oct. 30, 2012 12:57:49 P	@lilliebailey Errrr!!! Bus evacuation please!!! Lol x	false	□
route22toyota	Hillside, NJ	-1	40.6960149	-74.228657	Oct. 30, 2012 12:42:29 P	Mayor Roque and Board Of Commissioners Declare State of Emergency For Town	false	□
thedudesaysthis	Springfield N	-1	40.6992893	-74.325930	Oct. 30, 2012 11:52:11 A	RT @BizarreLazar: I don't feel that bad for the people who didn't heed the	false	□
newsnewarknj	Newark,New	-1	40.7241912	-74.172437	Oct. 30, 2012 11:40:51 A	Hurricane Sandy in Camden County: crews cleaning up fallen trees, assisting	false	□
jcp_l	Morristown,	-1	40.7965920	-74.477037	Oct. 30, 2012 10:35:26 A	Always assume downed power lines are live and dangerous. Never attempt to	false	□
priestsforlife	Staten Islan	-1	40.5834379	-74.149587	Oct. 30, 2012 9:44:39 AM	Just got word that our office building in the NYC evacuation zone was free of	false	□
sinycliving	Staten Islan	-1	40.5834379	-74.149587	Oct. 30, 2012 9:30:53 AM	Evacuation center Petrides School needs volunteers, sweatpants and socks.	false	□
newsnewarknj	Newark,New	-1	40.7241912	-74.172437	Oct. 30, 2012 7:54:58 AM	Palisades Medical Center under evacuation after two generators fail: Palisades	false	□
chrisonizzle	Moonachie,	-1	40.8408612	-74.058876	Oct. 30, 2012 7:50:38 AM	The evacuation crew just left without me and my mother...thanks assholes	false	□
hudsonctynews	Hudson Cou	-1	40.7309451	-74.075935	Oct. 30, 2012 7:03:33 AM	Palisades Medical Center under evacuation after two generators fail: Palisades	false	□
newdorpbeach	Staten Islan	-1	40.5834379	-74.149587	Oct. 30, 2012 6:31:32 AM	Officials report a 2nd death in one of Staten Island's Hurricane Sandy	false	□
newsnewarknj	Newark,New	-1	40.7241912	-74.172437	Oct. 30, 2012 6:13:56 AM	Officials report a 2nd death in one of Staten Island's Hurricane Sandy	false	□

Anonymization Level: [✓ No anonymization] US users only All users

○ TSV ○ XML ○ KML ● JSON

Figure 2.2: Search interface in TweetTracker. Search parameter choices are visible at the top, the list of returned tweets, and the export options are at the bottom.

This system enables the user to investigate data along multiple user-defined facets. It allows users to analyze tweets using user-defined themes. These themes are defined as a collection of query terms or hashtags, for example a theme on "evacuation" can be defined using the following terms: #traffic, evacuation, and shelters. The analysis of the data can then be performed along these themes to compare, contrast, and identify patterns in the collected crisis data. Figure 2.4, describes three themes defined by the user to investigate tweets collected during Hurricane Sandy: evacuation from the affected regions, requests for help, and infrastructure damage.

An innovative component of TweetXplorer is the information propagation network visible in the top-right panel of Figure 2.3. During a crisis studying how information propagates can provide valuable insights into the central individuals who help in information dissemination and who may be contacted for access to information in the future. Twitter is particularly suitable for such monitoring due to the concept of

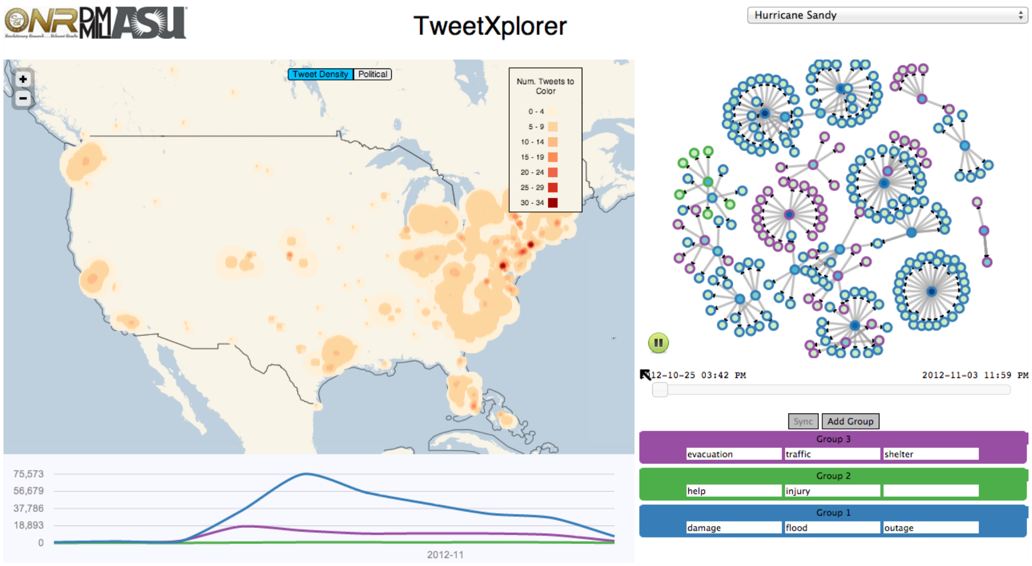


Figure 2.3: TweetXplorer: The top-left panel shows the User × User retweet network. The top-right panel shows the geographic distribution of geotagged tweets. The bottom-right panel shows the grouped keyword queries.

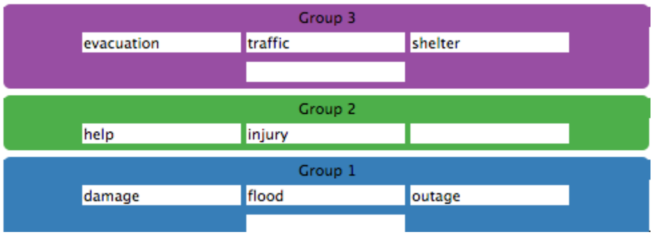
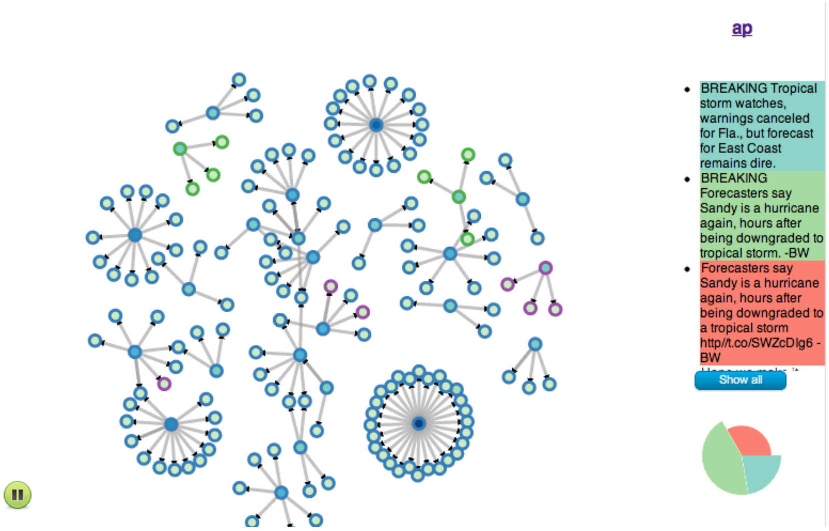


Figure 2.4: A Closer Look of the Keywords Used to Generate the Views

retweets, which naturally convey the direction of flow of information.

Figure 2.5 shows an example of a retweet network generated from a one hour sample of data during Hurricane Sandy. Each node represents a user and each edge represents a retweet relationship between two users. The direction of the edge indicates the flow of information between the users. By viewing this network, the analyst can identify important users who produced the most influential information and users who spread that information to smaller communities. A detailed view of the content



- BREAKING Tropical storm watches, warnings canceled for Fla., but forecast for East Coast remains dire.
- BREAKING Forecasters say Sandy is a hurricane again, hours after being downgraded to tropical storm. -BW
- Forecasters say Sandy is a hurricane again, hours after being downgraded to a tropical storm http//t.co/SWZcDIg6 -BW

Figure 2.5: Retweet Network Identifying the Direction and Involved Nodes in Information Propagation

generated by a user can be observed by selecting the user. The information panel on the right describes the tweets generated by the user and summarizes the influence of each tweets in terms of number of retweets. The tweets can themselves be selected to further identify the subset of users who have retweeted a particular tweet.

The map component allows a user to see geographic regions which received the most attention through heatmaps. The density of tweets in a region is indicated by the color of the region, with darker colors, indicating higher density. The map also serves as a summarization of the locations of users who are retweeting the selected user. Thus, allowing an analyst to focus on users who are in the region of interest.

The components of TweetXplorer also interact with each other to provide more contextual views of the data. For example, one can select a user in the retweet network to view the geographic origins of his retweets. This can give us an idea of the regions interested in the tweet's content. The selection of a region on the map similarly filters the retweet network to show only the users from that region.

2.4 Related Work

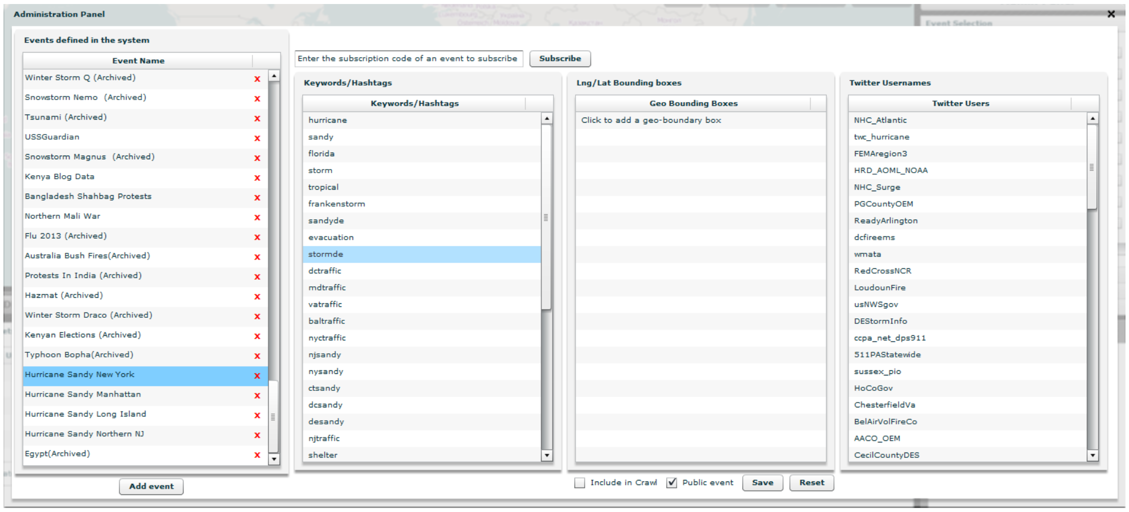
Building Twitter-based systems to solve real world problems is an active area of research. *Twitris* (Purohit and Sheth (2013)) is an online platform for event analysis using Twitter. The system combines geospatial visualization, user network visualization, and sentiment analysis to aid its users in analyzing events via different perspectives in near real-time. *TwitterMonitor* (Mathioudakis and Koudas (2010)) is a system to detect emerging topics or trends in a Twitter stream. The system identifies bursty keywords as an indicator of emerging trends, and periodically groups them together to form emerging topics. Detected trends can be visually analyzed through the system. *TEDAS* (Li *et al.* (2012)) is an event detection and analysis system focused on crime and disaster events. TEDAS crawls event related tweets using a rule-based approach. Detected events are analyzed to extract temporal and spatial information. The system also uses the location information of the author's network to predict the location of a tweet when the tweet is not geotagged. *SensePlace2* (MacEachren *et al.* (2011)) supports collection and analysis of Tweets for keyword searches on-demand. The system focuses on three primary views: text, map, and timeline, to enable exploration of data and to acquire situational awareness.

Geographical visualization systems have also been used to monitor non-Twitter data. In *BirdVis* (Ferreira *et al.* (2011)), the authors introduce a new tool to understand bird habitat preferences called Tagcloud lenses. This technique combines geospatial and temporal information with the textual information to highlight key differences in habitat preferences of birds over time. In Dykes *et al.* (2010), the authors extol the virtues of having dynamic map legends, exhibiting legends that go beyond mapping colors and glyphs to values. Instead, the authors create a mapping system, *Strategi*, which utilizes the legend to show the statistical properties of the graph. They also experiment with removing the legend entirely, allowing the colors on the map tell their own story. In Nocke *et al.* (2007), the authors use "brushing" on their maps to show the selected region in the form of "zoom and filter".

Network visualization is a popular topic of research and there are several systems, which can be used to visualize networks, and also compute various network measures to understand the network. Brandes *et al.* (2006) provide a history of network visualizations. Later they propose 5 characteristics that a network visualization should have, which they include in their system, Visone (Baur and Schank (2008)). Cytoscape (Shannon *et al.* (2003)) is a tool developed for visualizing biological networks. It is an open source software with a variety of network visualization layouts built into it and it also supports common file formats for data import/export. Cytoscape also scales very well to large networks as is the case in social network visualization. Gephi (Bastian *et al.* (2009)) is another open source graph visualization software which performs many of the functions available in Cytoscape. ORA (Carley *et al.* (2013)) is a commercial network visualization software which has an extensive range of metrics computable from the visualized network and is more focused on network analysis.

2.5 Tweet Aggregation to Crisis Insight: A Case Study

To provide an example of how the systems and their capabilities may be used in a real-world scenario, I present a study below highlighting a typical use case as observed from the first responders who have used the systems. Tweets for this case study were collected during Hurricane Sandy. A storm of this magnitude is highly unusual in this region of the United States, and as a result the disaster generated a tremendous amount of Twitter activity. The parameters to collect the discussion



Event Name	
Winter Storm Q (Archived)	x
Snowstorm Nemo (Archived)	x
Tsunami (Archived)	x
USSGuardian	x
Snowstorm Magnus (Archived)	x
Kenya Blog Data	x
Bangladesh Shahbag Protests	x
Northern Mali War	x
Flu 2013 (Archived)	x
Australia Bush Fires(Archived)	x
Protests In India (Archived)	x
Hazmat (Archived)	x
Winter Storm Draco (Archived)	x
Kenyan Elections (Archived)	x
Typhoon Bopha(Archived)	x
Hurricane Sandy New York	x
Hurricane Sandy Manhattan	x
Hurricane Sandy Long Island	x
Hurricane Sandy Northern NJ	x
Egypt(Archived)	x

Keywords/Hashtags
hurricane
sandy
florida
storm
tropical
frankenstorm
sandyde
evacuation
stormde
dctraffic
mdtraffic
vatraffic
baltraffic
nyctraffic
njsandy
nysandy
ctsandy
dcsandy
desandy
njtraffic
shelter

Geo Bounding Boxes
Click to add a geo-boundary box

Twitter Users
NHC_Atlantic
twc_hurricane
FEMAregion3
HRD_AOML_NOAA
NHC_Surge
PGCountyOEM
ReadyArlington
dcfireems
wmata
RedCrossNCR
LoudounFire
usNWSgov
DEStormInfo
ccpa_net_dps911
511PAStatewide
sussex_pio
HoCoGov
ChesterfieldVa
BelAirVolFireCo
AACO_OEM
CecilCountyDES

Figure 2.6: Parameters Used to Collect Tweets Related to Hurricane Sandy

were provided by volunteers from Humanity Road [1] and an analyst from the Office of Naval Research through TweetTracker and consisted of storm related keywords and Twitter usernames. The collection started on October 25, 2012 and continued through November 3, 2012, during which time I collected 5,639,643 tweets. Figure 2.6 shows the collaborative editing panel used by these users, where they can add and modify the parameters used to crawl the crisis tweets. An overview of the collected tweets from October 30 can be seen in Figure 2.1. Here, the content panel is indicative of the prominent topics of discussion, which include the words “power” and “water” indicative of the outages in the region and a popular topic of discussion among the people affected by the crisis.

2.5.1 Investigating the Data

Consider a scenario, where an analyst intends to investigate tweets to understand Hurricane Sandy’s impact. Clearly, the first step would be to identify regions of interest. This can be determined by analyzing the patterns in tweet traffic from the

[1] http://humanityroad.org/

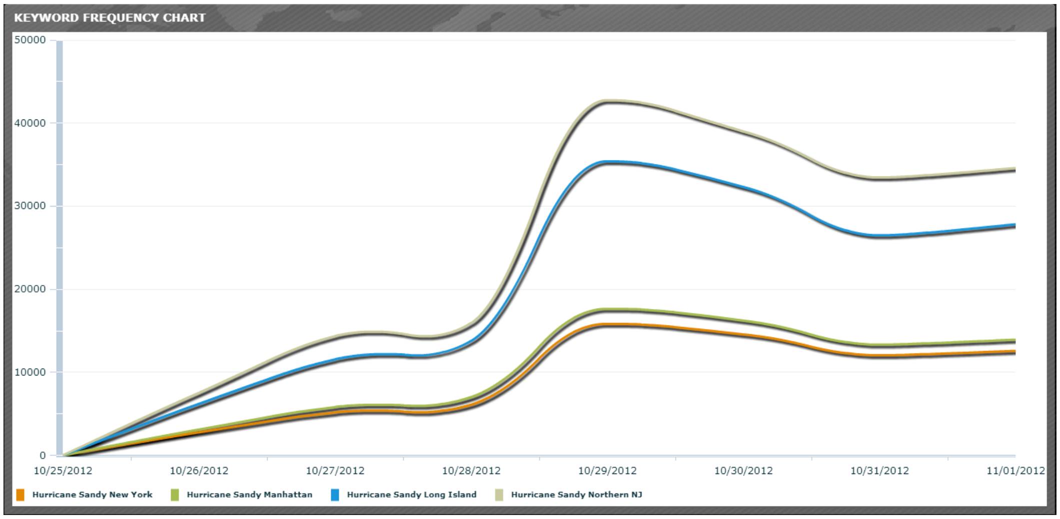
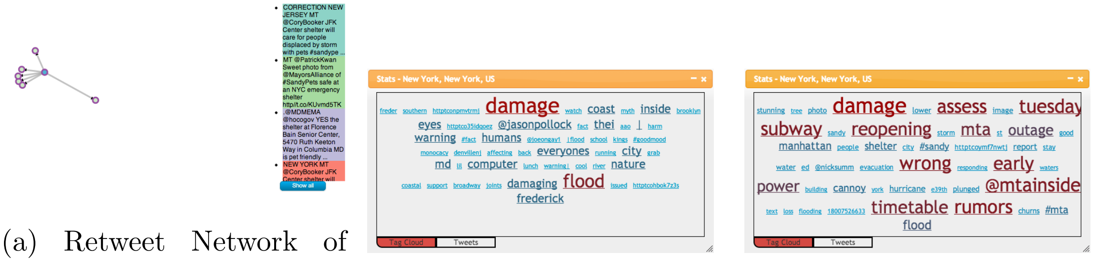


Figure 2.7: Traffic Trend for Severely Affected Areas of Hurricane Sandy

regions on Sandy's path. In Figure 2.7, I present a comparison of the traffic from different parts of New York and New Jersey, the most severely affected regions. The traffic patterns indicate that tweets from northern NJ indicate high interest on the topic. It is also clear that the volume of tweets is highest on the day of the landfall (October 29). The next step in this investigation would be to understand and contrast the patterns in the content of tweets before, during, and after the disaster by drilling into the data. Towards this I partition the dataset into three distinct epochs: pre-landfall (2012-10-29 00:00 - 2012-10-29 17:59), landfall (2012-10-29 18:00 - 2012-10-30 23:59), and recovery (2012-10-31 00:00 - 2012-11-01 12:00). To drive the analysis along the themes observed in TweetTracker in Figure 2.1, I identify indicative keywords as in Figure 2.3. The findings from the three epochs are presented below.

Pre-Landfall: In the hours leading up to Hurricane Sandy's landfall, we see discussions representing different issues. In Figure 2.8a, one of the most highly-retweeted tweets mentions the availability of pet shelters in evacuation areas, which indicates a concern from the pet owners regarding the safety of their pets during the storm. While the geotagged tweets produced during this epoch show generic


Stats - New York, New York, US
freder southern httptconpmvtrmj damage watch coast myth inside brooklyn
eyes httptco35idqoez @jasonpollock fact thei aao | harm
warning #fact humans @joeongay1 |flood school kings #goodmood
monocacy denvillenj affecting back everyones running city grab
md lil computer lunch warning| cool river nature
coastal support broadway joints damaging flood issued httptcohbok7z3s
frederick
Tag Cloud
Tweets
Stats - New York, New York, US
stunning tree photo damage lower assess image tuesday
subway sandy reopening storm mta st outage good
manhattan people shelter city #sandy httptcoymf7nwtj report stay
water ed @nicksumm evacuation wrong responding early waters
power building cannoy york hurricane e39th plunged @mtainside
text loss flooding 18007526633 timetable rumors churns #mta
flood
Tag Cloud
Tweets
Show all

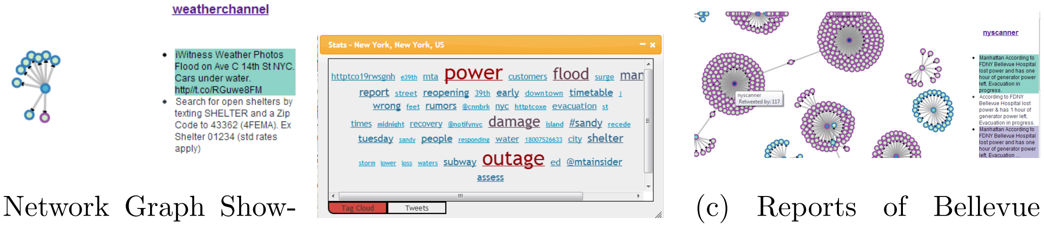
- CORRECTION NEW JERSEY MT @CoryBooker JFK Center shelter will care for people displaced by storm with pets #sandype ...
- MT @PatrickKwan Sweet photo from @MayorsAlliance of #SandyPets safe at an NYC emergency shelter http//t.co/KUvmd5TK
- .@MDMEMA @hocogov YES the shelter at Florence Bain Senior Center, 5470 Ruth Keeton Way in Columbia MD is pet friendly ...
- NEW YORK MT @CoryBooker JFK Center shelter will

(a) Retweet Network of @humanesociety Discussing Pet-Friendly Evacuation Shelters

(b) Tag Cloud of Frequent Words in New York 1 Day Before Landfall

(c) Tag Cloud of Frequent Words in New York; 1 Hour Before Landfall

Figure 2.8: Pre-Landfall Discussion of Hurricane Sandy


weatherchannel
Stats - New York, New York, US
httptco19rwsgnh e39th mta power customers flood surge man
report street reopening 39th early downtown timetable |
wrong feet rumors @cnnbrk nyc httptcoxe evacuation st
times midnight recovery @notifynyc damage island #sandy recede
tuesday sandy people responding water 18007526633 city shelter
storm lower loss waters subway outage ed @mtainsider
assess
Tag Cloud
Tweets
nyscanner
nyscanner
Retweeted by: 117

- iWitness Weather Photos Flood on Ave C 14th St NYC. Cars under water. http//t.co/RGuwe8FM
- Search for open shelters by texting SHELTER and a Zip Code to 43362 (4FEMA). Ex Shelter 01234 (std rates apply)

- Manhattan According to FDNY Bellevue Hospital lost power and has one hour of generator power left, Evacuation in progress.
- According to FDNY Bellevue Hospital lost power & has 1 hour of generator power left, Evacuation in progress.
- Manhattan According to FDNY Bellevue Hospital lost power and has one hour of generator power left, Evacuation ...

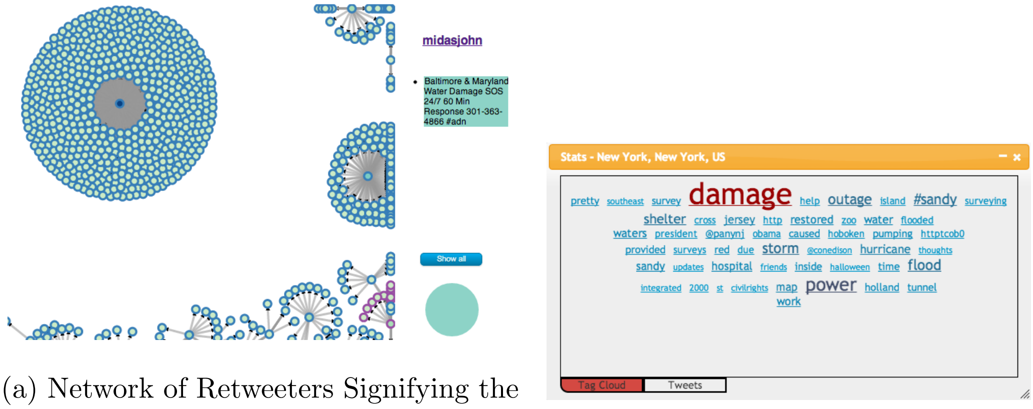
(a) Network Graph Showing the First Reports of Flooding in NYC

(b) Reports of Power Outages in and Around NYC.

(c) Reports of Bellevue Hospital Evacuation from @NYScanner

Figure 2.9: Discussion of Hurricane Sandy Immediately After Landfall

discussion with no clear topic as a focus. At the beginning of the epoch “damage”, and “flood” are ranked highly in the New York area. However, as the storm neared, the content in Figure 2.8c shows that specific issues such as “rumors”, “damage”, and “subway” became popular. **Landfall:** Hurricane Sandy made landfall on Oct 29, 2012 at 20:00 EST (NHC (2012)). First reports of flooding started to arrive around this time. As observable in Figure 2.9a, these reports contain links to images of flooding. As the storm progressed reports of power outage became prominent. Con Edison, New York’s power supplier, claimed that this was the worst power outage in their history. Figure 2.9b, shows the tweets centered on power outage and flooding


midasjohn
Show all
Stats - New York, New York, US
pretty southeast survey damage help outage island #sandy surveying
shelter cross jersey http restored zoo water flooded
waters president @panynj obama caused hoboken pumping httptcob0
provided surveys red due storm @conedison hurricane thoughts
sandy updates hospital friends inside halloween time flood
integrated 2000 st civilrights map power holland tunnel
work
Tag Cloud
Tweets

- Baltimore & Maryland Water Damage SOS 24/7 60 Min Response 301-363-4866 #adn

(a) Network of Retweeters Signifying the Importance of Recovery Resources After the Hurricane

(b) Tag Cloud of Frequently Used Words in New York After the Hurricane

Figure 2.10: Analyzing Discussion During the Recovery Phase of Hurricane Sandy

are predominantly from New York City and its surrounding areas. Due to the power outage we observed that at least two hospitals were forced to evacuate their patients. In Figure 2.9c, we can see two clusters of retweets connected by common retweeters. These two users are @NYScanner and @ABC and the tweets claimed that the Bellevue Hospital and the NYU Langone Medical center were being evacuated due to power failure.

Recovery: After the storm, people turned their attention towards the estimated $71 billion in damage (Russ (2012)) caused by the hurricane. Twitter activity after the event gives us the following insight: First, the most prominent tweets on the day after the hurricane are directing people to assistance in repairing the damage done to their homes as shown in Figure 2.10a, Second Figure 2.10b shows that much of the discussion in New York City focuses on the words “damage”, “power”, and “flood”, indicating that people have turned their attention to post-storm topics, such as power outage and post-storm cleanup.

2.6 Conclusion

In this chapter, I introduced the problem of monitoring and analyzing large volume of microblogging messages. Towards this goal, I proposed two systems which leverage visual analytics to tackle the challenges involved in this task. I introduced TweetTracker, which provides a visual interface to track crisis related tweets in a collaborative fashion and enables first responders to analyze the collected information to gain insight into the crisis. To perform deeper analysis of crisis tweets, I introduced TweetXplorer, which facilitates deeper analysis of big crisis data along multiple dimensions: information propagation network, geographical distribution, temporal patterns, and tweet content.

Chapter 3

IDENTIFYING EVENTS IN SOCIAL MEDIA STREAMS

3.1 Introduction

In recent world events, social media data has been shown to be effective in detecting earthquakes (Sakaki *et al.* (2010)), rumors (Mendoza *et al.* (2010)), and identifying characteristics of information propagation (Qu *et al.* (2011)). More recently, Twitter is being used to disseminate breaking news before traditional media. For example, during the Westgate Mall Attack in Kenya in 2013, first reports of the attack were published on Twitter (Khamadi Were (2013)). Thus, an automated approach to detect such events is desired as the volume of information is too high to manually identify crisis events. However, event detection approaches designed for documents cannot be directly applied to tweets due to the difference in the characteristics. Unlike traditional documents, tweets suffer from the informality of language, and differ in both volume and velocity at which the data is generated.

Existing approaches to detecting events in tweets focus on the problem in an offline setting, where the corpus is static and multiple passes can be employed in the solution. However, the streaming environment presents unique challenges, which prevent the direct application of existing approaches:

Informal use of language: Twitter users produce and consume information in a very informal manner compared with traditional media (Paris *et al.* (2012)). Misspellings, abbreviations, contractions, and slang are rampant in tweets, which is exacerbated by the length restriction (a tweet can have no more than 140 characters).

Noise: While traditional event detection approaches assume that all documents are relevant, Twitter data typically contains a vast amount of noise and not every tweet is related to an event (Pear Analytics (2009)).

Dynamicity: Twitter streams are highly dynamic with high volume and high velocity as typical characteristics. Event detection methods need to be scalable to handle this high volume of tweets.

An event is typically defined as "A non-trivial incident happening in a certain place at a certain time" (Allan *et al.* (1998a); Yang *et al.* (1998)). Due to the growing ownership of smartphones, every user on Twitter is a sensor. And as Twitter is a popular social media site with global visibility, the occurrence of a crisis event typically leads to the publication of tweets from several different users. Thus aggregation of tweets is an intuitive method to detect crisis events. Additionally, the diversity of the users indicating the occurrence of an event also lends credibility to the event and helps tackle noise which is prominently present in social media data such as tweets.

3.2 Problem Definition

Given an ordered stream of tweets $T = t_1, t_2, t_3, ...$, where each t_i is associated with a timestamp indicating its publication time, we need to detect events $E = e_1, e_2, ...e_m$, where $e_j = t_1, t_2, ..., t_k$, where $t_k \in T$ and $j \in [1, m]$.

Event: An event is formally defined as a set of similar tweets $E = t_1, t_2, ..., t_k$ with high user diversity.

User Diversity: of an event is the diversity of the user population who contribute to the event. The intuition here is that a diverse user population lends credibility to the event and helps us filter out noise. Entropy is a measure commonly used to compute the amount of information in a text and here I reformulate it to measure user diversity of an event. Given an event e, its *User Diversity* $H(e)$ is

$$H(e) = -\sum_{i} \frac{n_{u_i}}{n} log \frac{n_{u_i}}{n}, \tag{3.1}$$

where u_i is the ith user in the cluster. Here, n_{u_i} is the number of tweets published by user u_i, which are part of the cluster C, and n is the total number of tweets in the cluster.

3.2.1 Hardness of Event Detection

To detect events E, we aim to find a clustering C of the tweets T such that the user diversity of the resulting clusters and the distance between tweets in different clusters is maximized. Let us assume, that the number of events m is known. Then, the objective function can be defined as

$$\arg\max \sum_{e_i \in E} (\sum_{j} D(e_i, e_j)) + H(e)), \qquad s.t.\ |E| = m \tag{3.2}$$

where $D(e_i, e_j)$ is the maximum distance between any pair of elements in clusters e_i and e_j computed as

$$D(e_i, e_j) = max_{a,b}(D(e_i^{(a)}, e_j^{(b)})), \qquad s.t.\ |e_i| = a, |e_j| = b. \tag{3.3}$$

To show that this function is hard, I will prove that it is submodular using the following properties:

Monotone: The function is monotone as at each step a tweet is added to the nearest cluster, hence the summation of the distances between clusters cannot decrease.

Diminishing Returns: Let us assume that S and T are two clusters, where $S \subseteq T$. If a tweet x is added to S and T, then the change in $D(S, P)$ and $D(T, P)$, where P is any other cluster follows one of two scenarios:

- If both T and S contain the tweet which maximizes D(.) and x increases the distance to P. Then $D(T \cup x, P) - D(T, P)) = D(S \cup x, P) = D(S, P)$,

- Otherwise, if there is a tweet $x' \notin S$ but $x' \in T$, such that $D(S, P) < D(T, P)$, then $D(T \cup x, P) - D(T, P) \leq D(S \cup x, P) - D(S, P)$ because $D(S, P) < D(T, P)$, therefore the gain in the distance should be larger when the new tweet is added to S.

If a function satisfies the above properties it is submodular, therefore D(.) is submodular as it satisfies both properties. The Entropy function H(.) is submodular and the summation of submodular functions is submodular (Krause and Golovin (2012)). Therefore, the objective function in Equation 3.2 is submodular. Maximizing a submodular objective function under the cardinality constraint is NP-hard (Krause and Golovin (2012)). Therefore, event detection in streams where m is unknown and cannot be determined in a timely fashion, is at least as hard. Later, we will describe the approach which approximates the objective function by assigning tweets to the most similar cluster and determines events using the cluster's user diversity.

3.3 Identifying Events

During a real-world event, people use Twitter to tweet and retweet their experiences. Therefore, the information from these users can be aggregated/clustered to detect events. However, for streaming Twitter data extra care must be taken because (1) streaming tweets arrive continuously, traditional multi-pass clustering cannot handle streaming data, and (2) the informal language of tweets defies the standard preprocessing of text corpora such as stemming and vectorization and transformations are typically expensive in this environment.

To handle high volume and high velocity streaming data, clustering approaches must return the clusters in a single pass. Therefore, we require a clustering approach, which does not require multiple passes over data and which can continuously process the tweets as they arrive. In this paper, I use the single-pass clustering technique described in Rijsbergen (1979), to group related tweets into clusters as they arrive. This incremental clustering approach continually processes tweets as follows:

1. A tweet is compared with all the candidate clusters.
2. The tweet is added to the closest cluster, if the distance to the cluster is below a threshold.
3. Otherwise, a new cluster is created and the tweet is added to it as its first member.

To cluster the tweets, a distance measure appropriate for the characteristics of a tweet stream must be chosen. We require that the distance measure: be scalable to high-volume streaming data; avoid the need for expensive data transformations, be robust to informal language, and avoid the determination and maintenance of a vocabulary as the language is constantly evolving. Next, I will discuss compression-based distance, which addresses these requirements.

3.3.1 Tackling Data Informality

Compression distance computes the distance between two texts by measuring the compression gain obtained on the merging of the two texts. It has been shown to be both efficient and effective for clustering text in Keogh *et al.* (2004). Additionally, compression based distance has been shown to be effective on multilingual text. Although only discussed in the context of traditional documents, this distance measure is able to handle tweets due to its design. On Twitter, the advantage of compression distance over traditional distance measures such as cosine similarity, is its ability to handle the informal and evolving language in tweets. While cosine similarity requires the maintenance of a vocabulary and data transformation, compression distance can be applied to text without data transformation.

Compression distance is an approximation of the Kolmogorov complexity proposed in Keogh *et al.* (2004). In this work, I consider each tweet as a document. If C is any compressor, and $C(x)$ is the size of the compressed tweet x. Then the distance between two tweets x and y, $d(x, y)$ is

$$d(x, y) = \frac{C(xy)}{C(x) + C(y)}, \tag{3.4}$$

where C(xy) is the compression achieved by merging the two tweets.

Using the above definition of compression distance between tweets, the distance between two events $D(e_1, e_2)$ can be computed as the maximum pairwise distance between any pair of tweets in the clusters.

Choosing the Compressor: Existing literature recommends choosing a compressor appropriate for the problem domain. Here, I compared 3 compression algorithms: DEFLATE, Gzip, and QuickLZ for compression speed and compression ratio using a random set of 20,000 tweets. I found that DEFLATE was the most efficient algorithm in both criteria. Therefore, I will employ it as the compressor **C** in Equation 3.4.

3.3.2 Scaling to High Volume Data

Using public Twitter APIs, one can access a sample of (1%) tweet stream, which can lead to as many as several million tweets a day. Thus, detecting events in a stream necessitates a scalable solution. Here, I present detailed solutions to scalability.

Events are dynamic and it is essential to consider the temporal evolution of the events in the task of event detection on streaming data. The incorporation of a temporal model to detect events has the following advantages:

- capturing evolving events, and
- improving the efficiency of event detection.

A cluster representing an event can be considered to be active or inactive at any given time based on the arrival of new tweets. I propose a temporal model which can be used to make this decision. An event is modeled as Poisson process, which has been traditionally used to model the number of objects in an event at time t. In a Poisson process, the rate of arrival of tweets can be modeled as an exponential distribution. This rate is represented by the parameter λ. Let's consider a random variable X, where X measures the time between successive tweets. The variable X is modeled as an exponential distribution with parameter λ as

$$X \propto exp(\lambda). \tag{3.5}$$

Given an event e and the number of tweets in each time interval in the event $x_1, x_2, ..., x_n$, the likelihood function for the inter-arrival time is

$$L(\lambda | x_1, x_2, ..., x_n) \propto f(x_1, x_2, ..., x_n | \lambda) \propto \prod_{i=0}^{n} \lambda e^{-x_i \lambda}. \tag{3.6}$$

To obtain the λ_{MLE}, we take the derivative of the log-likelihood with respect to λ and set it to zero. Then, $\lambda_{MLE} = \frac{1}{\bar{x}}$, where $\bar{x}$ is the mean of the distribution. For each cluster c, if a tweet does not arrive in λ_c time units, the current estimate for cluster c, then c is considered inactive and removed from memory. The estimate for λ_c is updated every time a new tweet is added to the cluster.

3.3.3 Identifying Events from Clusters

Tweets are noisy and not every tweet in the stream is expected to be part of an event. Therefore, not every cluster identified by the algorithm can be an event. The volume of a cluster can help us identify events, but this is susceptible to noise. As a crowdsourced information sharing platform, the diversity of the users who publish tweets in a cluster lends credibility to the information within the cluster and can be

Algorithm 1 Crisis Events Detection in Twitter Streams Using a Single-Pass Clustering

```
Input: A stream of tweets $T$ and the Cluster Limit ($k$), the Tweet Limit ($l$), the Distance Threshold ($D_t$), and the Diversity Threshold ($H_t$).
Output: Detected events $E$.
  $E \leftarrow \{\}$
  $C \leftarrow \{\}$
  while tweet $t \in T$ do
      Identify active cluster $c \in C$, where $D(t, c) \leq D_t$
      if c exists then
          Add $t$ to $c$
          Update expected time of next tweet $\lambda_c$
          Update User Diversity ($H(c)$) of cluster $c$
          if $H(c) \geq H_t$ then
              Mark cluster as an event, Add $c$ to $E$
          end if
      else
          Create new cluster $c$ with $t$ as its first member
          Add $c$ to $C$
      end if
  end while
```

used to determine whether it is an event. A cluster is classified as an event, if its Diversity Score $H(c)$ is above the *Diversity Threshold* H_t.

3.3.4 Event Detection Framework

Using the strategies to handle informal language in tweets, temporal dynamics of events, and handling noise, we can detect events in Twitter streams using Algorithm 1. To improve the efficiency of the algorithm and to scale it to large Twitter streams I propose two heuristics:

Cluster Limit: The assignment of tweets to clusters requires a comparison with currently active clusters, but sequential search of all active clusters can be prohibitive. As a tweet is more likely to be similar to clusters with overlapping content we limit the comparisons to these candidate clusters. These clusters are identified by aggregating, sorting, and ranking clusters according to their overlap with the tweet. Then, we pick the top k clusters as the candidate clusters. $k = 100$, was empirically found to be effective in discovering reasonable clusters without sacrificing the speed of the algorithm.

Tweet Limit: The distance of a tweet to an event is computed as the maximum pairwise distance with the tweets contained in the event. Due to the timely nature of tweets, I propose to restrict the comparisons ordered by recency. This could be effective when clusters represent events which span an extended period of time and contain a large numbers of tweets. I propose to restrict the comparisons to at most l recent tweets in a cluster. In my implementation, I have set $l = 1000$.

Time Complexity: Given the number of tweets in the stream as N, the Cluster Limit (k) and the Tweet Limit (l), the time complexity of our algorithm is $O(Nkl)$. The most expensive operation in our algorithm is the assignment of tweets to clusters. For every tweet in the stream, it needs to be compared to at most l tweets in k clusters. As the values of k and l are much smaller than N, the algorithm allows us to process the tweets in near real-time. In the later sections, I will present empirical evidence of the algorithm's efficiency.

Selection of Parameters: Two thresholds are used in our framework to identify events. First, the distance threshold D_t is used to determine assignment of tweets to clusters. In a study on 20,000 random tweets, I found that the average self-similarity of tweets was 0.54 and a value of 0.8 was empirically found to be a suitable value to obtain reasonable clusters. Second, the diversity threshold H_t is used to decide which clusters can be labeled as events, as noise is a problem in tweet streams. Volume or the number of tweets in a cluster is also an important factor in determining whether a cluster is an event. Ideally, it is preferable for clusters to contain many tweets and have high user diversity. This threshold was set empirically as outlined later.

In the next section, I will discuss evaluation results along: 1) scalability to high volume and high velocity streams, and 2) quality of the detected events.

3.4 Evaluation

There are two specific challenges in evaluating events from Twitter: 1) Unlike traditional media such as broadcast news, where every event is reported, on Twitter there is less likelihood of finding tweets related to minor events, and 2) While traditional research on event detection has relied upon the availability of labeled corpora such as the TDT corpus for evaluation, no such corpus exists for Twitter. Due to the lack of ground truth the exact number or nature of the events is not easily available and manual labeling of a large Twitter dataset is expensive. Twitter streams can be collected in two forms: **topic streams** containing tweets related to a specific topic, where the number and type of events can be verified using external sources, and **random streams**, which contain randomly sampled tweets, where the number and type of events must be manually determined. In this section, I evaluate the proposed approach on both types of streams.

3.4.1 Detecting Events in Topic Streams

As a representative topic stream, I now introduce the *Earthquake* topic stream which consists of tweets related to earthquakes around the world. Due to the existing research demonstrating the use of Twitter during earthquakes (Sakaki *et al.* (2010); Mendoza *et al.* (2010)), we collected tweets referring to earthquakes between

June, 2011 to May, 2012 by monitoring the hashtags: #earthquake, #terremoto, and #quake. The data comprises of 1,007,417 tweets from 317,564 users.

To identify the real-world events spanned in this dataset, we must find an independent and external source, which can provide the ground truth at a reasonable cost as manual annotation is not practical. Towards this, I selected the major earthquakes in 2011 (Wikipedia (2011)) and 2012 (Wikipedia (2012)) on Wikipedia as the ground truth of the events in the dataset. These reports were manually compiled from several major news sources. In this work, I focus my effort on the days when a major earthquake resulted in at least 10 casualties, which are summarized in Table 3.1. For most events in 2011, only a few hundred tweets were collected which might be due to the popularity of regional hashtags. Therefore, I set $H_t = 5$ for this dataset.

Table 3.1: Major Earthquakes in 2011 and 2012

Day(UTC)	**Location**	**Magnitude**	**Death Toll**
Jul 19, 2011	Fergana Valley	6.2	14
Sept 5, 2011	Aceh, Indonesia	6.7	12
Sept 18, 2011	India-Nepal border	6.9	111
Oct 23, 2011	Van, Turkey	7.1	684
Nov 9, 2011	Van, Turkey	5.7	40
Feb 6, 2012	Visayas, Philippines	6.7	113
Apr 11, 2012	Aceh, Indonesia	8.6	10
May 20, 2011	Emilia-Romagna, Italy	6.1 & 5.8	27

Evaluating Scalability

To verify that the approach is scalable, I evaluated the rate at which the tweets in our dataset were generated and the time required by the proposed framework to identify events. Table 3.2 compares the measurements for the Earthquake dataset. Column 4 describes the tweet collection rate and Column 5 describes the rate at which tweets were processed. We find that the proposed approach is capable of handling high volume topic-specific Twitter streams by being able to process the tweets at a rate which is significantly higher than the rate at which tweets were generated.

Table 3.2: Efficiency of Event Detection: Earthquake

Day	**#tweets**	**Total processing time (Min)**	**Collection rate (Tweets/Min)**	**Processing rate (Tweets/Min)**
Jul 19, 2011	880	0.04	0.613	23,498.00
Sept 5, 2011	2,712	0.18	1.88	14,788.69
Sept 18, 2011	465	0.02	0.32	18,699.73
Oct 23, 2011	5,253	0.49	3.65	10,646.89
Nov 9, 2011	2,712	0.17	1.89	15,611.63
Feb 6, 2012	13,586	4.79	13.72	2,834.92
Apr 11, 2012	28,182	10.61	19.57	2656.06
May 20, 2012	20,509	6.40	14.33	3,204.44

Quality of Detected Events

Detected events are typically described using the frequent keywords from event tweets (Yang *et al.* (1998); Petrovic *et al.* (2010); Fung *et al.* (2005)). Therefore, I extracted the top keywords of each event as its description to verify whether they matched the ground truth. In Table 3.3, I present the most representative event corresponding to the events in Table 3.1. I also observed that the proposed approach was able to discover the evolution of events, which are represented by sub-events, which we will revisit later. Another observation which can be made from Table 3.3 is that our system was unable to detect any events on July 19, 2011 and Sept 18, 2011. On further investigation I found that the volume of tweets collected on these days was insufficient to identify an event.

Table 3.3: Events Detected in the Earthquake Dataset

Day	**Location**	**Key Event Terms**
Sept 5, 2011	Indonesia	sumatra, western, indonesian, island, #breakingnews
Oct 23, 2011	Turkey	#turkey, eastern, turkey, magnitude, news
Nov 9, 2011	Turkey	turkey, eastern, magnitude, rocks, usgs
Feb 6, 2012	Philippines	pray, visayas, philippines, struck, earlier
Apr 11, 2012	Indonesia	#indonesia, tsunami, magnitud, indonesia, sacudió
May 20, 2012	Italy	sentito, emilia, sono, cosa, chies

To quantify the effectiveness of our approach in detecting events, I compute the F_1 score which captures both the Precision and Recall. Precision is computed as the number of detected events that match the ground truth including sub-events. Recall is computed as the number of events from the ground truth which were successfully detected. The F_1 score for the Earthquake dataset was 0.77.

3.4.2 Detecting Events in Random Streams

Twitter streams can also be collected without any topic bias using the Twitter Sample API [1] , which returns a 1% random sample of the Twitter stream. The tweets in such streams include interpersonal conversations and discussions of real-world events. The task of event detection is harder in this case due to the presence of noise. To verify that our approach can be successfully applied to random streams, I collected sampled tweets from 11:02 AM on Apr 15 to 9:16 AM on Apr 16, 2013.

[1] https://dev.twitter.com/docs/api/1.1/get/statuses/sample

The data consisted of 4,212,333 tweets from 3,322,379 users. As there is no ground truth for these days, I will test the effectiveness of our framework by verifying that we can detect the top stories of the day. I begin by establishing the scalability of our approach. Here I set $H_t = 6.3$ due to the larger volume.

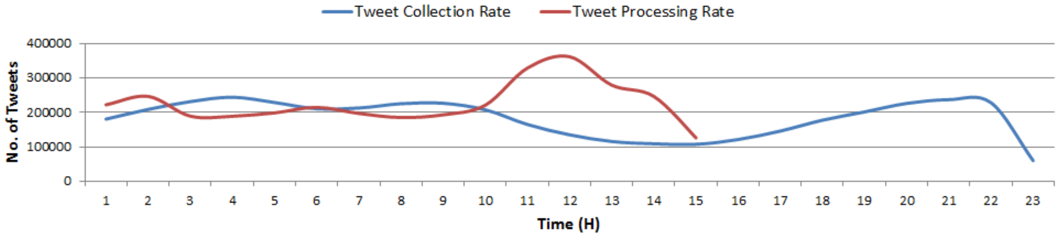


Figure 3.1: A Comparison of the Tweet Collection Rate and the Tweet Processing Rate in the Random Dataset

Evaluating Scalability

As in the case of the topic specific dataset, I test the efficiency of the proposed approach on a random stream by comparing the tweet generation rate and tweet processing rate. A comparison of these measurements is presented in Figure 3.1. The figure clearly shows that the processing speed for a majority of the collection period was on par with the collection speed and it often exceeded the tweet collection rate significantly. Nevertheless, the proposed approach was able to detect events in near-real time. This shows that our approach can be efficiently applied to a random Twitter stream.

Quality of Detected Events

As manually labeling the tweets is not practical, I evaluate the quality of the events based on the coverage of the two major events which occurred during this time period. From the random stream, I detected 167 events. A manual investigation of the events revealed 4 major types of events: events related to the Boston bombing incident, events related to the Presidential elections in Venezuela, events discussing a music festival, and finally events which represented banal Twitter chatter. An example of events expressing banal Twitter chatter included tweets from the fans of Justin Bieber, which resembled the characteristics of an event, but did not refer to a specific event. In Table 3.4, I present examples of the two main types of events: Boston marathon bombing and the Venezuelan Presidential elections. The first event discusses the controversy surrounding the counting of votes in the Presidential elections in Venezuela held on Apr 14, 2013 (Wikipedia (2013)). The other two events are related to the Boston marathon bombing incident. The first event contains reports of the bombing itself and the second event references the reactions of the Twitter users. The results show that the approach can detect reasonable events in the presence of large amount of noise.

Table 3.4: Events Detected in the Random Stream

Event	**Top Keywords**
Venezuelan Presidential elections voting controversy	votos, capriles, esto, #capriles-ganótibisaymintió, fraude
Boston marathon bombing incident	marathon, boston, explosion, finish, line
Support for the bomb victims starts pouring in	boston, marathon, explosion, heart, bombing

3.5 Related Work

Event detection in traditional media is also known as Topic Detection and Tracking (TDT) and a pilot study on this task is presented in Allan *et al.* (1998a). In Yang *et al.* (1998), news articles were modeled as documents to detect topics. The documents were transformed into vector space using the TF-IDF and two clustering approaches were evaluated: Group-Average Agglomerative Clustering (GAAC) for retrospective event detection, and Incremental Clustering for new event detection. The authors concluded that the task of new event detection was harder. In Allan *et al.* (1998b), the authors focused on online event detection. The authors approached the problem as a document-query matching problem. A query was constructed using the k most frequent words in a story. If a new document did not trigger existing queries then it was considered to be part of a new event. In Fung *et al.* (2005), the authors addressed the problem of detecting *hot* bursty events. They introduced a new parameter-free clustering approach called feature-pivot clustering, which attempted to detect and cluster bursty features to detect hot stories.

An attempt to detect earthquakes using Twitter users as social sensors was carried out by in Sakaki *et al.* (2010). The temporal aspect of an event was modeled as an exponential distribution, and the probability of the event was determined based on the likelihood of each sensor being incorrect. Becker *et al.* (2010) tackled event detection in Flickr. The authors leveraged the meta data of images to create both textual and non-textual features and proposed the use of individual distance measures for each feature. These features were used to create independent partitions of the data and finally the partitions were combined using a weighted ensemble scheme to detect event clusters. In Weng and Lee (2011), the authors constructed word signals using the Wavelet Transformation and used a modularity-based graph partitioning approach on the correlation matrix to get event clusters. Li *et al.* (2012) identified bursty segments in tweets and clustered the segments to identify events.

Few existing approaches are designed for streaming Twitter data and even fewer are scalable to real-time streams. In Sayyadi *et al.* (2009), the authors converted a stream of blog posts into a keyword graph, where nodes represented words and links represented co-occurrence. Community detection methods were applied on the graph to detect communities of related words or events. In Zhao *et al.* (2007), the authors model the social text streams including blogs and emails as a multi-graph and cluster the streams using textual, temporal, and social information to detect events. A hybrid network and content based clustering approach was employed in Aggarwal and Subbian (2012) to identify a fixed number of events in a labeled Twitter stream containing tweets from two events. Generally, the number of events is not known beforehand and obtaining the user network adds significant overhead, thus adding to the complexity of this method. In Petrovic *et al.* (2010), the authors recognized the need for faster approaches for first story detection in streams. The authors proposed a two-step process to identify first stories in streaming data. First, the nearest neighbor of each tweet is identified using locally sensitive hashing in constant time and space. Second, a clustering approach called Threading is applied to group related tweets into event clusters. The first tweet in a thread is presented as the first story and the thread itself is considered an event.

3.6 Discussion

While few approaches exist to capture events in a streaming environment, the *Threading* technique proposed in Petrovic *et al.* (2010) is the closest. Using the configuration recommended by the authors, I applied this technique to the Earthquake dataset. First, I compare the scalability of the two approaches. The results for this experiment are presented in Table 3.5. A comparison against our approach in Table 3.2 shows that our approach can process and detect events faster. Next, I evaluate the quality of the events. On all days, the Threading approach detected a greater number of events. Even using the ranking strategy proposed by the authors to retrieve the top 10 fastest growing events, I found that the F_1 score for the Threading

Table 3.5: Efficiency of Threading Technique: Earthquake

Day	**#tweets**	**Processing Time (Min)**	**Collection rate (Tweets/Min)**	**Processing rate (Tweets/Min)**
Jul 19, 2011	880	1.11	0.613	793.40
Sept 5, 2011	2,712	3.99	1.88	678.68
Sept 18, 2011	465	0.88	0.32	527.10
Oct 23, 2011	5,253	2.65	3.65	1,984.97
Nov 9, 2011	2,712	2.54	1.89	1,068.13
Feb 6, 2012	13,586	38.36	13.72	354.19
Apr 11, 2012	28,182	135.27	19.57	208.34
May 20, 2012	20,509	210.32	14.33	97.51

technique was 0.64 compared to 0.77 for the proposed approach. As the Earthquake dataset was the smallest among our datasets, the results show that the framework outperforms this approach. The proposed approach also successfully removed noisy tweets.

There are two additional advantages to using the framework to detect events over existing approaches. Firstly, the approach can detect the evolution of events in dynamic Twitter streams. The inclusion of a *temporal model* allows us to identify sub-events within a larger event. For example, the approach can not only detect that an earthquake has occurred, but also identify the topics that emerge as a result of the earthquake, such as damage reports as seen in Table 3.6, where I present 5 events from the tweets generated during the Indonesian earthquake on April 11, 2012. Secondly, we can directly apply the approach on a stream of multilingual tweets, which is essential due to the global popularity of Twitter.

Table 3.6: Evolution of the Crisis Event on April 11, 2012

Event	**Top Keywords**
Earthquake strikes Indonesia. Tsunami alert is issued	tsunami, #indonesia, #sumatra, scossa, allarme
Tremors felt in India	felt, singapore, thailand, indonesia, #tremors
Tsunami alert in Indian Ocean	tsunami, indian, ocean, move, alert
Sea water receding near the epicenter	aceh, quake, water, receding, island
Reports emerge that a tsunami is less likely	#indonesia, tsunami, moved, horizontally, vertically

3.7 Conclusion

In this chapter, I introduced a method to capture events in evolving Twitter streams. The proposed approach incorporates a temporal model to handle the evolution of topics and tackles the various challenges of working with Twitter streams including informality of text through the use of a suitable distance measure for comparisons. Through experiments on two forms of real-world Twitter streams: topic-specific and random, I showed that the approach is practical and is capable of detecting real-world events.

Chapter 4

IDENTIFYING RELEVANT USERS TO FOLLOW DURING A CRISIS

4.1 Introduction

Historically, in covering events with a large impact such as the Arab Spring movement, traditional media such as television and printed news provide a manicured view of the story to their audience backed with vetted, credible resources. While these media often provide a filtered (or edited) view of the story, the overhead incurred in the process results in a slower flow of information. The pervasive use of social media due to the low barrier to publication allows anyone to publish information at any time, making the details of an event instantly available. Instead of providing some edited, exclusive views of an event, social media provides not only timely information in the critical minutes and hours as an event develops, but also many different or inclusive views of the event. Meanwhile, social media generates mountains of data, at times mixed with noise.

In the context of noisy data, *how can we get fast access to relevant and useful information in social media during these events?* An inclusive approach to finding relevant information from messages is to identify relevant people in social media who are more likely to be the sources publishing useful information (*Information Leaders*) for dynamic events. In general, for a global-scale event, social media users can be naturally categorized into local users who witness the unfolding event and remote users who are connected via social media. Local users have first-hand experience, publishing specifics about the event. To answer this question, I seek to develop an effective way of solving the following problem.

Table 4.1: Parameters Used to Collect the Tweets

Country	Keywords/Hashtags	Geographic Boundary
Egypt	#egypt, #muslimbrotherhood, #tahrir, #mubarak, #cairo, #jan25, #july8, #scaf, #noscaf	(22.1,24.8),(31.2,34.0)
Tunisia	#tunisia,#tunisian,#tunez	(30.9, 9.1),(37.0,11.3)
Syria	#syria, #assad, #aleppovolcano, #alawite, #homs	(32.8,35.9),(37.3,42.3)
Libya	#libya, #gaddafi, #benghazi, #brega, #misrata, #nalut, #nafusa, #rhaibat	(23.4,10.0),(33.0,25.0)
Yemen	#yemen, #sanaa, #lbb, #taiz, #aden, #saleh, #hodeidah, #abyan, #zanjibar, #arhab	(12.9,42.9),(19.0,52.2)

Problem Statement. Given a social media site, and an event E, let C be the content associated with E and U be a set of corresponding users; find "information leaders" $S \subset U$ such that by following S, one can effectively obtain information about E.

Due to its effectiveness in recent studies (Mendoza *et al.* (2010); Gao *et al.* (2011)) and its rapid information dissemination capabilities (Sakaki *et al.* (2010)), Twitter is selected as the social media site under study. The content C is therefore represented using tweets (hereafter referred to as T) and the event in our case is the Arab Spring revolutions.

4.2 Data Collection

I systematically collected tweets from various countries within and outside the Middle East, which were related to the Arab Spring movement. This process involved the usage of certain variables, namely: keywords, hashtags, and geographic regions. I collected 12.9 million tweets which were generated about or from the countries: Egypt, Libya, Syria, Tunisia, and Yemen. The tweets were collected using TweetTracker over

Table 4.2: Characteristics of the Arab Spring Dataset

	Egypt	Tunisia	Syria	Yemen	Libya
#users	514,272	19,094	146,996	43,512	375,924
#tweets	6,184,346	86,437	2,916,449	381,386	3,418,485
#geolocated tweets	84,899	5,229	16,575	849	17,814
#retweets	2,821,864	31,392	1,253,551	142,103	1,919,540

the course of 7 months starting from February 1, 2011 to August 31, 2011. A full list of the variables used is presented in Table 4.1. Column 2 in the table contains the keywords and/or hashtags used. Column 3 contains the geographic boundary box surrounding each country used to crawl all the geolocated tweets from the region. The box is specified as the SW corner (longitude, latitude) of the geographic box followed by the NE corner (longitude, latitude) of the box, separated by a comma. More information on the characteristics of the collected data are presented in Table 4.2.

4.3 Data Preprocessing

The Arab Spring movement was not an isolated incident pertaining to a single country. The movement began and subsequently spread across several countries in the Middle East with prominent populations of Arabic, and English speakers. This mixture of language requires special care with respect to processing. As a result, the methods I chose to process the data are not specific to a language. In the preprocessing step I removed stop words (using a comprehensive list of stop words from the English and Arabic languages) and Twitter artifacts from the text such as hashtags, user mentions, and URLs. Next, I attempted to stem the words using three stemmers: the Arabic stemmer created by Larkey and Connell (2006), the Arabic stemmer provided with Apache Lucene [1] , and the Tashaphyne stemmer [2] . All three of the aforementioned stemmers produced inconsistent output that could not be understood by native Arabic speakers, making it impossible to know if their results were correct. Therefore, to remain consistent, I eliminated stemming from our preprocessing treatment for all languages.

Next, I will introduce the approach to identifying information leaders, or users to follow during an event.

4.4 Geo-Topical User Identification

Social media sites now have millions of users and information travels easily and quickly through this medium. Due to noise and credibility concerns, it is not sufficient to simply pick users who produce more information. Tracking all users is also not a viable option to acquire information. To identify a subset of the users who are likely to publish useful information on a crisis we need a more effective strategy. Two factors play an important role in a crisis: 1) the *topic of discussion* which relates the user to the event, and 2) the *location of the users* which is important to establish the credibility of the content being published by the user. Every user who has tweeted on a topic can be associated with each of these dimensions with a specific score that represents his relevance along that particular dimension. Below, I discuss the procedure to compute these scores and also explain the significance of scoring well along a particular dimension. Our first step is to identify the topics of discussion in the tweets.

[1] `http://lucene.apache.org/`

[2] `http://pypi.python.org/pypi/Tashaphyne/`

Table 4.3: Sample of Words from a Subset of Topics in Tunisia with Justification for Their Selection.

Topic Keywords	Selected	Reason
`forget, tonight, ..., proud, site`	No	Disagreeing
`police, protest, ..., situation, shot`	Yes	Agreeing

4.4.1 Topic of Discussion

Tweets can be considered as small documents of length at most 140 characters. The topic of discussion of the tweets can be manually labeled as one of several topics of discussion or factors that initiate these discussions. In the context of Arab Spring, these factors may include economic factors, torture and brutality, protest, etc. Alternatively, an automated approach of topic detection in documents is the Latent Dirichlet Allocation (LDA) (Blei *et al.* (2003)).

I used the Gibbs sampler LDA [3] to discover topics related to the Arab Spring movement. To tune the hyper-parameters on the Dirichlet priors (α, β) and the number of topics N, I performed several iterations of LDA using the Tunisia dataset and did manual inspection to determine parameter values which performed the best. I varied $\alpha = 0.1$ to 1.0 in intervals of 0.1, and $N = 10$ to 100 in intervals of 10 for a total of 100 iterations. Then, I manually identified the parameter values which resulted in the most relevant topics. As criteria, I looked to the coherency of the words in a topic to make up what can be viewed as a theme, regardless of the content. Using the values of α and N, the parameter β is tuned. To do this, I iterated $\beta = 0.1$ to 1.0 in intervals of 0.1 with $N = 40$ and $\alpha = 0.4$. After analyzing the results, I found that the best results resided between 0.1 and 0.2. Iterating between 0.1 and 0.2 at an

[3] `http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm`

interval of 0.01, I found that the best value for β was 0.11. However, some generated topics were not coherent. In the next section, I discuss how the irrelevant topics were trimmed, to ensure that all topics investigated present a coherent idea.

Topic Pruning: Upon inspection of the topics produced by LDA, I realized that many topics were unfit for further inspection, i.e., they contained unrelated keywords, or sets of keywords that did not describe a distinct topic. To remove the unrelated topics, I (along with native Arabic speakers), manually went through the topics and eliminated those that were not related to the event of that country. In Table 4.3, an example of an English topic for the events that were deemed appropriate for our studies and ones that were not is presented. After careful pruning, the following number of topics remained for each country: Egypt - 11, Libya - 23, Syria - 17, Tunisia - 14, Yemen - 21. Using the final set of topics, user relevancy can be identified through a topic affinity score.

Topic Affinity Score

Let S be the set of words that define the topic. These words are the top 25 most probable words for the topic, as determined by LDA, i.e., $|S| = 25$. Let $\mathcal{T}$ be the collection of a user's tweets. Let $T \in \mathcal{T}$ be a user's tweet, i.e., a set of words. A user's topic affinity score is defined in Equation 4.1.

$$\mathit{topic\_score}(S, \mathcal{T}) = \frac{\sum_{T \in \mathcal{T}} \mathit{sgn}(|S \cap T|)}{|\mathcal{T}|}, \tag{4.1}$$

where, *sgn* represents the sign function. Using this formulation we see that a user's topic affinity score is in the interval $[0, 1]$. Score value 0 indicates that they never tweeted in the topic and a score of 1 indicates that all of the tweets overlapped with the topic.

4.4.2 Location of the User

During a crisis, the location of the user is an important factor which can help determine which user is likely to publish information relevant to the crisis. For example, in an earthquake, tweets coming from a location closer to the earthquake are likely to be more pertinent to the crisis than tweets from outside the location. In the case of the Arab Spring, tweets coming from within the country are more likely to contain relevant information than those from outside the respective countries. To identify a user's relevancy to the event based on his location, I propose the *geo-relevancy score*.

Geo-Relevancy Score A user's location can be determined using the location from his tweets. The location of a tweet can be determined in one of two ways:

1. **Geolocated Tweet** - A tweet that has been located through the GPS sensor on a mobile device, or through IP location capabilities of the browser. This information is metadata that the individual tweeting chooses to share when publishing the tweet.
2. **Profile-located Tweet** - A tweet whose location data is obtained by analysis of the user's profile. Users can provide geographic location information in their profile, and I analyze this by geolocating it through the OpenStreetMaps [4] .

Using the location information from the user's tweets his geo-relevancy score is a value in the interval $[0, 1]$, calculated as follows:

1. If the user never produced a geolocated tweet, then his geo-relevancy score is the average number of his tweets that were profile-located to be within the crisis region. A user is represented as a tweet location vector $tweet\_loc \in \mathbb{R}^T$, where T is the number of tweets published by the user. $tweet\_loc_i = 1$, indicates that the user's profile information at the time of the ith given tweet resolves to within

[4] http://nominatim.openstreetmap.org/

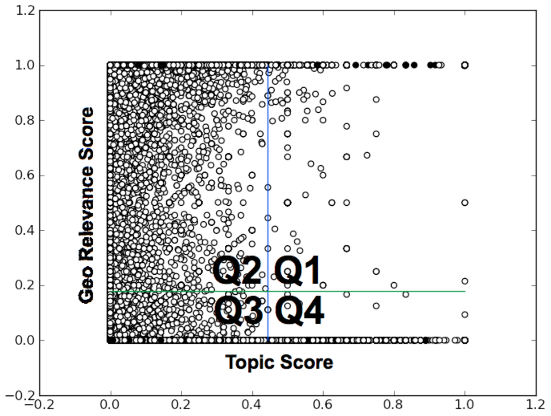


Figure 4.1: User Visualization of Geo-Relevancy and Topic Affinity for a Topic in Egypt.

the crisis region and a $tweet\_loc_i = 0$ indicates that the user was outside or that the location information was missing. Then, we can compute the geo-relevancy score as:

$$geo\_rel\_score(tweet\_loc) = \frac{||tweet\_loc||_0}{T}, \tag{4.2}$$

where $|| \cdot ||_0$ denotes the zero-norm.

2. If a user is geolocated and their location is within the crisis region, then his geo-relevancy score is 1.
3. Conversely, if a user produces a geolocated tweet that is not within the crisis region, then their geo-relevancy score is set to 0 as they have demonstrated that they are not within the location and do not have access to the temporally-sensitive information as someone experiencing the event firsthand.

It should be noted that a user may have a different topic affinity score for each topic in the revolution, but the same geo-relevancy score across the topics.

4.4.3 Visualizing Users in Two Dimensions

After obtaining the geo-relevancy score and topic score for each user in every topic, I create a scatter plot to see how users are related to each other. An example of one such plot is shown in Figure 4.1. In this plot, each dot is a user. The black dots are the users who received their score through geolocation (rules 2 and 3 of the previous section). The white dots are users who received their geo-relevancy score from resolving their profile information (rule 1 in the previous Section). The x-axis represents the user's topic affinity, and the y-axis represents the user's geo-relevancy score. The vertical and horizontal bars represent the averages for the distance and topic scores, respectively. In Figure 4.1 we can see that, based on the location of these average bars, the plot breaks down into four quadrants.

4.4.4 Understanding Users with the Quadrants

By laying out the quadrants as above, we observe that each quadrant has certain unique characteristics. Using the same numbering system as the Cartesian coordinate system, the following quadrants can be defined:

Quadrant I (Q1): This quadrant contains users with both topic and geo-relevancy scores above the average. This quadrant contains users who are both on the ground and actively discussing the topic at hand. These users are "*Eyewitness*" users.

Quadrant II (Q2): This quadrant contains users whose topic score is below average, but their location score is higher than average. These people are in the vicinity of the revolution, but not discussing the topic. These users are "*Topic Ignorant.*"

Quadrant III (Q3): This quadrant contains users with topic and geo-relevancy scores below the average. These users are "*Apathetic*", as they are neither within the region nor discussing the topic at hand.

Quadrant IV (Q4): This quadrant contains users with topic scores above the average, but geo-relevancy scores below. These users are outside of the country, but are still producing information relevant to the event. These users are "*Sympathizers*".

Users in Q1 can be considered the most relevant to the crisis, as they have high scores across both the dimensions. Users in both Q1 and Q4 are considered "*topic-aware*" as they have a better-than-average discussion rate on the given topic. These are users who have spent a lot of time talking about topics relevant to the Arab Spring. Hence, I propose to study the tweet characteristics of the users in Q1 and Q4. This study would clarify the utility of following Q1 for the purpose of obtaining information about an event.

4.5 Evaluation

In this section, I will show that users in Q1 generate higher quantity of information and later, I will evaluate the quality of the information generated by them.

4.5.1 Information Quantity from Q1 Users

To evaluate the quantity of information generated by Q1 users, one can measure the number of tweets published by Q1 users from each country. To show that these users produce more information and the quantity is statistically significant, I need to compare quantities produced with a set of representative users from within our dataset. Uniform sampling provides theoretical guarantees on generating accurate representative datasets. Hence, I uniformly sample an equal number of users from the dataset as contained in Q1 and consider it as a representative set. To avoid any sampling bias in the results of comparison, I generate 100 such sets of randomly selected users U_{Rand} and take the average of the number of tweets generated by them

Table 4.4: Comparison of the Quantity of Tweets Generated by Q1 Users and a set of Random Users U_{Rand}. All Results are Extremely Statistically Significant with p-value < 0.0001

	Tunisia		Egypt		Syria		Yemen		Libya	
	Q1	U_{Rand}	Q1	U_{Rand}	Q1	U_{Rand}	Q1	U_{Rand}	Q1	U_{Rand}
Feb	1,817	3,706	74,956	138,379	228	20	1,247	211	1,101	926
Mar	805	1,521	84,856	34,346	199	113	2,546	345	1,971	668
Apr	1,006	2,062	137,562	48,472	12,271	4,817	4,480	599	2,840	319
May	144	234	22,335	7,939	47,496	2,347	639	40	2,419	168
Jun	12	5	29,569	11,610	40,458	2,514	1,568	161	2,068	187
Jul	296	364	274,446	89,348	113,069	5,550	4,666	444	4,488	111
Aug	2,081	1,716	232,288	67,608	79,428	10,018	3,920	224	4,624	961

to the number of tweets generated by Q1. A comparison of the tweets generated by Q1 and U_{Rand} is presented in Table 4.4. Looking at the first two columns for each country, it is clear that the Q1 users generally tweet more than U_{Rand}. In cases such as Syria, I found that Q1 users comprised around 0.006% of the users and yet contributed more than 10% of all the tweets for the region. To show that the observed difference is also statistically significant I employ the χ^2 test. The null hypothesis is:

H_0: Q1 users and randomly selected users U_{Rand} generate similar numbers of tweets during a crisis.

Given 7 months of data, I ran the χ^2 test with 6 degrees of freedom and a significance level of $\alpha = 0.05$. As observed from Table 4.4, the null hypothesis for all the countries. The results of the χ^2 test show that the difference between the rate at which tweets were generated by Q1 and U_{Rand} is statistically significant.

4.5.2 Information Quality of Q1

The previous section established that Q1 users generate a significant amount of information. Here, I compare the quality of the information generated by Q1 users.

Meaningful Patterns: In this section, I will show that Q1 users generate information that captures the current trends in the region. Consequent to my methodology, these users are well placed to generate firsthand accounts, as they are in the crisis region and have access to information that most others outside the region do not. By meaningful information here, I mean information that does not correlate highly with the information an average *random* user concerned about the event would publish. Our assumption while performing this experiment is that information leaders should (1) post more often about the specific events *when these events exist* and at other times, (2) post information that is closer to the general discussion about the crisis. In this experiment, I compare the content from the tweets of Q1 users with the (i) general topic of discussion and to (ii) those of a randomly selected set of users. This experiment is performed over M days spanned by our dataset. Here $M = 212$. The topic of discussion for any set of users U is defined as a collection of the top 35 most popular keywords occurring in the tweets of U. Here Q1 users are the union of all Q1 users across all topics for a country, i.e., $Q1 = \cup_{j=1}^{n} Q1_j$, where $Q1_j$ is Q1 users for topic j of a specific country and n is the number of topics in that country.

- Let U_{Rand} represent a random set of users selected from the dataset. T_{Rand} represents the topic of U_{Rand}, T_{Q1} represents the topic of Q1 users and $T_{general}$ represents the general topic of discussion among all users.
- For day i, where $1 \leq i \leq M$, I compute the distance d_i between T_{Q1} and $T_{general}$ using Jaccard distance,

$$d_i = \frac{T_{Q1}^i \cap T_{general}^i}{T_{Q1}^i \cup T_{general}^i}. \tag{4.3}$$

Table 4.5: The Divergence of Q1 Users from the General Topic

	Egypt	Libya	Syria	Tunisia	Yemen
$\lvert Q1 \rvert$	19565	337	946	654	202
Position	2	1100	3307	1	3751

Then, daily distances can be represented using a vector, $D = (d_1, d_2, \ldots, d_M) \in \mathbb{R}^M$. We can generalize the distance to vector format using any vector norm. Here, I use the L_1-norm,

$$d(T_{Q1}, T_{general}) = ||D||_1 = \sum_{i=1}^{M} d_i. \tag{4.4}$$

$d(T_{Rand}, T_{general})$ can be calculated similarly. To remove random bias, I generated 5,000 random user sets $\{U^i_{Rand}\}_{i=1}^{5,000}$'s and their corresponding topics $\{T^i_{Rand}\}_{i=1}^{5,000}$'s.

The distance between $T_{general}$ and all 5,000 randomly generated topics $\{T^i_{Rand}\}_{i=1}^{5,000}$, i.e., $d(T_{general}, T^i_{Rand})$ was computed. Similarly, the distance between general topic $T_{general}$ distance to T_{Q1}, i.e., $d(T_{Q1}, T_{general})$ was computed. Now we can compare the 5,001 distances to the general topic (5,000 distances from random topics + 1 from Q1 users). The list of 5,001 distances is then sorted in ascending order. The first element in this list is the farthest away from the general topic of discussion, and the 5001st is the closest. The ranking of Q1 users is presented in Table 4.5. Q1 users deviated from the general topic of discussion in Egypt and Tunisia. In Syria and Yemen users in Q1 were closer to the general topic of discussion.

4.5.3 Unique Attributes of Q1 Users

It is important to distinguish the users found by the method from influential users found using other methods. These influential people are expected to generate

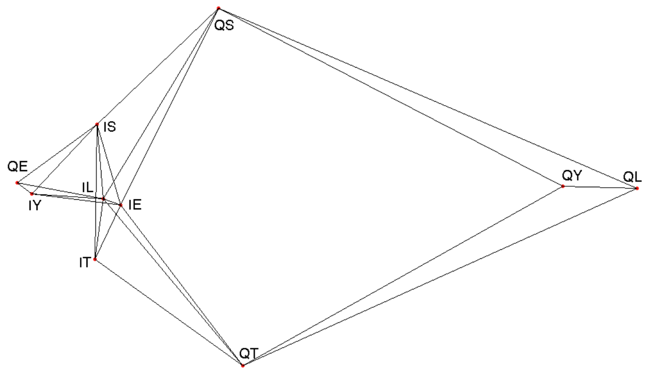


Figure 4.2: Group embedding (using Isomap) of influentials and Q1 users. The probability distribution is the frequency of top-words and the distance is computed using the square root of JS divergence. In this figure, each point is labeled by a two-character code. The first character is either 'I' or 'Q', indicating a Influentials or Q1 group, respectively. The second character is the first letter of the group's representative country. For example, QE represents the Q1 group in Egypt, and IY represents the Influentials group in Yemen.

crisis-relevant information. In this experiment, I show that Q1 users generate more information than influentials, later I will show that this information is also more focused compared to the influentials. To measure influence in a directed network such as Twitter we can consider the number of followers a user has accrued, although techniques, such as PageRank could also be employed.

To conduct this experiment, I first identify the number of tweets generated by users from Q1 and the Influentials from each country in each month m, spanned by our dataset. Our results are presented in Table 4.6. We can observe that the amount of information generated by Q1 is much higher than Influentials. To see if the difference is statistically significant, we run the χ^2 test on the results and find that the difference between the quantity of tweets generated by the two types of users is statistically significant in all cases.

Table 4.6: Evaluation of Tweet Quantity by Q1 and Followers. All Results are Statistically Significant with p-value < 0.0001

	Tunisia		Egypt		Syria		Yemen		Libya	
	Follow	Q1	Follow	Q1	Follow	Q1	Follow	Q1	Follow	Q1
Feb	1,117	1,817	204,023	74,956	48	228	237	1,247	1,720	1,101
Mar	570	805	41,275	84,856	135	199	364	2,546	1,670	1,971
Apr	664	1,006	46,187	137,562	4,967	12,271	497	4,480	616	2,840
May	105	144	9,310	22,335	5,793	47,496	148	639	270	2,419
Jun	5	12	14,928	29,569	5,265	40,458	135	1,568	170	2,068
Jul	152	296	122,505	274,446	12,036	113,069	461	4,666	336	4,488
Aug	855	2,081	67,525	232,288	12,316	79,428	240	3,920	1,603	4,624

To further investigate the uniqueness of Q1 users, the underlying word frequency probability distribution of their most-used words may be compared with that of the Influentials. Jensen-Shannon (JS) divergence (Lin (1991)) is a suitable measure to make this comparison,

$$JS(P||Q) = \frac{1}{2}[D(P||M) + D(Q||M)], \tag{4.5}$$

where $M = \frac{1}{2}(P + Q)$, and D is the Kullback-Leibler (KL) divergence (Cover *et al.* (1991)),

$$D(P||Q) = \sum_{i=1}^{|P|} P_i \cdot log(\frac{P_i}{Q_i}). \tag{4.6}$$

Here, P and Q are the normalized occurrences of the top 500 words used by each of the 10 groups ((Influentials + Q1s) × 5 countries). Using the JS divergence on can create a distance matrix between the 10 groups. From the distance matrix, an embedding of the groups can be generated based on embedding techniques. The embedding would demonstrate how different groups are situated with respect to one another in a 2-dimensional space. When seeking an embedding of the matrix, it is desirable to have a distance metric since distances will be comparable (due to triangle inequality). It has been proven that the square root of the JS divergence is a metric (Endres and Schindelin (2003)); therefore this is used instead.

For this work, I investigated classical embedding techniques such as the classical PCA or Multi-Dimensional Scaling (MDS), and decided to employ the more robust Isomap technique (Tenenbaum *et al.* (2000)), capable of extracting non-linear relationships using geodesic distances between points. The resulting 2-dimensional embedding can be seen in Figure 4.2. This figure shows that different Q1s are at a distance from each other surrounding a dense group of Influentials indicating the generality of the discussion of Influentials across countries.

4.6 Related Work

The related work to our research falls into three intertwined areas: topic models, event detection, and finally Twitter analysis under events, especially disasters.

Topic models have been studied extensively in short-messaging environments. Kireyev *et al.* (2009) analyzed tweets related to crises using topic models. Their approach employs topical clustering and their new technique, dynamic corpus refinement. They tune the term weights in order to get more accurate topic distributions and they also refine their corpus based on the initial topic distribution in order to get datasets that are more related to the disaster under study. Ramage *et al.* (2010) present a partially supervised learning model, called the Labeled LDA, that maps tweets into dimensions. They argue that these dimensions correspond to substance (topics about events, ideas, things, and people), social characteristics (social topics), style (broader trends), and status (personal tweets). Their models take into account both users and tweets. Besides their latent dimensions in Twitter that can help identify broad trends, in order to identify smaller ones, several classes of tweet- specific labels were applied to tweet subsets.In another effort (Hong and Davison (2010)), the authors attempt to adapt the standard topic model system to the microblogging environments. Their results show that models trained on aggregated messages result in higher performance in real-world scenarios. It is interesting to know how topics found by these models change from microblogs compared to the ones found in traditional media. This has been done in Zhao *et al.* (2011) where they compare Twitter based topic models and the ones found in traditional media. They found interesting differences in Twitter. For instance, Twitter acts as an invaluable source for "entity-oriented" topics. These are topics that have low-coverage in other sources of media. Another finding was that though Twitter users had low interest in international news, they actively engaged in helping spread important news.

Topic and event detection has also been an active area of research. In Cataldi *et al.* (2010), the authors introduce both a topic detection and a real-time topic discovery technique. The topics are described as a set of terms. Terms have a life cycle and a term or set of terms is considered emergent if its frequency increases in a specified time interval and was relatively rare in the past. They also weight content based on the PageRank of the authors and introduce a topic graph where users can identify semantically related emergent topics and keywords. In Popescu and Pennacchiotti (2010) the authors formalize controversial events and propose to solve it using regression methods. Controversial events are ones that provoke public discussions in which audience express opposing opinions. Their feature set includes Twitter-based (linguistic, structural, buzziness, sentiment, and controversy) and External features (News Buzz and Web News Controversy). Various systems have also been developed to monitor tweets and events on Twitter (Kumar *et al.* (2011a)). TwitterMonitor (Mathioudakis and Koudas (2010)) is a system that performs trend detection over a stream of tweets. The system detects emerging topics or trends by identifying bursty keywords, and provides meaningful analytics to analyze them.

Twitter, and in general microblogging, has shown to be highly effective when it comes to disaster relief and rapid communication during a natural disaster. Recent studies related to the disasters in Yushu (Qu *et al.* (2011)), Japan (Sakaki *et al.* (2010)), Chile (Mendoza *et al.* (2010)), and Haiti (Gao *et al.* (2011); Barbier *et al.* (2012)).

4.7 Conclusion

Identifying information quickly and efficiently is crucial during crises. In this chapter, an innovative approach to efficiently access information in social media is presented. Using Twitter as an example, I show that a subset of Twitter users *(Information Leaders)* who publish tweets about the event of interest can b identified and they can help provide quick access to relevant information.

Our approach is based on two natural dimensions along which a user can be categorized, namely: topic of discussion and the user's location. Specifically, our contributions are:

- A novel approach to find users who provide quick access to relevant event information.
- Different categories of users who provide different kinds of information. *Generalists* can be used to understand the global impact of a crisis. *Specialists* can be used to get access to information on various topics directly associated to a crisis from within the impact region.
- The method gives all users equal opportunity to be information leaders. In the event of a crisis, most useful information usually comes from people who have personally experienced the impact or have access to such information. These users are not expected to have a large number of followers or play a central role in the Twitter network outside of the crisis.

Through comparison with a reasonable measure of identifying information leaders, it is shown that users selected using our approach produce information in more quantity and with greater quality.

Chapter 5

IDENTIFYING POPULAR USERS THROUGH DIGITAL FIRST IMPRESSION

5.1 Introduction

Social media sites such as Facebook, Twitter, and Tumblr have millions of registered users. Facebook has more than 1.2 billion active users and Tumblr has more than 230 million active blogs [1] . The interactions between users have been used for a variety of applications. Dense connections among users provide low cost, visible, and high impact platform for applications such as advertising and marketing. Studies have shown that social media can be particularly effective for viral marketing (Miller and Lammas (2010)) campaigns where a few users are targeted so they can spread the message to other members of the network. A recent survey shows that the usage of Tumblr [2] is growing rapidly indicating its utility in such campaigns. Microblogging platforms, such as Tumblr and Twitter also play an important role in crises and situations of mass emergency. It has been discovered that new users join such platform during situations of mass emergency (Hughes and Palen (2009a)). Monitoring these events is a challenging problem due to the large volume of information and the large number of users involved. Metrics such as the number of shares/retweets received by a message typically used to identify popular and relevant content may only be computed after sufficient time has elapsed. Early awareness of popular users may help alleviate this problem to aid in crisis monitoring.

[1] http://bit.ly/1nioAZC

[2] http://mklnd.com/RE008J

A challenge these platforms face today is retaining the attention of new users. Studies have shown that only a fraction of all users on a site are regularly active. For example, Google+ now has more than a billion users, but only 359 million of these users are active monthly [3] . And this can be observed in other social media platforms as well due to the myriad choices available to a user. Transforming even a small fraction of these users into active users is in the interest of both the community and the platform. But, the scale of these social media sites makes it impossible to help all inactive users.

Hence, we must select a subset of such users, who can be offered incentives and promotions to help retain their attention on the site. An intuitive method to identify a subset of such users is by identifying users who are likely to be popular in future. However, the lack of information regarding an inactive or a new user's interests is a challenge. Activity information such as posts are unavailable. Typically, only the information provided at the time of registration, which includes the user identity/username is available. In this study, I will investigate the correlation between various weak signals which can extracted from a user's registration information and his popularity. Following are the contributions of this work:

- I introduce the concept of Digital First Impression, which can be used to describe a user when activity information is unavailable.
- I demonstrate that hidden patterns can be extracted from it to successfully predict user popularity. The approach is computationally inexpensive and complementary to existing approaches and works under the constraints of limited information.

[3] `http://bit.ly/1sNmfYX`

5.2 Digital First Impression (DFI)

The key assumption of this work, is that no activity information is available for a user. Therefore, typically the only the information provided by the user at the time of registration is known about them. When a new user registers on a social media site, they are required to create an identity via a unique username. This username is used by other users to identify the user and his actions on the platform. Therefore, this information is guaranteed to be available for every user on the platform and visible to other users when they first contact another user. Often, however additional information may be available about a user. For example, each Tumblr user is associated with a blog and a blogname [4] , and Twitter users provide a description of themselves as part of their profile [5] . A user's *Digital First Impression (DFI)* is defined as the collection of identifying information provided during registration, which is visible to other users on the platform. This information may be considered as the minimal information available about a user's interests, in the absence of any activity information. For example on the Tumblr, a user's DFI would consist of 1) his username and 2) the name of his primary blog. An illustrative example of DFI, can be observed in Figure 5.1, which represents the public profile of BBC News on Tumblr [6] . Here the username is "bbcnews" and the name of the blog is "BBC News".

5.3 Problem Statement

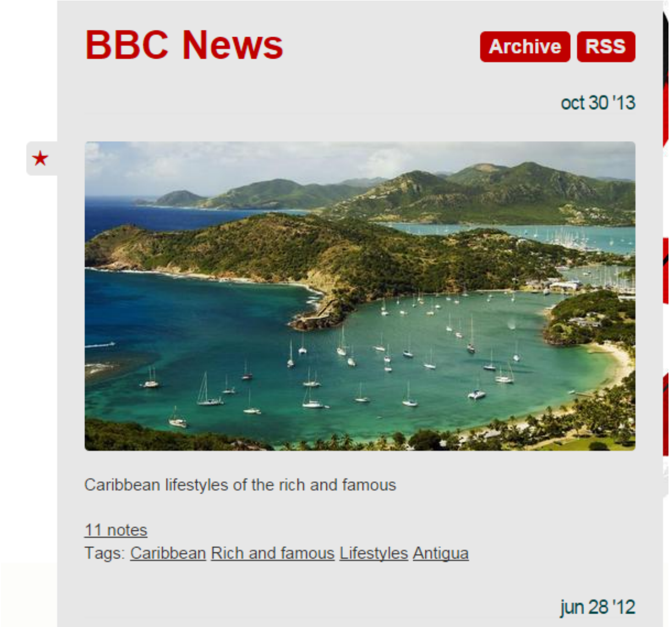
In most social media sites, majority of the attention is received by only a small part of all posts. This can be explained by the Pareto principle [7] , which is applicable

[4] `https://www.tumblr.com/docs/en/blog_management`

[5] `https://dev.twitter.com/overview/api/users`

[6] `http://bbcnews.tumblr.com/`

[7] `http://en.wikipedia.org/wiki/Pareto_principle`



BBC News

oct 30 '13

Caribbean lifestyles of the rich and famous

11 notes

Tags: Caribbean Rich and famous Lifestyles Antigua

jun 28 '12

Figure 5.1: Tumblr Profile of BBC News

to many forms of user behavior observed in social media sites, such as the number of followers of a user, or the number of likes received by a user. In the context of this study, “popularity” is assumed to mean the number of likes or approvals received by user generated content. Therefore, the users in a social media site can be grouped into two classes based on the attention received by their posts on the site. Therefore, the two classes of users are 1) popular class which consists of users who receive a majority of the attention, and 2) not popular class, which consists of a majority of the users.

Given a user u on a social media site with the digital first impression d_u as defined above. We need to decide whether the user is likely to be popular, that is if the user $u \in P$ or the user does not belong to the popular class, i.e., $t \notin P$.

5.4 Related Work

While the task of identifying popular users is not entirely novel, existing solutions to this problem leverage various activity information of the users to tackle it. In this work, this problem is investigated in the context of new and inactive users. As this is a novel problem, the related work is organized along methods to identify popular users in social media and studies on the characteristics of names and their correlation with social and economic factors.

Kunegis *et al.* (2009) investigated the identification of unpopular users in an explicitly signed network, Slashdot. Using Slashdot Zoo, where users explicitly indicate their friends and foes, the authors evaluated different popularity and centrality measures to identify unpopular users or trolls on Slashdot. This problem has also been investigated from the perspective of user influence in a network. In social media, a user's influence in the network is often measured through his reach or the number of connections a user has on a social network.Another interpretation of popular users is authoritative users. Often users whose content is well liked by the community can be considered as authoritative on a topic. Pal and Counts (2011) proposed an unsupervised method to identify authoritative users in microblogs. Using activity based features to capture topical propensity, they grouped users into two clusters, one of which represented authoritative users and used a ranking scheme to identify the top authoritative users.

Alternatively, we can also consider the identification of popular items. Items in social media receive disproportionate attention due to the power-law like distribution of user behavior in social media (Shirky (2003)). Therefore, identifying popular items at the right juncture is critical commercially as differential pricing strategies for content or ad placement can be employed to capitalize on the popularity of the item.

In Lerman and Hogg (2010), the authors studied the predictive ability of a social dynamics model to predict the popularity of news articles on the social news site Digg. The model was shown to perform better than a social influence based strategy. In Bandari *et al.* (2012), the authors investigated the utility of features constructed from the characteristics of an article, such as the subjectivity of an article and the nature of the source of the article in predicting its popularity.

In social psychology, studies have established that the characteristics of a person's name are indicative of a his behavior and his environment. And they can be used to predict user behavior in the real world. Kalist and Lee (2009) demonstrated through a study of people's first names that the popularity of the first name is an indicator of whether an individual is likely to commit crime. The authors observed that the popularity of first names was correlated with the economic backgrounds of individuals. While not confirming the causal relation, the authors verified that the correlation may be used to predict such behavior. Figlio (2005) discovered that individuals whose first-name includes an apostrophe, has a higher scrabble score, or contains several low-frequency consonants, are more likely to come from lower socio-economic backgrounds. In Aura and Hess (2010), the authors investigated the correlation between various "first name features", such as the phonetic features and the number of syllables in the name with the respondents race or financial status. The authors discovered that first-name features could independently predict lifetime outcomes such as income and social status of an individual.

Specifically, these studies show that first names and their characteristics may be indicators of user behavior in the real-world. In the context of social media, an important distinction from the real-world studies is that unlike first-names, usernames are chosen or created by individuals themselves. Thus, it is the first interaction a user has with the platform and may be used to predict user behavior on the platform. Our work is the first to study the relationship between usernames in the online world and the popularity of a user.

5.5 Dataset

Tumblr [8] is a microblogging and social networking website. Each registered user on the platform is associated with a blog and users share content with other users by publishing blogposts. Users demonstrate their approval for other users by "liking" blogposts from other users. Therefore, a straightforward method to measure the popularity of users on the platform is to measure the approval received by a user. In this study, we analyzed all Tumblr public posts published in April 2014. Each post was associated with a user and contained the number of likes received by the post from other users of the community. Popularity of users was measured by aggregating the likes received by a user's posts. Using this information, we created a dataset consisting of two class of users: popular and non-popular. The selection process for identifying popular users is described in the next section.

5.5.1 Identifying the Ground-Truth

As indicated in previous studies, the distribution of likes received by users follows a power-law distribution as observed in Figure 5.2. The Pareto principle suggests that a small fraction (20%) of the users contribute 80% of the observed behavior, such as the proportion of posts generated by the most active users of a community during a protest (Poell and Borra (2012); Shirky (2003)). Following this principle, we rank the users based on the number of likes received by their content and use the rank information to determine the top 20% of the users as popular users and the remaining users as not-popular users. Note that, we do not consider these users as

[8] `http://www.tumblr.com`

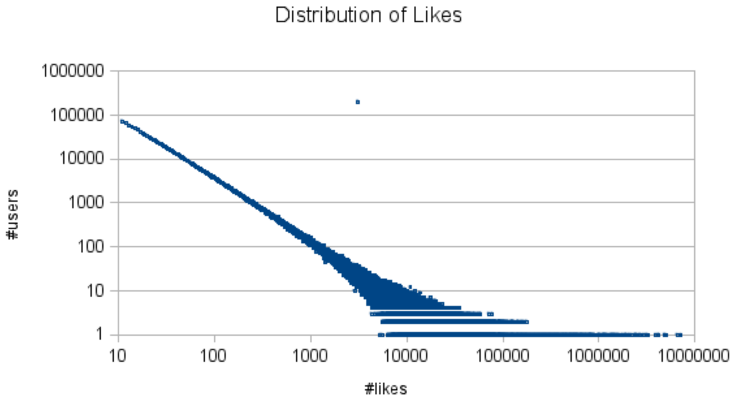


Figure 5.2: Distribution of *likes* in Tumblr blogs in log-log scale. The y-axis represents the number of users and the x-axis represents the number of likes received by the user's blog, which may include more than one post.

Table 5.1: Dataset Characteristics

Property	**Value**
Avg. likes	486.68
Median Likes	8.0
#Users	464,989
#Blogs	464,989
Avg. length of username	13.85
Avg. length of blogname	17.23

unpopular due to the absence of an explicit *dislike* function in Tumblr. From this data, we obtained a 10% stratified sample, where members of each class were sampled randomly, to conduct this study. Table 5.1, introduces some characteristics of this dataset.

5.6 Characteristics of Digital First Impression

Typically, content and social connections are reasonable sources used to determine the popularity of users on a social network. However, in their absence we evaluate the correlation between the weak signals which can be extracted from a user's DFI and his popularity. We present a general framework, which can be used to analyze DFI of a user on any site. From our observations, we will construct features which will be used to build a model for the task.

5.6.1 Structural Characteristics

Character Composition

Social media platforms are informal in nature. In this section, we will look at some patterns in the usage of different classes of characters to study their effect on the perception of users. We analyze both aspects of a user's DFI: blognames and usernames independently.

Special characters: In Figure 5.3, we present the distribution of the number of likes received by users(log-scale) and its variance with the number of special characters in the blogname and the username. The Pearson Correlation Coefficient (ρ) was 0.06 (username) and -0.25 (blogname). We observed a negative correlation between a user's popularity and the number of special characters in blognames. However, the number of special characters did not have a strong correlation with the popularity. Excessive use of special characters is typically indicative of informal content. Special characters including parentheses, and special symbols such as the '\$' symbol.

Numeric Sequences: The length of numeric sequences in the identity is increases the complexity of the name, such characteristics are typically not observed in genuine identities as it increases the cognitive load to remember the user. I discovered that

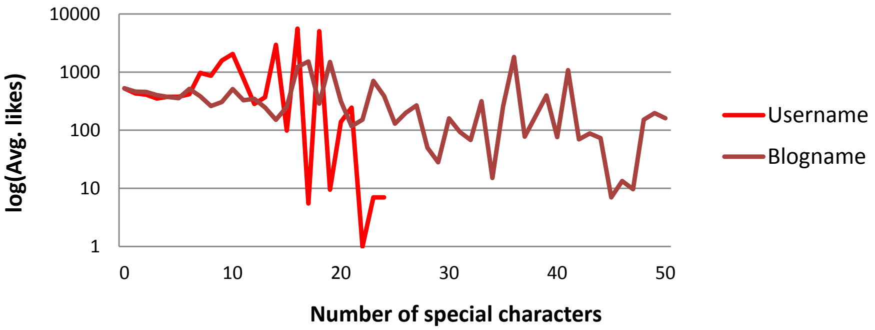


Figure 5.3: Effect of the Number of Special Characters on Popularity

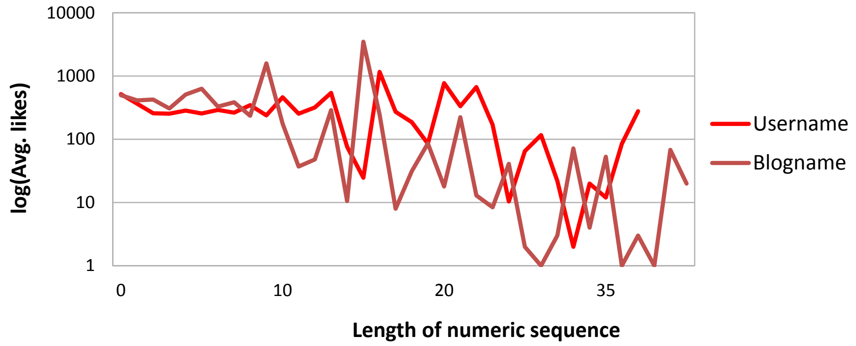


Figure 5.4: Effect of Numeric Sequence Length on Popularity

users with blognames containing numeric sequences greater than 8 characters had significantly lower popularity than other users. Figure 5.4 shows that the results are consistent for both identities of a user. The ρ was -0.32 (username) and -0.31 (blogname) indicating a clear negative correlation.

Other character classes investigated include upper case characters, digits, and emoticons and I observed similar patterns. In addition to the raw counts for these character classes, I also use the ratio of the characters in a class with respect to other content as features in our model.

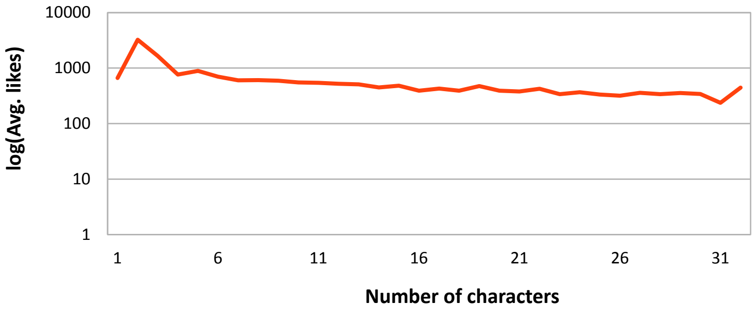


Figure 5.5: Effect of the Length of the Username on Popularity

Length

Figure 5.5 shows a comparison of the number of likes received by a user with number of characters in username and Figure 5.6 shows a similar study for the number of words in a blogname. The ρ was -0.56 (username) and -0.15 (blogname). The results suggest that there is a negative correlation between the two quantities, and this can be observed in both cases. Studies have shown that shorter texts are easier to remember (Baddeley *et al.* (1975)). In fact, memory span was shown to be inversely related to the length of the text. I suspect that this could be a factor in our observations.

Unique Characters: The effect of the number of unique characters on the popularity of a user in Figure 5.7 and it can be observed that the popularity decreases rapidly once the number of unique characters increases beyond 16 for the username and 25 for the blogname. The results were also found to be much stronger in this case with ρ values -0.88 (username) and -0.51 (blogname). In addition to the unique characters, the character distribution of the words in blogname summarized by the mean, max, min, median, and the standard deviation are also used as features in the method.

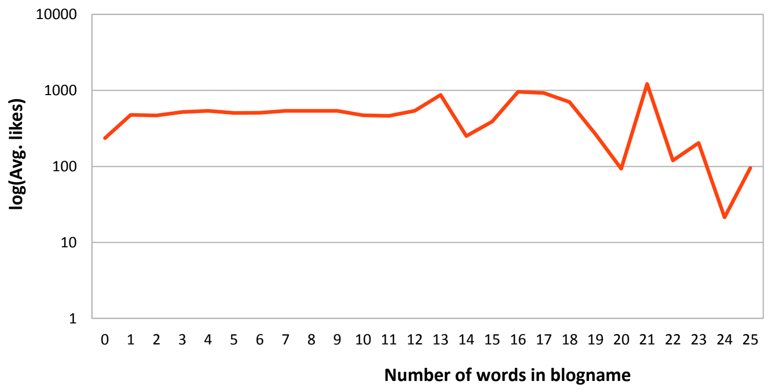


Figure 5.6: Effect of the Length of the Blogname on Popularity

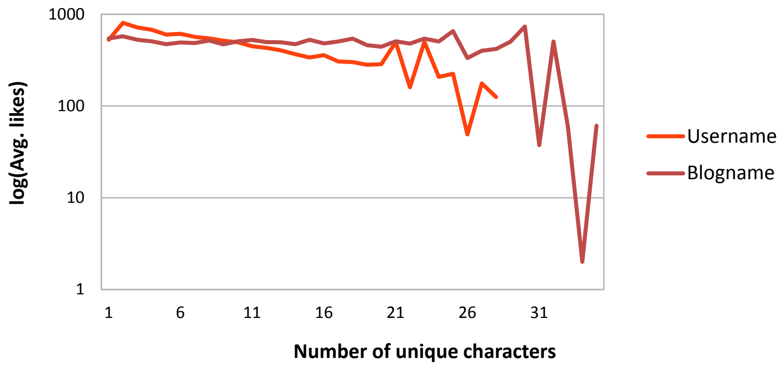
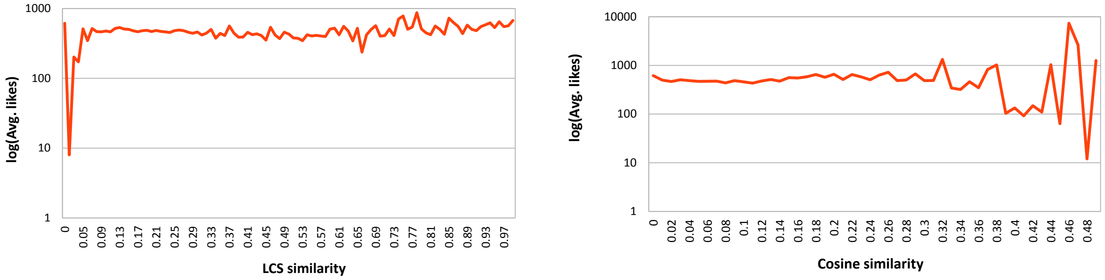


Figure 5.7: Effect of the Number of Unique Characters on Popularity

5.6.2 Relationship Between Identities

Similarity between identities: blogname and username, may indicate a closer association of the blog topic to the user. This is evident in the case of public figures and popular entities in the real world. For example, consider the example of BBC News discussed previously. Both the blogname and the username, in this case, have a high degree of similarity. The effect of similarity can be evaluated in two contexts:

- Cosine similarity is a well-known text similarity measure. It can be used to measure the character-level similarity between the two identities of a user's



(a) LCS similarity

(b) Cosine similarity

Figure 5.8: Effect of Similarity Between Blogname and Username on Popularity

DFI as the dot product between the distribution of two texts

$$Cosine(a, b) = \frac{|a \cdot b|}{|a| \times |b|}. \tag{5.1}$$

- Cosine similarity does not consider the order information. Therefore, we also compute similarity based on the length of the longest common subsequence(LCS). The LCS similarity between texts a and b with the length of the longest common subsequence l is

$$LCSSim(a, b) = \frac{2 * l}{|a| + |b|}. \tag{5.2}$$

In Figure 5.8a, we can observe that the increase in similarity of the longest common subsequence is correlated with the increased popularity of users. The ρ was 0.41. In the case of cosine similarity in Figure 5.8b, the popularity increase is gradual and we observe more variance when cosine similarity was greater than 0.3. The ρ was 0.25. Our observation confirms that there is indeed a positive correlation between the similarity of a user's identities and the likelihood of becoming popular. This shows that blogs with stronger ties to the user identity are more likely to enjoy increased popularity among its audience.

5.6.3 Formality of Text

Formality of text has often been associated with its quality, and this applies in the context of social media as well (Agichtein *et al.* (2008)). Social media are typically public and informal in nature. Therefore, a formal blogname may indicate additional effort from the author to convey the topic of the blog and possibly higher quality content.

In Heylighen and Dewaele (2002), suggested that textual information can be classified into contextual or non-contextual information based on it's content and proposed a measure to quantify the formality of text. Typically text with low contextual information are more formal. Contextual information in text can be inferred from the presence of parts-of-speech (POS) tags such as nouns and articles. Here only the blogname is considered as the username is too short to identify POS tags. We employ the Ark NLP tagger (Owoputi *et al.* (2013)) to identify POS in blognames as the tagger has been trained on short text. The formality score is

$$\begin{aligned} F = ((&P(noun) + P(adjective) + P(prep) + P(article) - P(verb) - P(adverb) \\ &- P(pronoun) - P(inter)) \times 100 + 100)/2. \end{aligned} \tag{5.3}$$

Our analysis shows that the means of the distribution of the formality score of the two classes are different and the difference is statistically significant under a two-tailed t-test (p-value=0.0019). The formality score and the individual POS counts for the blogname of a user are used as features.

In addition to the characteristics described above, the following are also employed:

Popularity of Names

The popularity of a name can be estimated as its probability of a name can be estimated using a bigram character-based language model, where the likelihood of each bigram is estimated using the Maximum-Likelihood approach as

$$P(w_{i-1}w_i) = \frac{c(w_{i-1}w_i)}{c(w_{i-1})}, \tag{5.4}$$

where c is the count in the corpus and W is the vocabulary. The probability of an identity a, $(P(a))$ is computed as

$$P(a) = \frac{1}{|W|}\Pi_{i=1}^{|W|+1}P(w_i|w_{i-1}). \tag{5.5}$$

This language model is smoothed using the Witten-Bell smoothing (Chen and Goodman (1999)), which uses lower-level language model probabilities for smoothing.

Entropy

The randomness in the identities is captured using character-level entropy. This might be an indication of automated usernames and bots, who are less likely to be popular. Given an identity a with character vocabulary W, it's entropy $H(a)$ is

$$H(a) = \frac{1}{|W|}\sum_{i=0}^{|W|} p_i \times \log(p_i) \tag{5.6}$$

5.7 Evaluation

In this section, the approach is evaluated on a real-world Tumblr dataset which was introduced earlier. First, I will introduce the baseline approach which will be used for comparison.

5.7.1 Baseline Strategy

A key challenge of the tackled problem is the unavailability of content or other activity information typically employed in prior works to identify popular users in a social network. Since our focus is on new users and inactive users of a network and existing approaches require some form of activity information, they cannot be applied to the proposed problem in this new context. In classification tasks, a commonly used strategy is the random selection of a class for test instances. Since there is class imbalance, it is more reasonable to predict the *majority class* for all test instances instead. Therefore, this strategy is adopted as the baseline.

Classification Algorithms

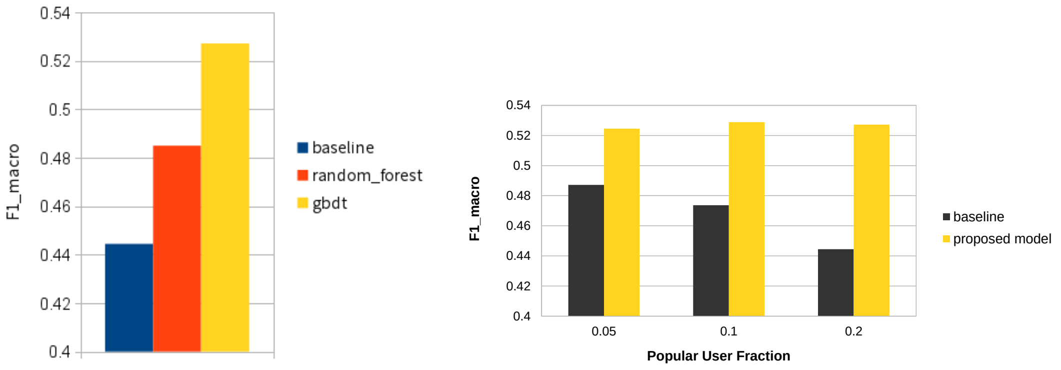
This problem can be solved using one of many existing classification algorithms. Two such popular algorithms are:

- *Random Forest* is an ensemble technique that uses the power of multiple weak learners to determine the final classification result. The implementation of Random Forests in Weka (Hall *et al.* (2009)) is used with the default parameters and no tuning.
- *Gradient Boosted Decision Trees (GBDT)* is another tree-based classifier that uses gradient boosting strategy to combine the results of weak classifiers sequentially. I used a parallel implementation of the technique proposed in Ye *et al.* (2009). In the following experiments, 300 trees were constructed with 20 leaf nodes using the logistic loss function.

Evaluation Measure

The F-measure, which implies high precision and recall, is typically employed to evaluate the performance of classifiers. However, the measure may not be appropriate when there is significant class imbalance. Manning *et al.* (2008) recommends that the macro-F measure, should be used in such cases. Therefore, the macro-F measure (F_{macro}) which is defined as

$$F_{macro} = \frac{1}{q}\sum_{\lambda=1}^{q} F_{\lambda}(tp_{\lambda}, fp_{\lambda}, tn_{\lambda}, fn_{\lambda}), \tag{5.7}$$



(a) Performance Comparison of Different Classifiers

(b) Performance Comparison for Different Thresholds to Determine Popular Users

Figure 5.9: Performance Evaluation

is employed as it considers performance individually on both classes to measure the effectiveness of an approach.Here q is the number of classes ($q = 2$ here), tp is the number of true positives, fp is the number of false positives, tn is the number of true negatives, and fn is the number of false negatives.

5.7.2 Classification Results

The models was trained and evaluated using 10-fold cross validation. Each user in the test set was classified as either a popular user or as a not-popular user. The classification results are presented in Figure 5.9a. The chart shows that the proposed DFI-based features perform significantly better than the baseline. The GBDT algorithm was observed to perform the best and the improvement over the baseline was 16.67%. Our method captures the behavioral patterns in a user's DFI and reflects a user's investment in his identity. It can be conjectured that these patterns reflect the effort a user might similarly apply towards generating content. In addition, the proposed features also capture the strength of the association between a user and his identities which are also indicative of the user's intent.

5.7.3 Effect of the Popularity Threshold

The performance of our approach is affected by the threshold used to determine the popular users in the short head. Thus far I have only discussed the performance on top 20% users as popular users. Therefore, we investigate the effect of varying this parameter on the performance of our method. We repeated the above experiment to construct the dataset by varying the threshold to select the top 10% and top 5% of users as popular users. From Figure 5.9b we can observe that the performance of our method remains relatively stable even when the popular users comprised a smaller fraction of the total population, thus demonstrating that it can be applied to more restrictive scenarios.

5.7.4 Feature Importance Analysis

We also investigated the most popular class of features and the importance of various features. Our observations show that characteristics of blogname were more important than those of username. We also observed that similarity between the two identities was an important factor in the classification task and both similarity measures featured in the top 10 features. The LCS similarity was found to be more important than cosine similarity in a user's DFI. For brevity, we present the top 3 most important features below:

1. The ratio of upper case characters in the blogname to the total number of characters.
2. The probability of the blogname computed using the smoothed language model.
3. The LCS similarity between a user's identities.

5.8 Conclusion

In this chapter, I introduced the concept of a user's Digital First Impression (DFI) comprising of a user's username and blogname on Tumblr. We demonstrated that the characteristics of DFI are indicative of user behavior and that it is indeed possible to identify popular users using the weak signals extracted from DFI. The evaluation results on a real-world dataset collected from Tumblr confirm that the approach is effective in the task of identifying popular users. Finally, it is observed that the relationship between a user's identities is important in the task of inferring user popularity.

Chapter 6

IDENTIFYING TWEETS FROM CRISIS REGIONS THROUGH USER BEHAVIOR ANALYSIS

6.1 Introduction

Twitter has been used with varying success in several recent crises and mass emergency situations. As pointed out in Chapter 2, during emergencies such as Hurricane Sandy in 2012, people published videos and images of the damage caused by the storm. The continued usage of Twitter as a platform to submit crisis related information motivates us to identify relevant information during a crisis. In the previous chapters, we introduced methods to identify relevant users in the context of crisis and when no historical information was available. Another perspective from which this problem can be addressed is by identification of relevant tweets generated during crisis. It is reasonable to assume that first-hand reports on a crisis including reports of damage, are more likely to originate from the crisis region. Therefore, in this chapter we investigate solutions to this problem through the analysis of user behavior.

While Twitter facilitates the tagging of tweets with geographical information, a recent investigation (Morstatter *et al.* (2013)) has shown that only a small fraction (∼1%) of all tweets contain location information. Thus, it is necessary for us to find alternate methods to identify a tweet's location. Recent approaches to tweet location identification leverage the geographic bias in the language of the tweet (Cheng *et al.* (2010)). However, these techniques do not consider the topic bias in the information stream during a crisis. Thus, it is much harder under these circumstances to determine whether a tweet is generated inside the crisis region. These challenges motivate us to to identify tweets from crisis region.

Identifying the user's location is an alternative solution, which can be found using social network information (Rout *et al.* (2013)) or user's historical tweets (Cheng *et al.* (2013); Mahmud *et al.* (2012)). Typically, identifying and extracting additional information such as a user's network, or his tweet history during a crisis is not practical due to the API constraints imposed by Twitter. These approaches also fail when there is insufficient network, or content history. Hence, we cannot apply these techniques effectively to identify tweets from crisis region. Recent studies have also shown that a user's location may not necessarily correspond to the location of the tweet (Cho *et al.* (2011)), due to user mobility. Thus, it is more reasonable to identify tweets generated from crisis regions.

Here, we conduct a study of crisis tweets to gain deeper insight into their characteristics and to specifically answer the following questions: 1) Do tweets inside crisis region express different behavioral patterns and can these patterns be used to identify tweets from crisis region when explicit location information is unavailable. Our contributions are the following:

- We formally define the novel problem of identifying tweets from a crisis region and highlight the challenges;
- We conduct a study of tweets from major crises to discover behavioral patterns in Section 6.2; and
- We propose an approach to identify tweets from crisis region in Section 6.3.

Problem statement: Given a crisis C associated with a crisis region R, and a collection of tweets relevant to the crisis T, where each tweet $t \in T$ contains tweet data t_d, including the tweet text and the user's profile information t_u, but does not include the location information. For each such tweet $t \in T$, we need to decide whether the tweet is generated from inside the crisis region, i.e., $t \in R$ or the tweet is generated from outside the crisis region, i.e., $t \notin R$.

6.2 Behavioral Patterns in Tweets

We begin with a study of the characteristics of crisis tweets to identify distinct behavioral patterns.

6.2.1 Datasets

To conduct this study, we collected tweets pertaining to major crises in 2011 and 2012. All the data was collected using our tweet monitoring platform TweetTracker through parameters specified by virtual volunteers from the NGO Humanity Road [1] and analysts from various governmental agencies monitoring the crises. In Table 6.1, we present a full list of the parameters used to collect the data along with a characterization of the events into different disaster types. Specifically, keywords, hashtags, geographic locations and Twitter user handles were used to collect the data between 2011 and 2012. The datasets comprise of tweets in various languages and span different geographic regions.

[1] `http://www.humanityroad.org`

Table 6.1: Parameters Used to Crawl the Datasets

Crisis Type	**Dataset name**	**Keywords/Hashtags**	**Geoboxes**	**Userids**
Earthquakes	EQ Turkey	#Van, #Ercis, #Turkey, #Magnitude		
	EQ Japan	#jpquake, #eqjp, #japaneq, #japantsunami, #fukushima		
Flooding	FL Mississippi River	#flood, #mississippiriver, #joplin, #MO	SW:(29.3, -92.2), NE:(35.5, -89.0)	
	FL Minot ND	#ndflood, #bisflood, #minotflood, #minot, #ndwx, #flood	SW:(46.5, -103.8), NE:(48.8, -99.7)	
Hurricanes	HUR Isaac	#isaac, hurricane isaac, storm isaac, #hurricane		
	HUR Lee	#lee, #hurricane, #lawx, #TSLee		
	HUR Sandy	hurricane, sandy, florida, storm, tropical, frankenstorm, sandyde, evacuation, stormde, dctraffic, mdtraffic, vatraffic, baltraffic, nyctraffic, njsandy, nysandy, ctsandy, dcsandy, desandy, njtraffic, shelter, damage, tree, treedown, outpage, linedown, power, flood, water, surge, outage, #hamptons, #northfork, #nofo		A list of 75 users representing NGOs and authorities from NJ and NYC
Socio- Political Events	SP London Riots	#londonriots, #clapham, #croydon, #peckham, #Hackney, #UKuncut, #Tottenham, #MetPolice, #liverpoolriots	SW:(51.2, -0.53), NE:(51.69, 0.2)	
	SP Occupy Wall Street (OWS)	#occupywallstreet, #ows, #occupyboston, #p2, #occupywallst, #occupy, #tcot, #occupytogether, #teaparty, #99percent, #nypd, #takewallstreet, #occupydc, #occupyla, #usdor, #occupysf, #solidarity, #15o, #anonymous, #citizenradio, #gop, #sep17, #occupychicago, #occupyphoenix, #occupyoakland		
Wildfires	WF AZ	#wildfires, #wallow, #wallowfire, #nmsmoke		
	WF CO	#LowerNorthForkFire, colorado wildfire, #colorado #wildfire, #waldofire, #waldocanyonfire, #cofire, #cofires, #highparkfire, #pyramidmtnfire		

Table 6.2: Dataset Characteristics

Dataset	**# Tweets**	**# Retweets**	**# Geo-tagged Tweets**	**# Inside Tweets**	**# Outside Tweets**
EQ Japan	2,734,431	1,223,609	105,669	44,119	28,953
FL Mississippi River	157,435	72,377	3,042	944	1,355
HUR Sandy	4,344,308	2,203,262	58,092	36,324	15,455
SP OWS	10,722,020	5,039,152	95,313	43,489	38,557
WF AZ	8,679	2,865	213	85	35

Preparing the dataset: To study the characteristics of the tweets, we must first identify tweets originating in the crisis region. The affected region for each crisis was decided based on the nature and scale of the crisis. For example, as Hurricane Sandy affected the entire East coast of the United States, we consider the region extending from Florida to New York as the crisis region R. Once a crisis region was determined, tweets from crisis region were identified through the geotagging information explicitly provided by the users. As the location information is voluntarily provided, we assume that it is accurate. Geotagged tweets which contained only a link or no content at all were removed. From Table 6.2, it can be observed that geotagged tweets typically comprised a small fraction of the data. The distribution of the tweets is presented in Columns 5 and 6 in Table 6.2.

6.2.2 Characteristics of Tweets from Crisis Regions

In this study, we investigate whether we can discover patterns which can help us identify tweets inside a crisis region. Previous studies have shown that the primary application of Twitter during a crisis involves information dissemination (Hughes and

Palen (2009b); Heverin and Zach (2010)). Existing studies have investigated the characteristics of tweets in other contexts such as the usage of mobile devices (Perreault and Ruths (2011)). Motivated by these studies, we propose to conduct a comprehensive study of the characteristics of tweets generated during crises along the three relevant dimensions:

- device and platform usage patterns (how),
- characteristics of the generated content(what), and
- the motivation to published content(why).

In the next sections, we will tackle these broad dimensions of user behavior individually. For each identified behavior, we will follow the below procedure to compare the behavior in the tweets. To compare the characteristics we propose to compute the likelihood of observing the behavior in tweets inside crisis regions and the likelihood of observing the behavior in tweets outside crisis regions. Then, we compare these quantities using the *Likelihood Ratio*. The *Likelihood Ratio* can tell us how likely are the tweets inside the region to demonstrate a behavior compared to tweets outside the crisis region. Given a behavior b, it's *Likelihood Ratio* is:

$$LR_b = \frac{P(b|inside)}{P(b|outside)}, \tag{6.1}$$

where $P(b|inside)$ is the likelihood of the behavior to be exhibited in tweets inside crisis region and $P(b|outside)$ is the likelihood of the behavior to be exhibited in tweets outside crisis region. If $LR_b > 1$, then the tweets inside the crisis region are more likely to express the behavior and the magnitude of the ratio indicates how likely this is. Further, to establish the observed differences in the behavior is significant we employ statistical tests.

Testing statistical significance: To establish that the observed differences in the behavioral patterns are statistically significant, we will employ the two tailed t-test to test the statistical significance of the results. For all the comparisons, we set the significance level $\alpha = 0.05$. Let μ be the mean of the number of tweets exhibiting the behavior inside crisis region and μ_0 be the mean of the number of tweets exhibiting the behavior outside crisis region. The null hypothesis H_0 for the test is:

H_0 : the tweets inside the crisis region and tweets outside the crisis region demonstrate similar behavior, i.e., $\mu = \mu_0$.

If the p-value of the test is below the chosen significance level, then we can reject the *null hypothesis* and say that the observed differences are statistically significant. Otherwise, we accept the *null hypothesis* and conclude that the observed differences in behavior are not statistically significant.

6.2.3 Platform and Device Usage

Crises are typically associated with failure of public utilities and other services, and studying how Twitter is accessed can help first responders create a more effective response and aid in dissemination of information. Tweets generated from mobile devices may be able to provide additional information such as images or videos of the destruction. Additionally, studying such tweets also enables us to reasonably estimate the mobility of users. It is now known that Twitter has more than 184 million mobile users [2] , thus making this a reasonable behavior to investigate.

Platform characteristics can also significantly influence the behavior exhibited by the users. For example, retweet is a popular mechanism by which information propagates on Twitter through social connections. Retweets enable users to highlight and promote content they find relevant and during crisis this can be interpreted as an endorsement of the content and this has been used to perform various tasks such as verify the credibility of information in the past (Castillo *et al.* (2011)). Therefore,

[2] `http://tnw.co/1nP4RPk`

we ask the following two questions to understand user behavior:

- Are mobile devices frequently used to publish tweets from inside crisis regions?
- Do users publish original content or retweets inside crisis regions?

Next, we investigate how these patterns vary in crisis tweets.

Is the usage of mobile devices more prevalent inside crisis regions?

Mobile devices are ubiquitous. The usage of capable mobile devices such as smartphones is increasing rapidly. In the United States alone, the usage of smartphones exceeded 60% of all mobile subscribers [3] .During hurricane Sandy it was noted that the usage of mobile devices significantly increased and overlapped with the peak of the crisis [4] . Therefore, we begin with an investigation of the mobile device usage in crises.

Each tweet is associated with a client/application which was used to publish the tweet. From the client, it is possible to detect whether a tweet was published using a mobile device by verifying whether the client used was a mobile client. As there are no explicit resources to distinguish between mobile clients and non-mobile clients, we present a strategy to perform this task.

Procedure to identify mobile clients: In our datasets, we observed that the popularity of clients followed the power-law distribution, where only a few clients were used to publish most tweets. This behavior can be observed in Figure 6.1, which shows the usage of clients in a log-log plot.

As the distinction between mobile and non-mobile clients needs to be performed manually and analyzing all the observed clients is not practical, we focus our effort on the most popular 100 clients from each dataset. To demonstrate that this is adequate,

[3] `http://bit.ly/1bgXMlX`

[4] `http://tcrn.ch/1r1483x`

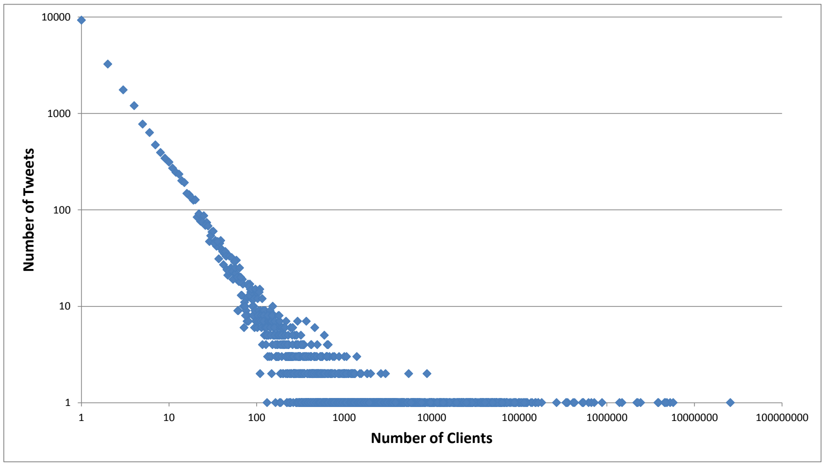


Figure 6.1: Distribution of Client Usage in Crisis Data

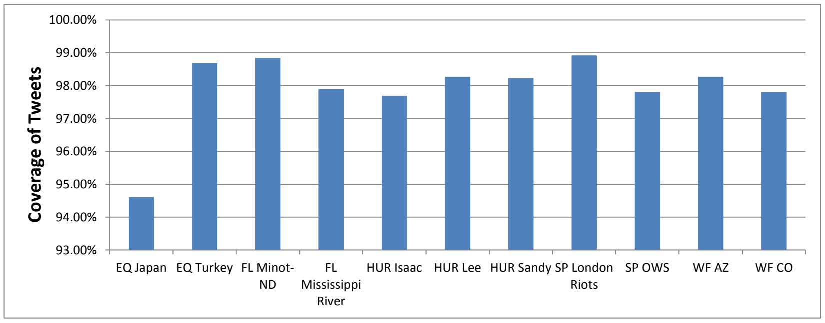


Figure 6.2: Crisis Data Generated from the Top 100 Clients

we present the coverage or the number of tweets which use any one of these clients in Figure 6.2. Our study shows that the top 100 clients account for more than 94% of all tweets generated in any dataset. To identify the mobile and non-mobile clients among these, we followed the procedure below: Some clients clearly indicate their mobile nature. For example *Ubertwitter for Android*. But, most client information is not as descriptive. However, each client is associated with a homepage, where additional client information can be found. We manually verified if a client was a mobile client by using the information on its homepage. If a client indicated that it was an API to publish tweets, provided a desktop tool, or was a bot, then it was classified as a non-mobile source. Using this procedure, we discovered 220 unique mobile clients and 556 unique non-mobile clients across all datasets.

Using the annotated clients, we investigated the number of tweets published in each dataset using mobile clients. We set the behavior b in Equation 6.1 to *mobile* and compute LR_{mobile}. In most datasets, we found that the tweets inside the crisis region were likely to be generated using mobile devices. In the case of *EQ Japan*, tweets inside crisis region were more than twice as likely to be published using a mobile device. Disasters and mass emergency situations are typically associated with the failure of utilities and increased mobility of users, such as in during Hurricane Sandy [5] . Moreover, as Twitter is associated with the publication of first hand reports, it is reasonable that tweets inside a crisis region are more likely generated using mobile devices. A summary of LR_{mobile} is presented in column 1 in Table 6.3. The t-test also confirmed that the observed differences are generally statistically significant.

Are tweets from the crisis region more likely to be retweets?

There are several methods to publish tweets on Twitter. When tweets posted by another user are forwarded by a user, the tweet is called a retweet. Retweets are characterized by the inclusion of the symbol "RT" at the beginning of the tweet and the original user. Typically retweets constitute a large part of crisis related tweets. However, they lack originality. Therefore, we compare the pattern of publishing in tweets inside and outside crisis regions. We measure $LR_{retweet}$ by setting the behavior b in Equation 6.1 to *retweet* for each dataset and summarize the observations in Column 2 of Table 6.3. The results show tweets from crisis regions are less likely to be a retweet and thus more original. Ten of the eleven datasets studied exhibited this pattern. During a crisis, we would expect that a tweet published from crisis region

[5] `http://nation.time.com/2012/11/26/hurricane-sandy-one-month-later/`

Table 6.3: Behavioral Characteristics in Tweets: I (* Indicates p-value < 0.05 and ** Indicates p-value << 0.05)

Dataset	LR_{mobile}	$LR_{retweet}$
EQ Japan	2.35 *	0.11 **
EQ Turkey	0.64 **	0.24 **
FL Minot-ND	1.16	0.08
FL Mississippi River	0.86 **	0.89
HUR Isaac	1.12 **	0.26 **
HUR Lee	1.55 **	0.86
HUR Sandy	1.07 **	0.40 **
SP London Riots	1.44 **	0.28 **
SP OWS	0.96 **	0.81 **
WF AZ	0.46 *	2.47
WF CO	1.10	0.68

to contain original information due to their exposure to the crisis. These users are more likely to have access to information that users outside the crisis regions may not have.

6.2.4 Motivation to Publish Tweets During A Crisis

Twitter provides a convenient platform to publish content and has thus been used in many recent crises. Some indications of the intention to use Twitter as a communication channel can be observed during the Arab Spring protests of 2011, when Twitter was employed to report the atrocities by the regime [6] or during Hurricane Sandy in 2012 to debunk rumors [7] . In this section, we investigate the underlying motivation behind using Twitter as a mechanism to publish content during a crisis.

[6] http://bit.ly/1wIPkqo

[7] http://bit.ly/1HB65sL

Table 6.4: Behavioral Characteristics in Tweets: II (* Indicates p-value < 0.05 and ** Indicates p-value $<< 0.05$)

Dataset	$LR_{conversation}$	$LR_{hashtag}$	$LR_{emotion}$
EQ Japan	1.86 *	0.06 *	2.51 **
EQ Turkey	0.89 *	0.47 **	0.67 *
FL Minot-ND	1.37 *	0.22 **	2.52 *
FL Mississippi River	1.09	0.72 **	0.47 *
HUR Isaac	0.75 **	1.15 **	0.91
HUR Lee	1.21	0.91 *	0.99
HUR Sandy	0.78 **	1.34 **	0.60 **
SP London Riots	2.06 *	0.26 **	1.99 **
SP OWS	0.96 **	0.86 **	0.45 **
WF AZ	0.94	2.59 **	0.10
WF CO	0.46 *	0.45	0.0

Specifically, we attempt to answer the following questions:

- Do users in crisis regions participate in conversations?
- Do tweets from crisis regions seek visibility?
- Are tweets from crisis regions more likely to express their emotions through tweets?
- Are tweets from crisis regions likely to indicate an action undertaken through tweets?

Are tweets from crisis regions more conversational?

Conversations are typically of relevance to the parties involved in the conversation and they do not contribute as much as other tweets to situational awareness. Therefore, we investigate whether this behavior is prevalent in crisis tweets. Conversational elements in a tweet include the mention of other users using the syntax “@username”. To initiate a conversation a user needs to begin the tweet with the mention of the target user. We perform a comparison of this behavior in the two types of tweets by measuring $LR_{conversation}$ by setting the b in Equation 6.1 to *conversation*, which uses the above pattern to identify conversations. $LR_{conversation}$ for all studied datasets are summarized in Column 1 of Table 6.4. From our study we observed that conversational elements were very likely to in both forms of tweets. In the context of a crisis, this behavior can be observed from people who have relatives or friends in the crisis region or from people inside the crisis region attempting to inform others of their status or provide updates on the crisis. For example, consider a scenario where users are attempting to check the condition of relatives inside the crisis region in an attempt to get more information on the impact of the crisis.

Are tweets from the crisis region more likely to seek visibility?

A very large volume of tweets is generated on Twitter every day. Therefore, finding content is a challenge for Twitter users. Twitter employs a mechanism known as hashtags to indicate the topic of a tweet and to facilitate searches. Therefore, using multiple hashtags allows the content to be visible to a wider audience (Page (2012)). To investigate this behavior, we compute $LR_{hashtag}$. The results in Column 2 of Table 6.4 indicate that in eight of the eleven datasets studied, the tweets from crisis regions were less likely to include multiple hashtags in the tweet. Although this may be surprising at first as seeking visibility would be an expected behavior. The usage of multiple hashtags significantly reduces the information which can be published in tweet due to the character constraints. Since, major crises are typically associated with popular hashtags especially in the context of hurricanes, where there is forewarning of the crisis, it might be sufficient to use fewer hashtags in the tweet.

Are tweets from the crisis region more emotional?

Emotional cues in tweets indicates the mood of the user. One way to measure sentiment is through the identification of emoticons, which has previously been used to detect emotion (Hu *et al.* (2013)). Here, we identify emoticons using the Ark POS Tagger (Owoputi *et al.* (2013)). As we are not concerned about the polarity of the emotion, we only measure the presence of an emotion in the tweets. To compare the two types of tweets, we compute $LR_{emotion}$, which is presented in Column 3 of Table 6.4. We found that tweets inside the crisis region are less likely to express emotion. This can be observed from $LR_{emotion}$, which is < 1, for eight of the eleven datasets studied. We suspect that this might be because tweets inside crisis regions often provide situational awareness and thus are more informational, whereas tweets outside the crisis regions tend to express emotional support with the people affected by the crisis.

Are tweets from the crisis region more likely to indicate an action?

Action words imply that the user publishing the tweet is performing an action. Verbs are typically used to indicate an action, such as leaving, moving etc. in the context of a crisis. Thus, we can investigate whether tweet is indicating an action by identifying verbs in the text. To detect verbs in a tweet, we use the Ark NLP Part-Of-Speech (POS) tagger and compute LR_{action}. The comparison is summarized in Column 4 of Table 6.4. Our results show that tweets inside the crisis region are less likely to contain action words compared to tweets generated outside the crisis region and that this behavior was consistent across most datasets. The results were also found to be statistically significant.

6.2.5 Message Content Characteristics

Twitter is often used as a news source by its users. It is already established that topics discussed by Twitter users often reflect the currently trending topics in traditional news media (Kwak *et al.* (2010)). However, it is necessary to establish that such behavior can be observed in tweets from crisis regions, so we can leverage them for the above mentioned task. Specifically in this section, we will analyze whether the content generated by Twitter users have differing message characteristics by answering the following questions regarding tweets from crisis regions:

- Do such tweets refer to physical entities more frequently?
- Do such tweets leverage external resources to elaborate their message with additional content?
- Do these tweets generate novel content from the perspective of the crisis?

Next we investigate these questions individually using the framework discussed earlier.

Are tweets from the crisis region more likely to reference entities?

Entities typically refer to the names of people, buildings, or specific locations. During a crisis, such entities may refer to people involved in the crisis response, landmarks in the crisis region etc. There are several methods to detect entities, but here we employ the Ark POS tagger to annotate the tweets and identify the proper nouns, since proper nouns are typically indicative of names of people and places. Therefore, by analyzing the proper nouns in the tweets we can determine whether the tweets reference entities. We set the behavior b in Equation 6.1 to *entity*, to compare the likelihood of tweets inside the crisis region to the tweets outside the crisis region. We would expect that first hand reports from the crisis region would be more likely to reference local entities

Table 6.5: Behavioral Characteristics in Tweets: III (* indicates p-value < 0.05 and ** indicates p-value << 0.05)

Dataset	$LR_{entities}$	$LR_{resource}$	LR_{action}
EQ Japan	0.43 *	2.37 *	0.34 *
EQ Turkey	1.20 **	0.64 **	0.70 **
FL Minot-ND	0.95	0.90	0.84 *
FL Mississippi River	0.86 **	0.79 **	1.10 **
HUR Isaac	0.93 **	1.02	0.96 **
HUR Lee	0.73 **	0.75 **	1.05
HUR Sandy	0.93 **	1.50 **	0.93 **
SP London Riots	0.76 **	0.73 **	0.89 **
SP OWS	0.98 **	1.03 *	1.08 **
WF AZ	1.06	1.51 *	0.97
WF CO	0.97	0.53 *	1.00

who may be affected/involved in the crisis, thus providing situational awareness to first responders and other responding agencies. However, LR_{entity} for the datasets shows that the inside the crisis region are almost as likely to contain a reference to entities as tweets outside the crisis region. In Column 1 of Table 6.5, we summarize the results. This might be due to the fact that as the awareness and visibility of a crisis grows, the information about local entities is propagated quickly and easily outside the region and used in tweets from both inside and outside the region. The difference in the behavior was found to be statistically significant in most datasets.

Are tweets from the crisis region more likely to share an external resource?

Twitter messages are restricted to 140 characters. Thus, longer content and media such as images and videos are shared by people through external references to describe their experiences. Therefore, we investigate whether tweets from crisis regions are more likely to share external resources. In this study, we measure $LR_{resource}$. We evaluated if the tweets inside the crisis region are more likely to contain URLs pointing to external media or resources. To avoid any bias from retweets, which can propagate an original tweet inside the crisis region around the world, we only compared the likelihood of original tweets to point to external resources. During a crisis, we expect that tweets emerging from the crisis region, contain links to resources such as videos and images pointing to the current state of their environment. For example, during Hurricane Sandy images of the flooding in the streets and the subway were widely shared by the local residents [8] . While tweets from crisis regions were more likely to contain URLs in half of the datasets, the observation was not consistent across all of them. Temporal factors might be one reason for this observation as tweets outside the region are likely to point to major articles and media already observed, whereas tweets inside the region are focused on new information. The $LR_{resource}$ for the datasets is summarized in Column 2 of Table 6.5.

Are tweets from the crisis region novel?

Novel content typically indicates original information and not a tendency to publish popular content. To answer this question, we construct a unigram language model assuming the tweets are constructed using a bag-of-words strategy. The likelihood of each word $P(w)$ is estimated using the Maximum-Likelihood approach as

$$P(w) = \frac{c(w \in W)}{\sum_w c(w)}, \tag{6.2}$$

[8] `http://bit.ly/1sgSj9h`

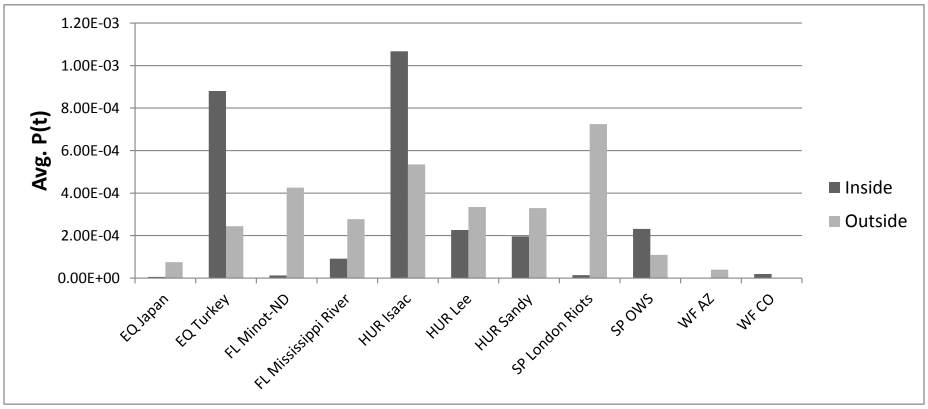


Figure 6.3: Average Probability of Tweets in Various Crises Data

where $c(w)$ is the count of the word in the corpus. The probability of a tweet $(P(t))$ is computed as

$$P(t) = \Pi_{w \in W} P(w). \tag{6.3}$$

In Figure 6.3, we present a comparison of the average probability of a tweet in each dataset. The experiment shows that tweets inside crisis region are novel.

6.2.6 Summary

In this study, we investigated user behavior in crisis tweets. We established that tweets from crisis regions exhibit different behaviors than tweets outside crisis regions and this difference has been observed to be statistically significant. Here we present some key insights gathered from the above study.

- Tweets from crisis regions are generally associated with original content and they are also more likely to discuss novel topics rather than popular topics, which reaffirms previous findings on the information dissemination behavior of tweets during crisis. This can be explained as the tendency of users inside crisis region to post information obtained first-hand from the region, whereas users outside the regions are less likely to have access to such information.

- We discovered that tweets from crisis regions are more likely to be published using mobile sources, which may be due to the availability and convenience of mobile devices to during a crisis.
- From the study, we found that tweets inside crisis region were less likely to be part of a conversation or express emotion, in both cases, it indicates a tendency to publish content which is more pertaining to the crisis and of interest to the general community rather than inter-personal conversations.
- The differences in the behavior observed in the wildfires datasets were generally found to be not statistically significant. We attribute this to the significantly smaller size of the datasets compared to others and we attribute this to the data collection strategy and the time period.

In the next section, we will elaborate on the task of predicting whether a tweet originates from crisis region using the observed patterns.

6.3 Evaluation

Using insights from the study, we want to answer the following question: can we use the observed behavioral patterns to decide whether a tweet originated from crisis regions. To answer this question, we created various features to describe a tweet based on the behavioral patterns observed in the previous section. These features are summarized below:

Mobile Features: Using a manually annotated list of 220 mobile sources and 556 non-mobile sources as described earlier, we create a boolean feature to identify whether a tweet is published using a mobile client.

Resource Features: As discussed in the study previously, tweets from crisis region are likely to contain links to external resources. Therefore, features indicating the presence of a URL and the number of URLs are used as representative resource features. In addition, we identify whether the URL points to an image or a video using regular expressions and popular image and video hosting domains. The presence of Foursquare location references are considered separately as an indication of location information.

Textual Features Patterns contained in tweets, such as whether a tweet is a retweet, a directed message, or contains a user mention as well as the usage of hashtags. Positive and negative emoticons are distinguished using a list of popular happy and sad tags which are compiled and listed in Wikipedia [9] . The occurrence of punctuations was used as an indication of the quality of the text. The length of a tweet represented by both the character length and the word length are also considered as features.

Linguistic Features: Having studied the usage of various parts of speech in the crisis tweets, we use both the presence and the frequency of various part-of-speech tags in the tweet. These include proper nouns, verbs, and pronouns, which indicate references to actions, and entities in a tweet.The presence of emotion is also used as a feature by detecting the presence of emoticons. Additionally, given a tweet t with vocabulary W, we compute and use the probability of the most probable word ($max_w P(w)$ and the least probable word $min_w P(w)$) as features. The probability of the tweet, $P(t)$, computed from the language model described previously is also used as a feature.

User Features: Typically user related information that accompanies a tweet includes the user's complete profile, which contains information about the user. From this information, we created features to include information such as, the user's previously published number of status messages which indicates his familiarity with the medium and the user's network activity which is captured using the number of connections (both followers and friends).

[9] `http://en.wikipedia.org/wiki/List_of_emoticons`

In the next section, I will discuss the baseline approaches against which a comparison could be made. Later, I will present experimental results on multiple datasets.

6.3.1 Comparison with Other Approaches

Given the constraints of our problem, where the only information available to us is the information which is collected during a crisis that includes the tweets themselves and the profile of the publishing user, we cannot directly compare with existing location prediction approaches, such as Mahmud *et al.* (2012) and Cheng *et al.* (2010). These methods require additional historical data to be applicable to this task and as pointed out in Section 6.5, collecting a user's historical tweets or historical tweets from a specific geographic location may not be feasible in the context of a crisis. Since, in this work we assume that we have access to only the tweets collected during a crisis and the user profile collected as metadata, we seek alternative approaches to perform a fair comparison with our method. Therefore, we propose the following approaches which can also operate on limited information:

- A common baseline employed in classification tasks is selecting a random class as the prediction for the test instances, where each class has equal probability of being selected as the prediction. Since there are two classes in our problem, this would imply that the probability of selecting each class is 0.5. However, in our dataset, clearly there is a bias in the distribution of the tweets where in most cases tweets inside the crisis regions occur more often. Therefore, it is more appropriate for us to use the **majority** class as a baseline. This information is summarized in Table 6.2.
- Tweet content has been shown to be a good indicator for the prediction of a user's location or a tweet's location. Additionally, the objective of identifying tweets from the crisis regions is to reduce information overload by adopting a

strategy to prioritize the investigation of some tweets over others. Therefore, we choose to employ a content based approach as an alternative strategy for comparison. Verma et al. Verma *et al.* (2011) introduced a content based method to identify tweets containing situational awareness. In this approach the authors investigated and demonstrated that linguistic features such as the unigrams extracted from tweet text and their raw frequency as well as the Parts-Of-Speech (POS) tags are effective in identifying tweets containing situational awareness. Here we will use unigrams and their raw frequency along with the POS tags computed using the Ark POS Tagger to distinguish tweets from crisis regions from other tweets and denote this approach as (LF). All stopwords and Twitter elements(hashtags, URLs, and usernames) were removed during preprocessing. A key drawback of adopting a content based approach over the proposed approach is that the vocabulary of tweets may continue to grow rapidly with new tweets due to the informal nature of the language in tweets. In our dataset, we observed that the number of features was typically proportional to the number of instances, thus requiring significantly more time for model construction and storage.

Below we present the performance comparison with these baseline approaches. All the experiments were performed using the implementations available in Weka (Hall *et al.* (2009)). The results were generated using 5-fold cross validation with 80-20 split, where 80% of the data was used for training and 20% was reserved for testing. All experiments were carried out independently on each dataset. Default parameters were used to train the model and no tuning was performed.

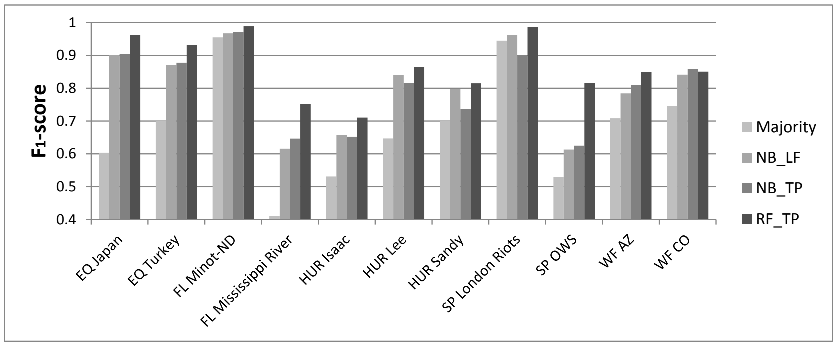


Figure 6.4: A Comparison of the F_1 Score

6.3.2 Evaluating the Performance

To evaluate the utility of the proposed features we employed the naive Bayes classifier to identify tweets inside a crisis region. The accuracy of our approach and the baseline methods is presented in Figure 6.4. To compare the performance of the methods, we employ the F_1 score, which can be computed as

$$F_1 = \frac{2 * (precision * recall)}{precision + recall}. \tag{6.4}$$

The results in Figure 6.4, describe the performance of the naive Bayes classifier applied to the baseline features NB_LF, and those constructed from our study NB_TP. In most datasets, we find that our approach performs better than the baseline with significantly greater efficiency due to the reduced number of features. In addition, we also trained a Random Forest classifier (RF_TP) using the proposed features and found that it performed considerably better than the naive Bayes classifier in general.

To account for the class imbalance, we also present the weighted AUC score for the datasets in Figure 6.5. Since the "majority" baseline is not as effective as the other baseline method, we omit it from the following results. Here, the improvement in performance can be observed more clearly as the proposed approach outperforms the baseline in all but one of the datasets. To verify that the improvement is sta-

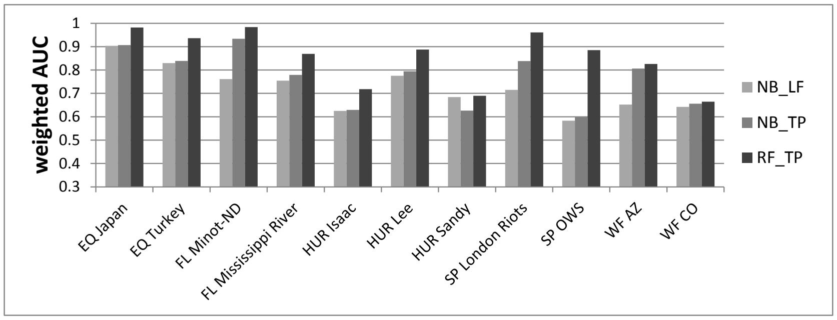


Figure 6.5: A Comparison of the Weighted AUC Score

tistically significant, we conduct the Wilcoxon-Signed Rank test. As the results are obtained using the same data, we compare the baseline (NB_LF) and our approach (NB_TF), treating them as a paired sample. The observed p-value was 0.0019 and the improvement was statistically significant at $\alpha = 0.05$.

Although the results show that the linguistic features are good indicators for location prediction, our results show that using the behavioral characteristics identified in the study above, we can better predict whether a tweet is inside a crisis region. Additionally, we can perform this with greater efficiency because using the linguistic features requires us to maintain a vocabulary as the content of the tweets are likely to change as a crisis develops. This requires constant maintenance of the vocabulary and may require frequent reconstruction of the model to reflect the evolution of the content. However, using the proposed approach we have shown that this can be achieved with significantly less overhead as no vocabulary needs to be maintained. The proposed approach is also more efficient compared to using the linguistic features as it can be constructed and applied to data much faster due to the very large number of features in the LF baseline.

6.3.3 Feature Importance Analysis

To evaluate the importance of different feature classes in the task, we constructed a Logistic Regression classifier. This classifier learns a weight for each feature, which can be interpreted as a measure of the feature's importance for the prediction task. The features can then be ranked based on the absolute value of the learned weight for each individual dataset. We ranked the proposed features based on the sum of the ranks of each individual feature across all datasets to identify its overall prediction power. Lower rank values indicate that a feature is important across the datasets, while larger values indicate lower importance of the feature. We discovered that linguistic features were found to be the most important class of features for the task. This can be attributed to the fact that tweets inside the crisis region are novel and less likely to contain emotions. Textual features such as the use of punctuations was more useful than reference to entities and action words. Resource related features were also found to be generally as important in distinguishing tweets from the crisis region, which can be explained by the tendency of tweets inside the crisis region to contain links to external resources, particularly Foursquare. User related features were typically the least important class of features for the task, thus suggesting that prediction can be reasonably performed with just the information contained in the tweet even when the user information is unavailable.

6.4 Case Study: Application on Arizona Wildfires Data

Tweets from crisis region can be used to obtain situational awareness and can be used to generate post-crisis summarization of the event from the perspective of tweets. In the task of event summarization, the goal is to identify a small number of representative tweets from the entire corpus, which can describe the event or a

Table 6.6: Arizona Wildfires Summary Created Using Z_i

Summary Tweets
it's really smokey and hazy today. #wallowfire
smoke near eagar #wallowfire http://twitpic.com/5ci9i7
wildfire info: wallow fire pm update 6/19/11 (wallow wildfire) http://bit.ly/mdoigp #azfire #wallowfire
#wallow fire swept thru greer.
glenwood gazette - breaking news: #wallowfire 06/10/11 map http://t.co/xr8e23b

crisis. In this case study, we will use the task of event summarization to demonstrate that the application of our approach enables the generation of a more meaningful summary of the crisis.

Extracting representative tweets from topics derived from the tweets is a commonly used approach to event summarization (Chua and Asur (2013)). To illustrate the differences between the two sets of tweets, we will summarize the Arizona Wildfires (*WF AZ*). First, the proposed model is used to classify all tweets whose location information is unknown. Then, the following procedure is applied:

- Extract 10 topics Z_i from tweets inside and Z_o from tweets outside crisis region.
- For each detected topic, rank tweets t with vocabulary w by its perplexity score defined as $perplexity(t) = exp\left(\frac{-logP(t|z)}{|w|}\right)$.
- Create a summary of the crisis by picking the 5 most relevant tweets from the top 10 tweets in the topic.

The extracted summaries in Tables 6.6 and 6.7 show that the summary created using Z_i has more relevant information and it highlights the relevance of our approach.

Table 6.7: Arizona Wildfires Summary Created Using Z_o

Summary Tweets
wildfires wreaking havoc in arizona. http://bit.ly/jsgwpv
#arizona - y su bonito glowing bird suena en radio paranoia :)
rt @radionoisefm cel mai devastator incendiu din a.. http://bit.ly/jtshyl #12 #ore #arizona #devastator #dublat #incendiu
hello #arizona, #bringit :— http://instagr.am/p/f4ife/
1600 quadratkilometer wald durch brand vernichtet #arizona

6.5 Related Work

Social media services have been extensively studied as social sensors to monitor important events occurring in the real world. In particular, recent research has focused on the analysis of the use of social media during emergencies (Sakaki *et al.* (2010)), including earthquakes (Mendoza *et al.* (2010)), riots (Panagiotopoulos *et al.* (2012)), wildfires (Sinnappan *et al.* (2010)), etc. Seeking high-quality social media data pertaining to crisis serves as the basis of these studies and motivates this study to identify tweets from crisis regions.

Identifying a user's home location using social media data (Hecht *et al.* (2011)) is an interesting and important problem. The existing research on this topic can be divided into two groups. The first set of research methods assume that a user's tweets might contain distinct features due to their proximity to the region. Cheng *et al.* (2013) estimated that a Twitter user's home city based on the content of their tweets. Mahmud *et al.* (2012) used an ensemble of statistical and heuristic classifiers to infer the home location of Twitter users at different granularities by using the content information and their tweeting behavior. However, topic specific variation of content has not been investigated. These approaches also rely on the availability of a user's tweet history, which is not readily available during a crisis. The second set of research methods assume that a user's home location is strongly correlated with his friends' home location. Backstrom *et al.* (2010) estimated the home location of Facebook users using user-supplied address data and the network of associations between members. But, due to the API limitations it is not practical to extract network information during a crisis under time constraint. Therefore, these approaches cannot be directly applied to our data.

The problem of recognizing eyewitness tweets was independently investigated in Morstatter *et al.* (2014). While the authors evaluated whether linguistic features could be used to identify such tweets, here we analyzed several kinds of behavioral patterns in tweets from crisis regions.

6.6 Conclusion

Identifying tweets from crisis regions is becoming increasingly important due to information overload on Twitter. In this chapter, we investigated tweets from several crises to study user behavior exhibited in tweets published from crisis regions. From our study, we observe that tweets from crisis regions demonstrate distinct user behavior compared to tweets outside crisis regions. Using the findings from this study, we develop a novel framework which to analyze crisis tweets and introduce a predictive model to identify tweets originating from crisis regions to empower first responders and analysts to tackle information overload. Our experiments confirm that the proposed method can successfully identify tweets from crisis regions.

Chapter 7

CONCLUSION

7.1 Summary of Contributions

In this thesis, I have investigated the growing use of social media during crises around the world and its impact on activities such as disaster response. Specifically, I have investigated novel and challenging problems faced by first responders and analysts when employing social media, particularly microblogging platforms to obtain situational awareness during emergencies for planning effective response strategies. As social media gains prominence as an alternative media outlet, where individuals act as sources of information, there is increasing demand for methods and platforms to aid in handling the information generated. This thesis makes the following contributions:

- The TweetTracker system for gathering and analyzing tweets from crises in a collaborative environment (Chapter 2) using visual analytics. The TweetXplorer system, a visual analytics based platform to facilitate and guide the users to analyze microblogging posts generated during crises. More information on the systems can be found at `http://tweettracker.fulton.asu.edu`.
- An introduction to the characteristics and challenges of streaming data and a novel approach to identify crisis events in the context of dynamic Twitter streams under these challenges in Chapter 3. The approach was also evaluated on both random streams and topic specific streams encountered in the real-world.
- An investigation of the characteristics of users who publish tweets during a crisis (Chapter 4). To identify relevant users, I propose a geo-relevancy score and a

topic affinity score. Experimental results show that the classification approach proposed using these scores can successfully identify information leaders. These information leaders publish more information and higher quality information that other users and this knowledge can be used to facilitate efficient monitoring of the progress of crisis events.

- Due to the time constraints and API restrictions it is often not possible to extract historical information on user behavior. In this context, we investigated whether it was possible to identify users who could be followed with no historical information in Chapter 5. I introduced the concept of Digital First Impression (DFI) of a user which consists of information provided by the user to identify himself to the platform. Through a study of user behavior expressed in a user's DFI I found that it is possible to determine whether a user will be popular and thus followed for access to information.
- During crises, one way to access relevant information efficiently is through information generated from crisis regions. However, the lack of geographical information in the tweets is a challenge. In Chapter 6, I introduce the novel problem of identifying tweets from crisis regions, in the absence of historical information. I addressed this challenge by investigating the user behavior in tweets from crisis regions. I found that tweets published from crisis regions have distinct behavior patterns with respect to the user's intention and the characteristics of the content. I also proposed a method which leverages such differences to efficiently identify tweets lacking geographical information which are generated from crisis regions.

7.2 Future Research Directions

Many of the studies outlined in this thesis introduce novel problems and an attempt at tackling the challenges and problems. However, the frequent usage of microblogging and other social media platforms is creating new avenues for research. Here, I will outline a few directions which can be pursued in the context of the usage of social media for situational awareness towards crisis response.

The integration of different forms of information in a meaningful way is a challenge and an active area of research. In rich social media, this problem is exacerbated by the encouragement of the users by the platforms to publish information in varied forms of text, images, and videos. Identifying and integrating information from these varied sources to acquire "social intelligence" remains a challenge and visual analytics are an intuitive method to tackle this problem. The platforms introduced in Chapter 2 are an attempt towards a solution to these challenges, which employ network, content, and temporal information to achieve this goal. We intend to explore different forms of network information such as the hashtag co-occurrence network and the friendship network to detect dynamic communities during events.

Another challenge in using social media is the lack of information about individuals. Social media data is often considered to be big data due to the population of these networks. However, they can be interpreted as a collection of a very large number of small data, where each individual is a single datum. Often, we are able to acquire a large volume of information about the community as a whole but fail to find sufficient information about individuals to infer their behavior and preferences. Thus approaches to infer user behavior from limited information is of interest to the community. In Chapter 5, I discussed an approach to identify users who may become popular in future. The next step would be to verify if the approach can be generalized to other social media platforms. Additionally, quantifying user behavior such as the number of prediction of the number of likes received by the user's content. Another avenue of research is the application of Digital First Impression to predict other forms of user behavior on social media. For example, the principle of homophily suggests that similar users are more likely to be connected. Here, the similar patterns in their First Impression then could be used to predict the likelihood of two new users to be connected.

REFERENCES

Aggarwal, C. C. and K. Subbian, "Event Detection in Social Streams", in "SDM", pp. 624–635 (2012). 3.5

Agichtein, E., C. Castillo, D. Donato, A. Gionis and G. Mishne, "Finding High-Quality Content in Social Media", in "Proceedings of the 2008 International Conference on Web Search and Data Mining", pp. 183–194 (ACM, 2008). 5.6.3

Allan, J., J. Carbonell, G. Doddington, J. Yamron and Y. Yang, "Topic Detection and Tracking Pilot Study Final Report", Tech. rep. (1998a). 3.1, 3.5

Allan, J., R. Papka and V. Lavrenko, "On-Line New Event Detection and Tracking", in "Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval", pp. 37–45 (ACM, 1998b). 3.5

Aura, S. and G. D. Hess, "What's in a Name?", Economic Inquiry **48**, 1, 214–227 (2010). 5.4

Backstrom, L., E. Sun and C. Marlow, "Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity", in "Proceedings of the 19th International Conference on World Wide Web", pp. 61–70 (2010). 6.5

Baddeley, A. D., N. Thomson and M. Buchanan, "Word Length and the Structure of Short-Term Memory", Journal of Verbal Learning and Verbal Behavior **14**, 6, 575–589 (1975). 5.6.1

Bandari, R., S. Asur and B. A. Huberman, "The Pulse of News in Social Media: Forecasting Popularity", in "Proceedings of 6th AAAI International Conference on Weblogs and Social Media", (The AAAI Press, 2012). 5.4

Barbier, G., R. Zafarani, H. Gao, G. Fung and H. Liu, "Maximizing Benefits from Crowdsourced Data", Computational & Mathematical Organization Theory pp. 1–23 (2012). 4.6

Bastian, M., S. Heymann and M. Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Networks", in "Proceedings of 3rd AAAI International Conference on Weblogs and Social Media", vol. 2 (The AAAI Press, 2009). 2.4

Baur, M. and T. Schank, *Dynamic Graph Drawing in Visone* (Univ., Fak. für Informatik, 2008). 2.4

Becker, H., M. Naaman and L. Gravano, "Learning Similarity Metrics for Event Identification in Social Media", in "Proceedings of the Third ACM International Conference on Web Search and Data Mining", pp. 291–300 (ACM, 2010). 3.5

Blei, D., A. Ng and M. Jordan, "Latent Dirichlet Allocation", JMLR **3**, 993–1022 (2003). 4.4.1

Bowman, S. and C. Willis, “We Media: How Audiences Are Shaping the Future of News and Information”, The Media Center at the American Press Institute (2003). 1

Brandes, U., P. Kenis and J. Raab, “Explanation Through Network Visualization”, Methodology: European Journal of Research Methods for the Behavioral and Social Sciences **2**, 1, 16–23 (2006). 2.4

Carley, K. M., J. Pfeffer, J. Reminga, J. Storrick and D. Columbus, “ORA User’s Guide 2013”, Tech. rep., DTIC Document (2013). 2.4

Castillo, C., M. Mendoza and B. Poblete, “Information credibility on twitter”, in “Proceedings of the 20th International Conference on World Wide Web”, pp. 675–684 (ACM, 2011). 6.2.3

Cataldi, M., L. Di Caro and C. Schifanella, “Emerging Topic Detection on Twitter Based on Temporal and Social Terms Evaluation”, in “Proceedings of the Tenth Int. Workshop on Multimedia Data Mining”, p. 4 (2010). 4.6

Chen, S. F. and J. Goodman, “An Empirical Study of Smoothing Techniques for Language Modeling”, Computer Speech & Language **13**, 4, 359–393 (1999). 5.6.3

Cheng, Z., J. Caverlee and K. Lee, “You Are Where You Tweet: A Content-Based Approach to Geo-Locating Twitter Users”, in “Proceedings of the 19th ACM International Conference on Information and Knowledge Management”, pp. 759–768 (2010). 6.1, 6.3.1

Cheng, Z., J. Caverlee and K. Lee, “A Content-driven Framework for Geolocating Microblog Users”, ACM Trans. Intell. Syst. Technol. **4**, 1, 2:1–2:27 (2013). 6.1, 6.5

Cho, E., S. A. Myers and J. Leskovec, “Friendship and mobility: User Movement in Location-Based Social Networks”, in “Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining”, pp. 1082–1090 (2011). 6.1

Chua, F. C. T. and S. Asur, “Automatic Summarization of Events From Social Media”, Technical Report (2013). 6.4

Cover, T., J. Thomas *et al.*, *Elements of Information Theory* (Wiley Online Library, 1991). 4.5.3

Dykes, J., J. Wood and A. Slingsby, “Rethinking Map Legends with Visualization”, IEEE Transactions on Visualization and Computer Graphics **16**, 6, 890–899 (2010). 2.4

Endres, D. and J. Schindelin, “A New Metric for Probability Distributions”, IEEE Transactions on Information Theory **49**, 7, 1858–1860 (2003). 4.5.3

Ferreira, N., L. Lins, D. Fink, S. Kelling, C. Wood, J. Freire and C. Silva, “BirdVis: Visualizing and Understanding Bird Populations”, IEEE Transactions on Visualization and Computer Graphics **17**, 12, 2374 –2383 (2011). 2.4

Figlio, D. N., “Names, Expectations and the Black-White Test Score Gap”, Tech. rep., National Bureau of Economic Research (2005). 5.4

Fung, G. P. C., J. X. Yu, P. S. Yu and H. Lu, “Parameter Free Bursty Events Detection in Text Streams”, in “Proceedings of the 31st International Conference on Very Large Data Bases”, pp. 181–192 (VLDB Endowment, 2005). 3.4.1, 3.5

Gao, H., G. Barbier and R. Goolsby, “Harnessing the Crowdsourcing Power of Social Media for Disaster Relief”, Intelligent Systems, IEEE **26**, 3, 10–14 (2011). 4.1, 4.6

Guskin, Emily and Hitlin, Paul, “Hurricane Sandy and Twitter”, `http://www.journalism.org/2012/11/06/hurricane-sandy-and-twitter/`, [Online; accessed 4-December-2014] (2012). 1

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, “The WEKA Data Mining Software: An Update”, SIGKDD Explorations Newsletter pp. 10–18 (2009). 5.7.1, 6.3.1

Hecht, B., L. Hong, B. Suh and E. H. Chi, “Tweets From Justin Bieber’s Heart: The Dynamics of the Location Field in User Profiles”, in “Proceedings of the SIGCHI Conference on Human Factors in Computing Systems”, pp. 237–246 (2011). 6.5

Heverin, T. and L. Zach, “Microblogging for Crisis Communication: Examination of Twitter Use in Response to a 2009 Violent Crisis in the Seattle-Tacoma, Washington, Area”, in “Proceedings of the 7th International ISCRAM Conference”, vol. 1 (2010). 6.2.2

Heylighen, F. and J.-M. Dewaele, “Variation in the Contextuality of Language: An Empirical Measure”, Foundations of Science **7**, 3, 293–340 (2002). 5.6.3

Hong, L. and B. Davison, “Empirical study of topic modeling in twitter”, in “Proceedings of the First Workshop on Social Media Analytics”, pp. 80–88 (ACM, 2010). 4.6

Hu, X., L. Tang, J. Tang and H. Liu, “Exploiting Social Relations for Sentiment Analysis in Microblogging”, in “Proceedings of the 6th ACM International Conference on Web Search and Data Mining”, pp. 537–546 (2013). 6.2.4

Hughes, A. L. and L. Palen, “Twitter Adoption and Use in Mass Convergence and Emergency Events”, International Journal of Emergency Management **6**, 3, 248–260 (2009a). 5.1

Hughes, A. L. and L. Palen, “Twitter Adoption and Use in Mass Convergence and Emergency Events”, International Journal of Emergency Management **6**, 3, 248–260 (2009b). 6.2.2

Kalist, D. E. and D. Y. Lee, “First Names and Crime: Does Unpopularity Spell Trouble?”, Social Science Quarterly **90**, 1, 39–49 (2009). 5.4

Keogh, E., S. Lonardi and C. Ratanamahatana, "Towards Parameter-Free Data Mining", in "Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining", pp. 206–215 (ACM, 2004). 3.3.1

Khamadi Were, D., "How Kenya turned to social media after mall attack", http://edition.cnn.com/2013/09/25/opinion/kenya-social-media-attack/index.html?hpt=hp_c4, [Online; accessed 27-January-2014] (2013). 1, 3.1

Kim, S., "Twitter's IPO Filing Shows 215 Million Monthly Active Users", http://abcnews.go.com/Business/twitter-ipo-filing-reveals-500-million-tweets-day/story?id=20460493, [Online; accessed 26-February-2014] (2013). 1

Kireyev, K., L. Palen and K. Anderson, "Applications of Topics Models to Analysis of Disaster-Related Twitter Data", in "NIPS Workshop on Applications for Topic Models: Text and Beyond", (2009). 4.6

Krause, A. and D. Golovin, "Submodular Function Maximization", Tractability Practical Approaches to Hard Problems **3** (2012). 3.2.1

Kumar, S., G. Barbier, M. A. Abbasi and H. Liu, "TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief.", in "Proeedings of 5th AAAI International Conference on Weblogs and Social Media", (The AAAI Press, 2011a). 2.1, 4.6

Kumar, S., F. Morstatter, R. Zafarani and H. Liu, "Whom Should I Follow? Identifying Relevant Users During Crises", in "Proceedings of the 24th ACM Conference on Hypertext and Social Media", pp. 139–147 (ACM, 2013).

Kumar, S., R. Zafarani and H. Liu, "Understanding User Migration Patterns in Social Media", in "AAAI", (2011b).

Kunegis, J., A. Lommatzsch and C. Bauckhage, "The Slashdot Zoo: Mining a Social Network with Negative Edges", in "Proceedings of the 18th International Conference on World Wide Web", pp. 741–750 (ACM, 2009). 5.4

Kwak, H., C. Lee, H. Park and S. Moon, "What is Twitter, a Social Network or a News Media?", in "Proceedings of the 19th International Conference on World Wide Web", pp. 591–600 (ACM, 2010). 6.2.5

Larkey, L. S. and M. E. Connell, "Arabic Information Retrieval at UMass in TREC-10", Tech. Rep. ADA456273, University of Massachussetts (2006). 4.3

Lerman, K. and T. Hogg, "Using a Model of Social Dynamics to Predict Popularity of News", in "Proceedings of the 19th International Conference on World Wide Web", pp. 621–630 (ACM, 2010). 5.4

Li, C., A. Sun and A. Datta, "Twevent: Segment-Based Event Detection from Tweets", in "Proceedings of the 21st ACM International Conference on Information and Knowledge Management", pp. 155–164 (ACM, 2012). 2.4, 3.5

Lin, J., "Divergence Measures Based on the Shannon Entropy", IEEE Transactions on Information Theory **37**, 1, 145–151 (1991). 4.5.3

MacEachren, A., A. Jaiswal, A. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang and J. Blanford, "SensePlace2: GeoTwitter Analytics Support for Situational Awareness", in "2011 IEEE Conference on Visual Analytics Science and Technology (VAST)", pp. 181 –190 (2011). 2.4

Mahmud, J., J. Nichols and C. Drews, "Where Is This Tweet From? Inferring Home Locations of Twitter Users", in "Proceedings of 6th AAAI International Conference on Weblogs and Social Media", (The AAAI Press, 2012). 6.1, 6.3.1, 6.5

Manning, C. D., P. Raghavan and H. Schütze, *Introduction to Information Retrieval*, vol. 1 (Cambridge university press Cambridge, 2008). 5.7.1

Mathioudakis, M. and N. Koudas, "TwitterMonitor: Trend Detection over the Twitter Stream", in "SIGMOD", pp. 1155–1158 (ACM, 2010). 2.4, 4.6

Mendoza, M., B. Poblete and C. Castillo, "Twitter Under Crisis: Can We Trust What We RT?", in "Proceedings of the First Workshop on Social Media Analytics", pp. 71–79 (2010). 3.1, 3.4.1, 4.1, 4.6, 6.5

Miller, R. and N. Lammas, "Social Media and its Implications for Viral Marketing", Asia Pacific Public Relations Journal **11**, 1, 1–9 (2010). 5.1

Morstatter, F., N. Lubold, H. Pon-Barry, J. Pfeffer and H. Liu, "Finding Eyewitness Tweets During Crises", in "Workshop on Language Technology and Computational Social Science", (2014). 6.5

Morstatter, F., J. Pfeffer, H. Liu and K. M. Carley, "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose", Proceedings of 7th AAAI International Conference on Weblogs and Social Media (2013). 2.1, 6.1

NHC, "Post-Tropical Cyclone SANDY", `http://www.nhc.noaa.gov/archive/2012/al18/al182012.update.10300002.shtml` (2012). 2.5.1

Nocke, T., M. Flechsig and U. Bohm, "Visual Exploration and Evaluation of Climate-Related Simulation Data", in "Simulation Conference, 2007 Winter", pp. 703–711 (IEEE, 2007). 2.4

Owoputi, O., B. O'Connor, C. Dyer, K. Gimpel, N. Schneider and N. A. Smith, "Improved Part-Of-Speech Tagging for Online Conversational Text with Word Clusters", in "Proceedings of NAACL-HLT", pp. 380–390 (2013). 5.6.3, 6.2.4

Page, R., "The Linguistics of Self-Branding and Micro-Celebrity in Twitter: The Role of Hashtags", Discourse & Communication **6**, 2, 181–201 (2012). 6.2.4

Pal, A. and S. Counts, "Identifying Topical Authorities in Microblogs", in "Proceedings of the fourth ACM International Conference on Web Search and Data Mining", pp. 45–54 (ACM, 2011). 5.4

Panagiotopoulos, P., A. Z. Bigdeli and S. Sams, "“5 Days in August”–How London Local Authorities Used Twitter during the 2011 Riots", in "Electronic Government", pp. 102–113 (Springer, 2012). 6.5

Paris, C., P. Thomas and S. Wan, "Differences in language and style between two social media communities.", in "Proceedings of 6th AAAI International Conference on Weblogs and Social Media", (The AAAI Press, 2012). 1, 3.1

Pear Analytics, "Twitter Study", `http://www.pearanalytics.com/wp-content/uploads/2012/12/Twitter-Study-August-2009.pdf`, [Online; accessed 27-January-2014] (2009). 3.1

Perreault, M. and D. Ruths, "The Effect of Mobile Platforms on Twitter Content Generation", in "Proceedings of 5th AAAI International Conference on Weblogs and Social Media", (The AAAI Press, 2011). 6.2.2

Petrovic, S., M. Osborne and V. Lavrenko, "Streaming First Story Detection with Application to Twitter", in "Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics", vol. 10 (Citeseer, 2010). 3.4.1, 3.5, 3.6

Poell, T. and E. Borra, "Twitter, YouTube, and Flickr as Platforms of Alternative Journalism: The Social Media Account of the 2010 Toronto G20 Protests", Journalism **13**, 6, 695–713 (2012). 5.5.1

Popescu, A. and M. Pennacchiotti, "Detecting Controversial Events from Twitter", in "CIKM", pp. 1873–1876 (2010). 4.6

Purohit, H. and A. P. Sheth, "Twitris v3: From Citizen Sensing to Analysis, Coordination and Action", in "Proceedings of 7th AAAI International Conference on Weblogs and Social Media", (The AAAI Press, 2013). 2.4

Qu, Y., C. , P. Zhang and J. Zhang, "Microblogging After a Major Disaster in China: A Case Study of the 2010 Yushu Earthquake", in "Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work", pp. 25–34 (2011). 3.1, 4.6

Ramage, D., S. Dumais and D. Liebling, "Characterizing Microblogs with Topic Models", in "Proceedings of 4th AAAI International Conference on Weblogs and Social Media", (The AAAI Press, 2010). 4.6

Rijsbergen, C. J. V., *Information Retrieval* (Butterworth-Heinemann, Newton, MA, USA, 1979), 2nd edn. 3.3

Rout, D., K. Bontcheva, D. Preotiuc-Pietro and T. Cohn, "Where's @Wally?: a Classification Approach to Geolocating Users Based on their Social Ties.", in "Proceedings of the 24th ACM Conference on Hypertext and Social Media", pp. 11–20 (2013). 6.1

Russ, H., "New York, New Jersey Put $71B Price Tag on Sandy", `http://news.msn.com/us/new-york-new-jersey-put-dollar71b-price-tag-on-sandy` (2012). 2.5.1

Sakaki, T., M. Okazaki and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors", in "Proceedings of the 19th International Conference on World Wide Web", pp. 851–860 (2010). 3.1, 3.4.1, 3.5, 4.1, 4.6, 6.5

Sayyadi, H., M. Hurst and A. Maykov, "Event Detection and Tracking in Social Streams", in "Proceedings of 3rd AAAI International Conference on Weblogs and Social Media", (The AAAI Press, 2009). 3.5

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks", Genome research **13**, 11, 2498–2504 (2003). 2.4

Shirky, C., "Power Laws, Weblogs, and Inequality", Clay Shirky's writings about the Internet **8** (2003). 5.4, 5.5.1

Sinnappan, S., C. Farrell and E. Stewart, "Priceless Tweets! A Study on Twitter Messages Posted During Crisis: Black Saturday", in "ACIS 2010 Proceedings", (AIS Electronic Library, 2010). 6.5

Tenenbaum, J., V. De Silva and J. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction", Science **290**, 5500, 2319–2323 (2000). 4.5.3

Verma, S., S. Vieweg, W. J. Corvey, L. Palen, J. H. Martin, M. Palmer, A. Schram and K. M. Anderson, "Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency", in "Proceedings of 5th AAAI International Conference on Weblogs and Social Media", (The AAAI Press, 2011). 6.3.1

Weng, J. and B. S. Lee, "Event detection in twitter", in "Proceedings of 5th AAAI International Conference on Weblogs and Social Media", (The AAAI Press, 2011). 3.5

Wikipedia, "Earthquakes in 2011", `http://en.wikipedia.org/wiki/Earthquakes_in_2011`, [Online; accessed 27-January-2014] (2011). 3.4.1

Wikipedia, "Earthquakes in 2012", `http://en.wikipedia.org/wiki/Earthquakes_in_2012`, [Online; accessed 27-January-2014] (2012). 3.4.1

Wikipedia, "Venezuelan Presidential Election, 2013", `http://en.wikipedia.org/wiki/Venezuelan_presidential_election,_2013`, [Online; accessed 27-January-2014] (2013). 3.4.2

Yang, Y., T. Pierce and J. Carbonell, "A Study of Retrospective and On-Line Event Detection", in "Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval", pp. 28–36 (ACM, 1998). 3.1, 3.4.1, 3.5

Ye, J., J.-H. Chow, J. Chen and Z. Zheng, “Stochastic Gradient Boosted Distributed Decision Trees”, in “Proceedings of the 18th ACM Conference on Information and Knowledge Management”, CIKM ’09, pp. 2061–2064 (ACM, New York, NY, USA, 2009). 5.7.1

Zhao, Q., P. Mitra and B. Chen, “Temporal and Information Flow Based Event Detection From Social Text Streams”, in “AAAI”, vol. 7, pp. 1501–1506 (2007). 3.5

Zhao, W., J. Jiang, J. Weng, J. He, E. Lim, H. Yan and X. Li, “Comparing Twitter and Traditional Media Using Topic Models”, Advances in Information Retrieval pp. 338–349 (2011). 4.6