# Designing and Evaluating Techniques to Mitigate Misinformation Spread on Microblogging Web Services

By
Aditi Gupta

Under the Supervision of Dr. Ponnurangam Kumaraguru

Indraprastha Institute of Information Technology Delhi
March, 2015

©Indraprastha Institute of Information Technology (IIITD), New Delhi
2015

# Designing and Evaluating Techniques to Mitigate Misinformation Spread on Microblogging Web Services

By
Aditi Gupta

Submitted

in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

to the

Indraprastha Institute of Information Technology Delhi
March, 2015

# Certificate

This is to certify that the thesis titled "**Designing and Evaluating Techniques to Mitigate Misinformation Spread on Microblogging Web Services**" being submitted by **Aditi Gupta** to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of Doctor of Philosophy, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

Supervisor Name: Dr. Ponnurangam Kumaraguru,"PK"
March, 2015
Department of Computer Science
Indraprastha Institute of Information Technology Delhi
New Delhi 110 020

# Acknowledgments

My advisor, Ponnurangam Kumaraguru, has helped me immensely throughout my Ph.D. life. All of the research I have conducted in completing my thesis wouldn't have been possible without his vision, encouragement and support. I am extremely fortunate to have him as my advisor.

I would like to thank my collaborators and mentors Prof. Anupam Joshi, Dr. Carlos Castillo and Dr. Patrick Meier for their continued advice and guidance in my research. Their expertise and counsel helped me achieve new milestones during my Ph.D.

I also express my sincere gratitude to all faculty, students and members of IIIT-D, CERC and Precog. The resources and environment provided by them really helped me in my research. They have made my research life at IIIT-D great fun. I would really like to thank Dr. Pankaj Jalote for providing the vision and motivation for doing research.

This thesis would not have been possible without the support of my parents and parents-in-law, sisters and my husband Sameer. Sameer provided encouragement and optimism to me throughout the tough times in my Ph.D.

Without friends it is really difficult for any one to achieve anything; it has not been any different with me. I had many wonderful friends who have made my graduate life a memorable one: Niharika Sachdeva, Anupama Aggarwal and Siddhartha Asthana.

I had a great opportunity to closely work with some brilliant undergraduate students: Hemank Lamba, Mayank Gupta and Nilaksh Das.

Even though others have contributed to my thesis work, I am responsible for the content, conclusions, errors, or omissions in my thesis.

**Abstract**

Online social media is a powerful platform for dissemination of information during important real-world events. Beyond the challenges of volume, variety and velocity of content generated on online social media, veracity poses a much greater challenge for effective utilization of this content by citizens, organizations, and authorities. Veracity of information refers to the trustworthiness / credibility / accuracy / completeness of the content. Over last few years social media has also been used to disseminate misinformation in the form of rumors, hoaxes, fake images, and videos. We aim to address this challenge of veracity or trustworthiness of content posted on social media. The spread of such untrustworthy content online has caused the loss of money, infrastructure and threat to human lives in the offline world. We focus our work on Twitter, which is one of the most popular microblogging web service today.

We provide an in-depth analysis of misinformation spread on Twitter during real-world events. We propose and evaluate automated techniques to mitigate misinformation spread in real-time.

The main contributions of this work are: (i) we analyzed how true versus false content is propagated through the Twitter network, with the purpose of assessing the reliability of Twitter as an information source during real-world events; (ii) we showed the effectiveness of automated techniques to detect misinformation on Twitter using a combination of content, meta-data, network, user profile and temporal features; (iii) we developed and deployed a novel framework for providing indication of trustworthiness / credibility of tweets posted during events. We evaluated the effectiveness of this real-time system with a live deployment used by real Twitter users.

First, we analyzed Twitter data for 25+ global events from 2011-2014 for the spread of fake images, rumors, and untrustworthy content. Some of the prominent events analyzed by us are: Mumbai blasts (2011), England Riots (2011), Hurricane Sandy (2012), Boston Marathon Blasts (2013), Polar Vortex (2014). We identified tens of thousands of tweets containing fake images, rumors, fake websites, and by malicious user profiles for these events. We performed an in-depth characterization study of how this false versus the true data is introduced and disseminated in the Twitter network. Second, we showed how features of meta-data, network, event and temporat from user-generated content can be used effectively to detect misinformation and predict its propagation during real-world events. Third, we proposed and evaluated an automated methodology for assessing credibility of information in tweets using supervised machine learning and relevance feedback approach. We developed and deployed a real-time version in *TweetCred*, a system that assigns a credibility score to tweets. TweetCred, available as a browser plug-in, has been installed and used by 1,808 real Twitter users. During ten months of its deployment, the credibility score for about 12 million tweets

was computed, allowing us to evaluate TweetCred in terms of accuracy, performance, effectiveness and usability.

The system TweetCred built as part of this thesis work is used effectively by emergency responders, firefighters, journalists and general users to obtain credible content from Twitter. This thesis work has shown that measuring credibility of the Twitter content is possible using semi-automated techniques, and the results can be valuable to the real-world users. The insights obtained from this research and deployment provide a basis for building more sophisticated technology to tackle similar problems on different social media.

# Contents

**List of Figures** **vi**

**List of Tables** **ix**

**1 Introduction** **1**

1.1 Aim of thesis . . . . . 1

1.2 Methodology . . . . . 3

1.3 Thesis contribution . . . . . 3

1.3.1 Theoretical . . . . . 4

1.3.2 Real world impact . . . . . 4

1.4 Outline of thesis . . . . . 4

**2 Background** **6**

2.1 Online social media: Volume and impact . . . . . 6

2.2 OSM characteristics . . . . . 7

2.3 Misinformation on online social media during real-world events . . . . . 8

2.4 About Twitter . . . . . 9

2.4.1 Twitter glossary . . . . . 10

2.4.2 Twitter data collection APIs . . . . . 11

**3 Literature review** **15**

3.1 Quality assessment of content posted on OSM . . . . . 15

3.1.1 Emergence of Twitter as a news media . . . . . 16

3.1.2 Spam and phishing detection . . . 16
3.1.3 Trust and credibility assessment . . . 18
3.1.4 Rumor detection . . . 20
3.1.5 Hate and inflammatory content identification . . . 20
3.2 Analyzing Twitter data during real-world events . . . 21
3.2.1 Twitter based early warning systems . . . 21
3.2.2 Extracting situational awareness from Twitter . . . 25
3.3 User Modeling on Online Social Media . . . 26
3.4 Summary . . . 26

**4 Characterizing and detecting fake content** **28**
4.1 Fake content on Twitter during crisis events . . . 28
4.2 Methodology . . . 31
4.2.1 Data collection . . . 31
4.2.2 Temporal and network analysis . . . 32
4.2.3 Building algorithms for fake image detection using meta-data . . . 33
4.3 Results . . . 34
4.3.1 Characterization results . . . 34
4.3.2 Classification results . . . 36
4.4 Discussion . . . 39

**5 Analyzing and predicting propagation of misinformation** **40**
5.1 Propagating misinformation on Twitter . . . 40
5.2 Methodology . . . 42
5.2.1 Data collection . . . 43
5.2.2 Rumors and true tweets . . . 45
5.3 Characterizing fake content propagation . . . 46
5.3.1 Temporal analysis . . . 46
5.3.2 Fake tweet seed user profiles . . . 48
5.3.3 Tweet source analysis . . . 49

5.3.4 Role of user attributes in fake content identification . . . . . 50
5.3.5 Role of user attributes in propagating fake content . . . . . 50
5.4 Suspended profiles analysis . . . . . 53
5.5 Discussion . . . . . 56

**6 Analyzing role of Twitter during India centric events** **59**
6.1 Analyzing Twitter data during India centric events . . . . . 59
6.2 Case study: Mumbai blasts, 2011 . . . . . 60
6.2.1 Data collection . . . . . 61
6.2.2 Analysis . . . . . 62
6.3 Twit-Digest: An online solution for analyzing and visualizing Twitter . . . . . 69
6.3.1 Twit-Digest functionality . . . . . 69
6.4 Discussion . . . . . 72

**7 Identifying and characterizing user communities** **74**
7.1 Identifying communities of users on Twitter during crisis events . . . . . 74
7.2 Theory and methodology . . . . . 75
7.3 Data description . . . . . 77
7.4 Results and evaluation . . . . . 77
7.4.1 Clustering evaluation . . . . . 77
7.4.2 Top users evaluation . . . . . 78
7.4.3 Community characterization . . . . . 78
7.5 Discussion . . . . . 81

**8 Credibility ranking of Tweets during real-world events** **83**
8.1 Using supervised learning methods to assess credibility of content . . . . . 83
8.2 Background on relevance ranking for tweets . . . . . 85
8.3 Methodology . . . . . 86
8.3.1 Data collection . . . . . 87
8.3.2 Events selection . . . . . 87
8.3.3 Annotation scheme . . . . . 88

8.4 Analysis . . . 90
8.4.1 Types of features . . . 91
8.4.2 Pseudo relevance feedback . . . 92
8.4.3 Evaluation metric . . . 93
8.5 Experimental results . . . 93
8.5.1 Regression analysis . . . 94
8.5.2 Evaluation of ranking . . . 94
8.6 Discussion . . . 95

**9 TweetCred: A tool to measure credibility on Twitter in real-time 98**
9.1 Building real-time credibility model for Twitter . . . 98
9.2 Training data collection . . . 100
9.3 Credibility modeling . . . 102
9.3.1 Feature extraction . . . 103
9.3.2 Learning scheme . . . 104
9.3.3 Evaluation metric . . . 105
9.3.4 Evaluation . . . 105
9.4 Implementation and performance evaluation . . . 107
9.4.1 Implementation . . . 107
9.4.2 Back-end . . . 107
9.4.3 Front-end . . . 109
9.4.4 Response time . . . 111
9.5 User testing . . . 111
9.5.1 User feedback . . . 112
9.5.2 Credibility rating bias . . . 113
9.5.3 Usability survey . . . 114
9.5.4 User comments . . . 114
9.6 Discussion . . . 115

**10 Conclusion and future work 121**

10.1 Summary . . . . . 121
10.1.1 Characterizing misinformation and fake content . . . . . 121
10.1.2 Effectiveness of automatic methods to identify misinformation . . . . . 122
10.1.3 Design and build systems to assess credibility of content on Twitter . . . . . 123
10.2 Future work and limitations . . . . . 124

**Bibliography** **125**

**List of publications** **135**

**Appendix** **137**

# List of Figures


2.1 Sample tweets containing rumors posted during the England riots of 2011. . . . . 8

2.2 A tweet by a fake account @_Boston Marathon, created during Boston marathon blasts, 2013. . . . . 9

2.3 Sample Twitter profile of a user. . . . . 10

2.4 A tweet by user CNN Breaking News (username: @cnnbrk). . . . . 12

3.1 A Chrome based browser plugin to detect phishing URLs on Twitter. . . . . 17

3.2 A real-time system to study information diffusion on Twitter. . . . . 19

3.3 The Information Integration Landscape. . . . . 22

3.4 Screenshot of the Emergent system, which detects true, false and unverified topics. . 23

3.5 Social network analysis performed by Cheong et al. on Twitter dataset during 2011 Australian floods. . . . . 24

4.1 Some of the fake pictures of Hurricane Sandy. . . . . 30

4.2 Spatial distribution of tweets on Hurricane Sandy. . . . . 32

4.3 Details of data collected for the tweets with fake images. . . . . 36

4.4 CDF of retweets of the fake image tweets by the users. . . . . 37

4.5 Spread of tweets with fake images (retweet and reply graph). . . . . 38

5.1 Two sample tweets containing fake content. . . . . 42

5.2 A picture clicked during the Boston marathon blasts. . . . . 43

5.3 Architecture diagram describing the methodology followed in this chapter. . . . . 43

5.6 The log distribution for the number of the total, fake, true information and NA category tweets. . . . . 48

5.7 Tweet source analysis for the three categories of tweets. . . . . 49
5.8 Distribution of user attributes in fake content identification. . . . . 51
5.9 Regression results of the overall impact of the users in previous time quantum. . . . 53
5.10 Number of suspended profiles created in the hours after the blast. . . . . 54
5.11 Network of suspended accounts (retweets / replies / mentions) created during the Boston blasts. . . . . 55
5.12 Tag cloud of user description and tweets posted by the suspended accounts. . . . . 56

6.1 The first tweets captured by us about the attacks after the Mumbai blasts. . . . . 61
6.2 Tag clouds showing the vocabulary analysis for the tweets. . . . . 63
6.3 Distribution of tweets by users during the Mumbai blasts. . . . . 65
6.4 The variation in average number of friends and followers for the top 500 users. . . . 66
6.5 The activity regarding the rumors of 13th July (13:00 to 23:59 hours) being Kasab's birthday through the network of Twitter. . . . . 67
6.6 For each tweet credibility score is computed and tweets are marked as spam or not spam. . . . . 70
6.7 Sample geographical distribution of the tweets and users corresponding to a query. . 71
6.8 The network graph, with users as nodes and edges representing the retweets and replies by one user to another. . . . . 72
6.9 The tag-cloud of hashtags present in the tweets containing the query word. . . . . 72

7.1 The architecture diagram for the methodology followed. . . . . 76
7.2 The graph shows the Mean Average Precision score for the top central users' detection for the three crisis events. . . . . 79

8.1 Sample tweets in our dataset for the event Hurricane Irene. . . . . 84
8.2 Methodology and analysis performed in this chapter. . . . . 86
8.3 Screenshot of annotation portal. It shows the options presented to the annotators per tweet. . . . . 90
8.4 Performance Evaluation of ranking . . . . . 96

9.1 Screenshot of timeline of a Twitter user when TweetCred browser extension is installed. 99
9.2 Architecture diagram for this research work. . . . . 100

9.3 Screenshot of the annotation task using CrowdFlower. . . . . 103
9.4 Data flow steps of the TweetCred extension and API. . . . . 108
9.5 The interface webpage to access the TweetCred API. . . . . 109
9.6 User feedback. . . . . 110
9.7 CDF of response time of TweetCred. . . . . 111
9.8 Distribution of number of followers and friends of TweetCred users. . . . . 112
9.9 Distribution of credibility scores. . . . . 114
9.10 Chrome Web Store page of TweetCred . . . . . 116
9.11 Posts on Twitter about the TweetCred Chrome Extension . . . . . 117
9.12 Article in Washington Post about TweetCred System . . . . . 118
9.13 Article in NewYorker Magazine about TweetCred System . . . . . 119
9.14 Article in DailyDot about TweetCred System. . . . . 120

# List of Tables


2.1 The complete list of Twitter API rate limits, it shows the allowed number of requests by an application. . . . . 14

4.1 Descriptive statistics of the Twitter dataset for Hurricane Sandy. . . . . 31

4.2 Descriptive statistics of the tweets with fake and real images URLs. . . . . 33

4.3 User and tweet based features used for classification of fake and real images of Sandy. 35

4.4 Results of the Algorithm *Compute_overlap*. . . . . 36

4.5 Classification results for tweets containing fake image and real images. . . . . 39

5.1 Descriptive statistics of dataset for Boston blasts, April 15th - 19th, 2013. . . . . 44

5.2 Top 20 most popular tweets (retweeted and replied) posted during Boston Marathon blasts. . . . . 47

5.3 Descriptive statistics of the four user accounts that were the seeds of the six fake content tweets. . . . . 49

6.1 Some prominent India centric events analyzed by us. . . . . 60

6.2 Descriptive statistics for the Twitter dataset for 13th, 14th and 15th July, 2011. . . 62

6.3 Top 10 users in the user network for tweets on Mumbai blasts by betweenness, closeness and degree centrality. . . . . 68

7.1 Data statistics for the three crisis events. . . . . 77

7.2 Top 20 words from each community (cluster) for the three crisis events. . . . . 80

7.3 Screen name and user description of top users in Community 2 and 3. . . . . 81

8.1 Descriptive statistics of the Twitter dataset. . . . . 87

8.2 Fourteen events selected for the time period from 12 July to 30 August, 2011. . . . . 89

8.3 Content and source based features that were considered for ranking. . . . . 91

9.1 Number of tweets and distinct Twitter users from which data was collected for the purposes of creating a training set. . . . . 101

9.2 Distribution of labels over tweets. . . . . 102

9.3 Features used by the credibility model. . . . . 104

9.4 Evaluating ranking algorithms in terms of Normalized Discounted Cumulative Gain (NDCG) and execution times. . . . . 106

9.5 Performance of SVM-rank algorithm in credibility ranking of tweets using 2011 and 2013 data. . . . . 106

9.6 Top 10 features obtained using SVM-rank for ranking tweets according to their credibility. . . . . 107

9.7 Summary statistics for the usage of TweetCred. . . . . 112

9.8 Feedback given by users of TweetCred on specific tweets. . . . . 113

# Chapter 1

# Introduction

## 1.1 Aim of Thesis

Online Social Media (OSM) has emerged as one of the most effective communication and information sharing platforms today. Some sample users of OSM services are head of states, journalists, organizations, celebrities and citizens. OSM is a platform, which spans across barriers of country, region, religion, race and language; it has revolutionized the way people access information and news about current events. Some of the popular OSM are Facebook, Twitter, YouTube, Google+, etc. [1] Many of these online social media services have their own unique purpose, attributes and user base. Huge volume of content is generated on OSM every day, for e.g. 300 hours of video are uploaded to YouTube every minute and more than 500 million tweets are posted on Twitter every day. [2] [3] Any content posted on OSM has the potential to reach millions of users within seconds. Over past few years, media and researchers have taken a key interest in analyzing the content and attributes of OSM activity during real-world events such as elections, sports tournaments, earthquakes, floods, and blasts.

Volume, variety, velocity and veracity of data in online social media pose a major challenge in effective utilization of this data by organizations, authorities, and citizens. Veracity of data refers to the trustworthiness / credibility / accuracy / completeness of the content posted on online social media. Over last few years, social media has been used to disseminate misinformation in the form of rumors, hoaxes, fake images, and videos. We aim to address this challenge of trustworthiness of content posted on social media services. The spread of such untrustworthy content online has caused losses of money, infrastructure and poses a threat to human lives in the offline world. During

[1] `https://www.facebook.com/`, `https://twitter.com/`, `https://www.youtube.com/`, `https://plus.google.com/`

[2] `http://www.youtube.com/yt/press/statistics.html`

[3] `https://about.twitter.com/company`

crisis and emergency events, people are often more susceptible to fall for rumors / fake content due to heightened anxiety and emotional vulnerabilities. In one of the most recent incidents in U.S.A., Dow Jones index plunged 140 points due to a rumor tweet posted from a news agency's (Associated Press) Twitter account [21]. The estimated temporary loss of market cap in the S&P 500 totaled $136.5 billion. The rumor mentioned that U.S.A. President Barack Obama has been injured in twin explosions at the White House. In case of England Riots, online social media was responsible for spreading and instigating violence amongst people. Many rumors propagated during the riots, which resulted in large-scale panic and chaos among the citizens [79]. Two people were also sentenced for spreading false posts on Facebook during the riots [17]. In another incident in Venezuela, some people had spread rumors on Twitter, to destabilize the banking system of the country [111]. In one of the extreme case, Twitter terrorists in Mexico were given thirty years sentence for spreading rumors about a fake shooting by gunmen in schools [2]. As parents rushed to get their children from school the incorrect information caused a massive chaos in the city, which led to 26 road accidents and jammed telephone lines. In the case of Boston marathon blasts, the rumors, resulted in fake and fraud charity fund creations and incorrect news about a young boy dying in the blasts [77]. In another incident in India, OSM was used during ongoing communal riots to spread inflammatory and provoking false content against the government [95]. During the widespread outbreak of Ebola last year, there was a hoax on online social media that salt water can prevent or cure Ebola, this false information led to many deaths [75].

We focus most of our work on microblogging services such as Twitter; it is a microblog web service with over 600 million users all across the globe. Twitter has gained a reputation over the years as a prominent news source, often disseminating information faster than traditional news media. Users on Twitter, create their public / private profile and post messages (also referred as tweets or statuses) via the profile. The maximum length of the tweet can be 140 characters. Each post on Twitter is characterized by two main components: the tweet (content and associated metadata) and the user (source) who posted the tweet. Studies have explored and highlighted the role of Twitter as a news media and a platform to gauge public sentiments [16, 19]. Millions of tweets are posted on Twitter everyday; for our work we analyze tweets generated during real-world events. The opening match of Soccer world cup in 2014 between host country Brazil and Croatia recorded about 12.2 million tweets while the match was being played [9]. During crisis events such as snow storms, floods and bomb blasts, tweets are posted regarding information such as updates, pictures or videos, help requests, condolences and volunteer messages.

One major difference between dissemination of news or information through traditional media and Twitter is that, Twitter is a crowd-sourced medium. In contrast to television or print or news websites where the source of information are few and known (i.e. credible), users on Twitter act as its sensors, and fill in the information gap about an event. Due to the anonymous and unmonitored nature of the Internet, a lot of content generated on Twitter maybe incredible. During an event,

when a user types a query in the Twitter search (e.g. UK riots) or clicks on a related trending topic (e.g. #ukriots), all tweets matching the query words are displayed to the user. The search results display tweets ordered from top to bottom, in reverse chronological order. When an event of a sizable magnitude and impact occurs, thousands of tweets are posted per hour [9, 65, 82]. The information shared and accessed on social media such as Twitter, is in real-time, the impact of any malicious intended activity, like spreading fake images and rumors need to be detected and curbed from spreading immediately. Filtering out spam, phishing content, detecting rumors, extracting credible information, are some of the research problems that address the challenges of analyzing trustworthiness information on Twitter. The aim of this research is to use automated methods to build real-time solutions for mitigating misinformation spread on microblogging websites.

## 1.2 Methodology

To achieve the objective of mitigating misinformation spread on Twitter, we first established a round the clock data collection framework using Twitter APIs. Using this data, we analyzed multiple events from 2011-14. For most of our analysis, we considered crisis events like earthquakes, hurricane, bomb blasts, etc. In total we collected about 4 Billion tweets occupying 5.6 TeraBytes of disk space. Data statistics for some of the events: Boston Marathon Blasts (7,888,374 tweets by 3,677,531 users), Hurricane Sandy (1,782,526 tweets by 1,174,266 users), Washington Navy Yard shootings (484,609 tweets by 257,682 users). We first identified different forms of misinformation such as fake images, rumors, spam and malicious URLs. We characterized various demographic, temporal, network and other attributes of such content in Twitter network. Next, we explored and evaluated the use of supervised and unsupervised techniques such as classification, clustering, ranking and regression to distinguish fake content from the real content. The insights obtained from above experiments led us to formulate and design a real-time solution that can help mitigate misinformation spread on Twitter. We envisioned that such a system should provide an indication to Twitter users about trustworthiness of tweets in real-time within the environment of Twitter ecosystem. With this aim in mind, we developed and deployed our system *TweetCred* with real Twitter users.

## 1.3 Thesis Contribution

This thesis is both timely and needed to reduce the negative consequences of misinformation propagation in online social media. Results from this work can be used to build solutions for the problem of trustworthiness of user-generated content on different online social networks. The insights obtained and system built by us are effectively used by emergency responders, firefighters, journalists

and general users to identify trustworthy content on Twitter. We present how computational techniques and methods can be deployed in a usable system for real users to help them make judgments about trustworthiness of a post.

### 1.3.1 Theoretical

- Analyzed how real and fake content is propagated through the Twitter network, with the purpose of assessing the reliability of Twitter as an information source during real-world events.
- Evaluated the effectiveness of automatic methods to detect misinformation on Twitter using a combination of content, meta-data, network, user profile and temporal features.
- Developed and deployed a novel framework for providing indication of trustworthiness / credibility of tweets posted during events. Evaluated the framework for accuracy, performance and usability.

### 1.3.2 Real World Impact

- The real-time system *TweetCred* built as part of this thesis to assess credibility of content on Twitter is used by 1,808 real Twitter users. It had computed and displayed credibility scores for more than 14.2 million tweets.
- A unique data set of thousands of fake images, rumor tweets and malicious profiles for 25+ real-world events. We have shared this data with multiple researchers across the globe to aide in their research.
- An online framework to collect, analyze and visualize Twitter data called *Twit-Digest* was built and installed at five government agencies.

## 1.4 Outline of Thesis

The next chapter discusses the fundamentals of Twitter, challenges of doing OSM based research and real-world examples of misinformation spread on Twitter. Chapter 3 discusses the relevant literature in assessing trustworthiness of content on OSM and analyzing real-world events. In Chapter 4 we present the analysis of characterizing and detecting tweets containing fake images on Twitter during Hurricane Sandy. Chapter 5 discusses the propagation of rumors and true news tweets during Boston Marathon blasts and demonstrates the use of linear regression to predict virality of rumors on Twitter. The focus of Chapter 6 is on detecting rumors and fake content

posted on Twitter in events related to India. Chapter 7 demonstrates the use of spectral clustering in identifying user communities on Twitter during crisis events. Chapter 8 presents a novel algorithm to assess and measure credibility of content on Twitter in posthoc environment. Chapter 9 discusses the design, implementation and evaluation of *TweetCred*. Finally, Chapter 10 presents conclusions from this thesis work.

# Chapter 2

# Background

In this chapter we discuss the various Online Social Media (OSM), the characteristics of OSM, and real-world examples of misinformation spread on OSM. For the better understanding of the work done in rest of this thesis, we would also describe the Twitter social network, its terminology, attributes and data collection APIs.

## 2.1 Online Social Media: Volume and Impact

Online social media provides a platform for its users to interact, share information, opinions and updates with one another. According to recent statistics, 74% of online adults use OSM services [18]. Many different kinds of OSM exist today: Twitter is a microblogging website, [1] Facebook is primarily used by people to connect with their friends, [2] LinkedIn is a social media where users's create a professional network, [3] YouTube is used by people to share videos, [4] and Pinterest to share pictures. [5] There are many similarities and differences between these different services. Data posted on OSM has the potential to reach millions of people within minutes. Real-world events result in a huge amount of corresponding data being generated on OSM. Mining this data created during an event can be useful to authorities, journalists and people affected. During last few years, OSM services have been used successfully to improve disaster response [62]. FEMA (Federal Emergency Management Agency) reported that during Hurricane Sandy, even when the telephone and mobile lines where down during the peak of the storm, more than 20 million tweets were sent by users about the hurricane [31]. OSM is also effectively used for branding, promotion, engagement and sharing content from various domains like politics, sports and entertainment. The FIFA world cup,

[1] `https://twitter.com/`
[2] `https://www.facebook.com/`
[3] `https://www.linkedin.com/`
[4] `https://www.youtube.com/`
[5] `https://pinterest.com/`

2014 broke most OSM records with the huge volume of content posted: Facebook reported that 350 million people interacted about the event on its service generating 3 billion posts [9], it became the most tweeted event with over 672 million tweets posted [83]. OSMs are often considered as a network, where the a user profile, a message post, a video or a picture is a node and the relationship (friends, followers, like, retweet, share, etc.) between the nodes form the links. User generated content on OSM are the user profiles, status messages, posts, videos or pictures uploaded by users. Information that is posted on these OSM services spreads very quickly in the same service and / or across services.

## 2.2 OSM Characteristics

In this section, we discuss some of the characteristics of OSM. All of these characteristics are also challenges for any research based on microblogging websites such as Twitter.

- *Volume*: Most of the popular online social websites have users of the order of hundreds of millions. A huge amount of content is generated every second, minute and hour of the day. Any algorithms or solutions build to analyze OSM data should be scalable enough to order content and user data upto the order of millions and billions.
- *Velocity*: Impact of malicious activities in online social media, such as spread of spam, phishing or rumors, causes vast amount of damage within hours of being introduced in the media. Hence, solutions and algorithms build need to be able to solve and detect such content in real-time. Post-analysis may be able to capture concerned criminals, but would not be able to contain the damage.
- *Variety*: Unlike websites and online news articles, users on OSM often have space limitations (like a tweet can be of maximum 140 characters long on Twitter) and is of informal nature. In addition, use of mobile phones to access social media websites, have resulted in people using mostly slang and shorthand language in posting content. Content generated by formal organizations and professionals are of well formed English or any other language. Standard tools such as Stanford's NLP parser do not perform well on OSM data. Thus, there is a dire need to construct special and customized toolkits to analyze OSM data.
- *Veracity*: Another key challenge on online social networks is determining the truthfulness or trust on content posted on them. Internet, by itself is an anonymous medium. Lack of tracing technologies, distributed framework and private corporation monopoly worsens the condition. Anybody can create one or multiple profiles on most OSM, with little or no verification required. Hence, validation and authenticity of any information source on OSM is extremely challenging.

In our work, we aim to enhance computational techniques to address the fourth challenge of *veracity* of content and build solutions that are scalable and real-time to address the challenges of *volume* and *velocity* respectively. In order to make our solution robust enough to handle *variety* of data we use large number of attributes of data and meta-data of content on OSM.

## 2.3 Misinformation on Online Social Media during Real-world Events

The content on OSM can provide rich information about the event, however, this vast resource of information is often diluted with misinformation such as rumors, spam, fake and other malicious content. The unmonitored and anonymous nature of online social media makes it difficult to assess the quality, accuracy or trustworthiness of the information. Spread of misinformation on OSM can often have adverse effects on ground to real people. On the other hand, extracting timely, actionable and accurate information can be very useful in dealing with high impact events.

Thousands of posts, pictures and videos are uploaded on OSM during a real-world event, but a lot of this content is fake and malicious. These can be in form of text messages, pictures or videos. Figure 2.1 contains some rumor tweets posted during the England riots of 2011. In our work, we analyze various types of misinformation spread on OSM during real-world events. Fake or fraudulent user profiles are created by users during real-world events. Figure 2.2 shows a tweet by a fake user profile on Twitter posted during Boston marathon blast. The tweet received more than 50,000 retweets.

Figure 2.1: Rumor tweets posted during the England riots of 2011. Misinformation like the wild animals from the zoo being freed caused immediate panic among the locals of the area. False news such as a 16 year old girl being beaten by police, fueled the riots further.

Over recent years OSM have also been used to spread hate or inflammatory content. Such content

Figure 2.2: A tweet by a fake account @_Boston Marathon, created during Boston marathon blasts, 2013. The tweet asked for retweets in order to collect donations to victims. The account was later deleted by Twitter. The tweet received more than 50,000 retweets.

if propagated during some existing real life volatile situations can have major adverse implications. Media such as Twitter, which is a microblog is most suited for dissemination and sharing news based information, since it is mostly public, and gives a wide range of audience for the content posted. Hence, we focus on Twitter as our medium of study for this thesis. More about misinformation on Twitter is provided in rest of the document.

## 2.4 About Twitter

Twitter is one of the largest online social media service today. Twitter is a microblogging service, which has gained popularity as a major news source and information dissemination agent over last few years. Users on Twitter, create their public / private profile and post messages (also referred as tweets or statuses) via the profile. The maximum length of the tweet can be 140 characters. Each post on Twitter is characterized by two main components: the tweet (content and associated metadata) and the user profile (source) who posted the tweet. Users can subscribe to other user's tweets using the 'follow' feature As per the 2015 official Twitter statistics, it has more than 288 million monthly active users and around 500 million Tweets are sent per day. [6] Twitter supports 33 different languages. Katty Perry is the most followed user of Twitter with approximately 67 million followers, which implies that any tweet by her can reach millions of her followers within seconds. [7] Figure 2.3 shows a sample profile of a user on Twitter. Figure 2.4 shows a sample tweet by *CNN Breaking News* user account on Twitter.

[6] `https://about.twitter.com/company`
[7] `https://twitter.com/katyperry`

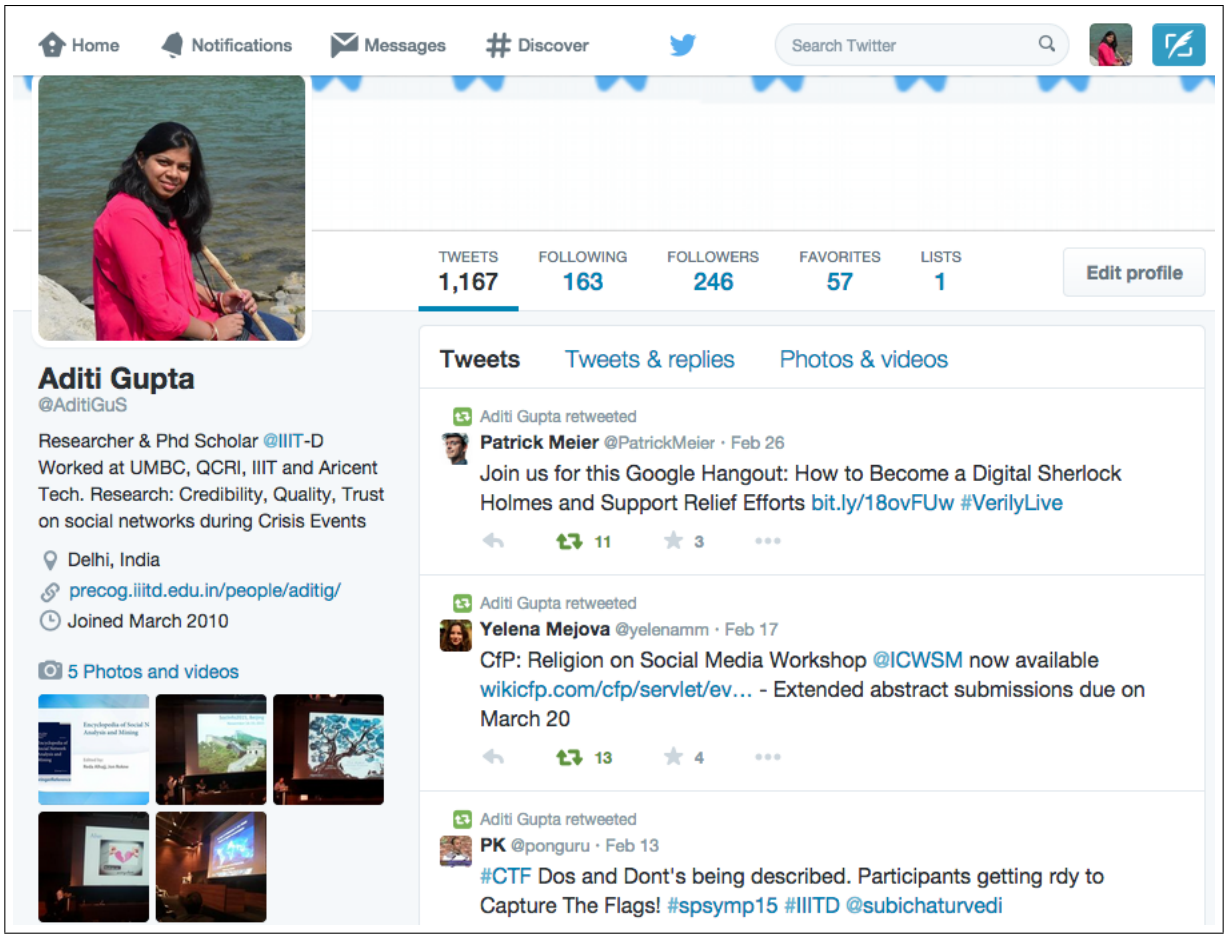



Figure 2.3: Sample Twitter profile of a user, the left column contains the user details like username, user description, website, etc. and the right part contains the tweets posted or retweeted by the user.

### 2.4.1 Twitter Glossary

Now we present a brief glossary of Twitter terminology quoted / adapted mostly from official Twitter website. [8]

- *Tweet*: A tweet is a message posted on Twitter which can be of maximum 140 characters long. Each tweet has attributes like time of creation and a corresponding user who posted the tweet.
- *User Profile*: A user's profile on Twitter displays information that they choose to share publicly, as well as all of the Tweets they have posted. A user's profile along with his @username identifies him on Twitter.

[8] `https://support.twitter.com/articles/166337-the-twitter-glossary`

- *Follower / Friend*: A *user A* is said to be a follower of *user B* if *user A* has subscribed receive *user B*'s Tweets in his home stream. In this case, *user B* is considered *user A*'s friend.
- *Retweet*: It is an action in which a user shares another user's Tweet to all of his followers by clicking on the Retweet button.
- *Hashtag*: A hashtag is any word or phrase immediately preceded by the # symbol. When a user clicks on a hashtag, all tweets containing the same keyword or topic are shown.
- *Mention*: It is used to explicitly refer / include a user in a tweet. 'username' is said to be a mention to a user in a tweet.
- *List*: A user can create a group list of other Twitter users by topic or interest (e.g., a list of friends, coworkers, celebrities, athletes). Twitter lists also contain a timeline of Tweets from the specific users that were added to the list, offering the users a way to follow individual accounts as a group on Twitter.
- *Favorite*: A user can tap the star icon to favorite a Tweet and the author of the tweet will see that he liked it.
- *Username*: Unique name string assignment to each user profile on Twitter.
- *Reply*: A response to another user's Tweet that begins with the @username of the person that a user is replying to is known as a reply. Reply by clicking the "reply" button next to the Tweet one would like to respond to.
- *Verified Accounts*: A process whereby a Twitter account receives a blue check icon to indicate that the creator of these Tweets is a legitimate source. Verified users include public figures and those who may have experienced identity confusion on Twitter.
- *Trends*: A Trend is a topic or hashtag determined algorithmically to be one of the most popular on Twitter at that moment. A user can choose to tailor Trends based on his location and who he follows.
- *Direct Messages*: Direct Messages are private messages sent from one Twitter user to another Twitter users. A user can use Direct Messages for one-on-one private conversations, or between groups of users.

### 2.4.2 Twitter Data Collection APIs

Twitter provides an extensive framework of APIs and their corresponding documentation to aid researchers and developers in building applications, embedding Twitter components in websites

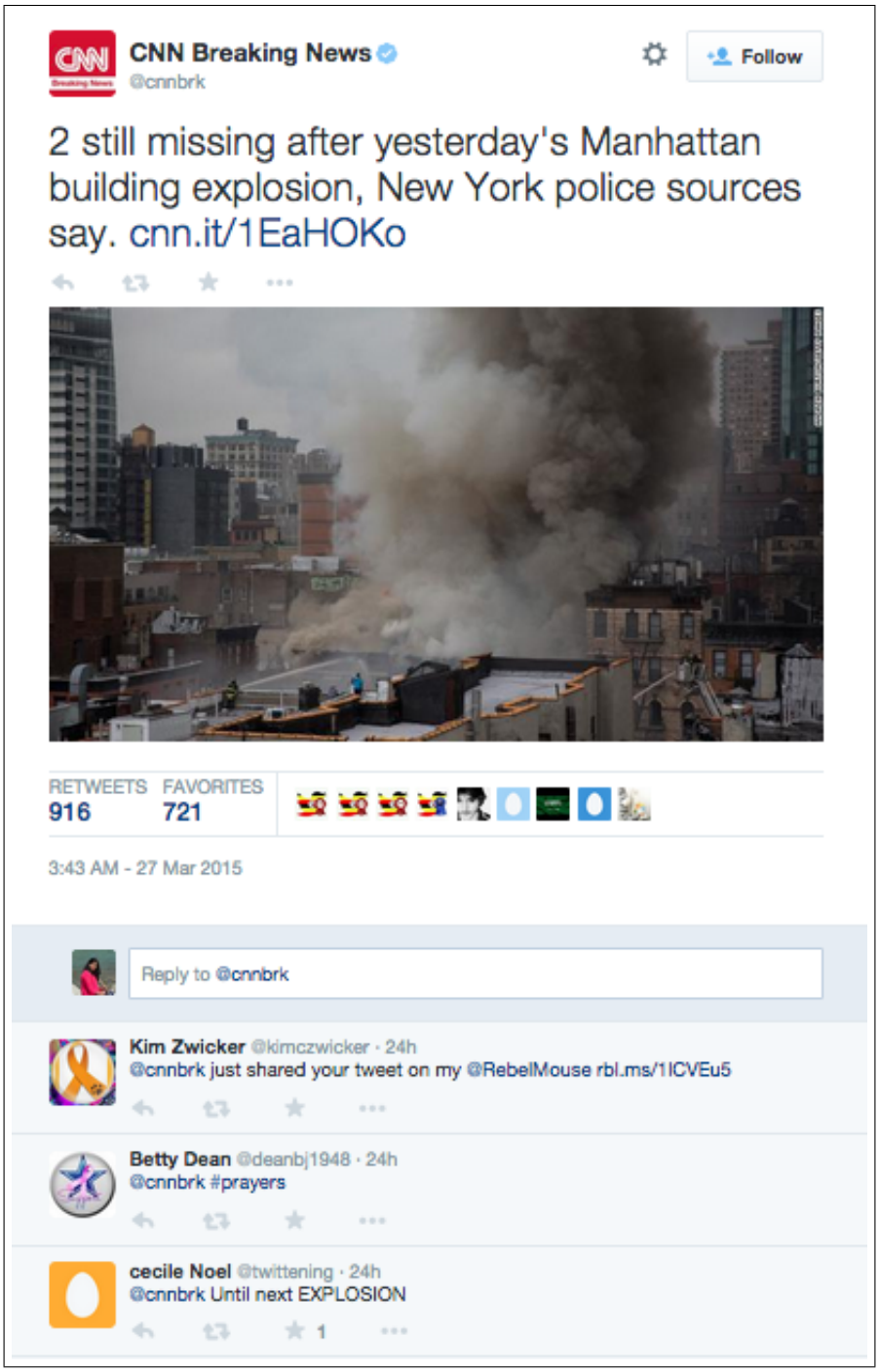



Figure 2.4: A tweet by user CNN Breaking News (username: @cnnbrk) containing an embedded image and replies (or comments) by other users. The tweet is retweeted by 961 users and favorited by 721 users.

and collecting data. [9] The OAuth endpoints provided by Twitter are used by users to connect to

[9] `https://dev.twitter.com/`

Twitter and send secure and authorized requests to the Twitter API. In order to collect data, users create applications on Twitter; for each application a user receives an API access token. There are two modes of authentication in Twitter API: application-user and application-only authentication. In application-user authentication the request authenticates the application and the corresponding user on whose behalf the request is made. It is the most common form of authentication used. In application-only authentication only the application makes request on its own behalf, without the user context. The data provided the APIs is mostly rate limited by time, based on the number of requests from a particular application access token. Table 2.1 presents the complete details of rate limits of Twitter API.

**REST APIs** Twitter data can be read or written programmatically using the REST APIs provided by Twitter. The APIs use OAuth for authentication and JSON format is used for the response by the API. There are two main types of Twitter APIs used for tweets data collection by developers and researchers.

- *Search APIs*: The search API provides historical tweets for the given set of keyword(s). The results can be further filtered based on location and language of tweets. We can provide a maximum of ten keywords and operators at a time. The request returns either last 1500 tweets or tweets posted during last 7 days, whichever is less. [10]
- *Streaming APIs*: The Streaming APIs provides real-time stream of tweets. To filter which tweets are returned is determined by a comma-separated list of phrases are mentioned . As stated by Twitter documentation: "a phrase may be one or more terms separated by spaces, and a phrase will match if all of the terms in the phrase are present in the Tweet, regardless of order and ignoring case. By this model, you can think of commas as logical ORs, while spaces are equivalent to logical ANDs (e.g. 'the twitter' is the AND twitter, and 'the,twitter' is the OR twitter)."

We collected data from Twitter corresponding to tweets about a particular event or topic. We obtained the latest tends from Twitter using the *Trends API* by Twitter. For each tweet, we get the following data: tweet details like tweet content, creation date, meta-data, etc. and user details like profile creation date, number of followers, lists, etc. A sample complete tweet JSON is given in Appendix 10.2. Each tweet is uniquely identified by a Tweet ID. We stored our data in MySQL database.

[10] As per the Twitter API 1.0. This is subject to change by Twitter API updates.

Table 2.1: The complete list of Twitter API rate limits, it shows the allowed number of requests by an application.

| Title | Resource family | Requests / 15-min window (user auth) | Requests / 15-min window (app auth) |
|---|---|---|---|
| GET application/rate_limit_status | application | 180 | 180 |
| GET favorites/list | favorites | 15 | 15 |
| GET followers/ids | followers | 15 | 15 |
| GET followers/list | followers | 15 | 30 |
| GET friends/ids | friends | 15 | 15 |
| GET friends/list | friends | 15 | 30 |
| GET friendships/show | friendships | 180 | 15 |
| GET help/configuration | help | 15 | 15 |
| GET help/languages | help | 15 | 15 |
| GET help/privacy | help | 15 | 15 |
| GET help/tos | help | 15 | 15 |
| GET lists/list | lists | 15 | 15 |
| GET lists/members | lists | 180 | 15 |
| GET lists/members/show | lists | 15 | 15 |
| GET lists/memberships | lists | 15 | 15 |
| GET lists/ownerships | lists | 15 | 15 |
| GET lists/show | lists | 15 | 15 |
| GET lists/statuses | lists | 180 | 180 |
| GET lists/subscribers | lists | 180 | 15 |
| GET lists/subscribers/show | lists | 15 | 15 |
| GET lists/subscriptions | lists | 15 | 15 |
| GET search/tweets | search | 180 | 450 |
| GET statuses/lookup | statuses | 180 | 60 |
| GET statuses/oembed | statuses | 180 | 180 |
| GET statuses/retweeters/ids | statuses | 15 | 60 |
| GET statuses/retweets/:id | statuses | 15 | 60 |
| GET statuses/show/:id | statuses | 180 | 180 |
| GET statuses/user_timeline | statuses | 180 | 300 |
| GET trends/available | trends | 15 | 15 |
| GET trends/closest | trends | 15 | 15 |
| GET trends/place | trends | 15 | 15 |
| GET users/lookup | users | 180 | 60 |
| GET users/show | users | 180 | 180 |
| GET users/suggestions | users | 15 | 15 |
| GET users/suggestions/:slug | users | 15 | 15 |
| GET users/suggestions/:slug/members | users | 15 | 15 |

# Chapter 3

# Literature Review

In this chapter, we would discuss and present the current state of the art in research for analyzing and measuring trustworthy information from Twitter. Section 3.1 discusses the research work to assess, measure, quantify and detect good quality content from Twitter. In section 3.2, we would present the research done to characterize the role of Twitter during real world events. In the last section, we summarize the implications and research gaps in analyzing trustworthy information from Twitter during real-world events.

## 3.1 Quality Assessment of Content Posted on OSM

This section presents the research work done in the space of extracting and analyzing trustworthy and credible information from Twitter during real world events. One major challenge in consuming content from Twitter is that it is difficult to filter out good quality content from the large volume of content created. The quality of content on Twitter is polluted with the presence of phishing, spam, advertisements, fake images, rumors and inflammatory content. Media such as Twitter, which is a micro-blog is more suited for dissemination and sharing news based information, since it is mostly public, and gives a bigger range of audience for the content posted. Hence, majority of the work discussed in this survey, is centered around Twitter. Researchers have used various classical computational techniques such as classification, ranking, characterization and conducting user studies, to study the problem of trust on Twitter. Some of the researchers who applied various kinds of classifiers (Naive Bayes, Decision Tree, SVM) to identify spam, phishing and not credible content on Twitter, using message, user, network and topic based features on Twitter [4, 16, 41]. Ranking algorithms have been applied and fine tuned by researchers for questions pertaining to trust related problems such as credibility and spam [2, 40]. Each of the above mentioned work are

discussed in detail, later in the chapter.

### 3.1.1 Emergence of Twitter as a News Media

Computer science research community has analyzed relevance of online social media, in particular Twitter, as news disseminating agent. Kwak et al. showed the prominence of Twitter as a news media, they showed that 85% topics discussed on Twitter are related to news [59]. Their work highlighted the relationship between user specific parameters v/s the tweeting activity patterns, like analysis of the number of followers and followees v/s the tweeting (re-tweeting) numbers. Zhao et al. in their work, used unsupervised topic modeling to compare the news topic from Twitter versus New York Times (a traditional news dissemination medium) [112]. They showed that Twitter users are relatively less interested in world news, still they are active in spreading news of important world events. Lu et al. showed how tweets related to news event on Twitter can be mapped using energy function [61]. The methods proposed act like novel event detection techniques. The study analyzed 900 news events through 2010-2011. Castillo et al. performed qualitative and quantitative analysis on online social media activity about news articles [15]. They concluded that news articles describing breaking news events have more repetitive social media reactions, than in-depth articles.

### 3.1.2 Spam and Phishing Detection

Presence of spam, compromised accounts, malware, and phishing attacks are major concerns with respect to the quality of information on Twitter. Techniques to filter out spam / phishing on Twitter have been studied and various effective solutions have been proposed. Phishing is one of the most prominent problem on the social media such as Twitter and Facebook. Legitimate users lose millions of dollars each year to phishing scams. Chhabra et al. highlighted the role of URL shortener services like *bit.ly* [1] in spreading phishing; their results showed that URL shorteners are used for not only saving space but also hiding the identity of the phishing links [20]. The study showed how social media websites have into the focus of phishers and becoming as popular as e-commence websites like PayPal. This study used blacklisted phishing URLs from PhishTank as their ground truth. In a followup study Aggarwal et al. further analyzed and identified features that indicate to phishing tweets [4]. They used a variety of features such as tweet based, user based, URL based, WhoIs based features for the above classification. Using them, they detected phishing tweets with an accuracy of 92.52%. One of the major contributions of their work, was the Chrome Extension they developed and deployed for real-time phishing detection on Twitter. Figure 3.1 shows the *PhishAri* system developed by them.

Grier et al. characterized spam spread on Twitter via URLs. They found that 8% of 25 million

[1] https://bitly.com/

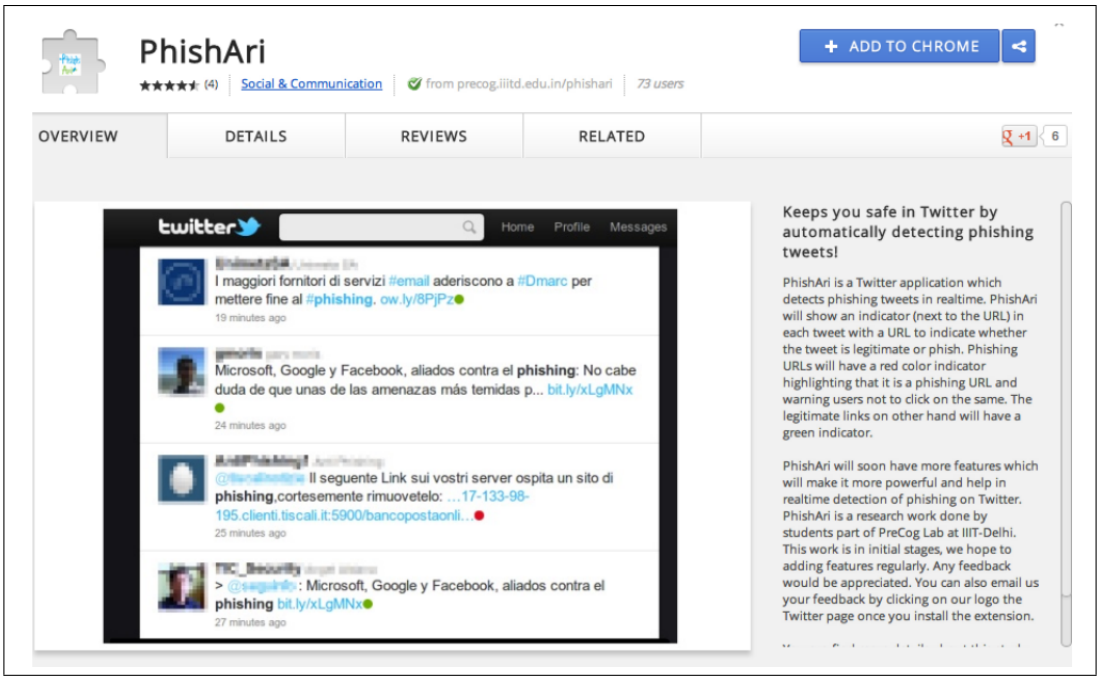



Figure 3.1: A Chrome based browser plugin to detect phishing URLs on Twitter [4].

URLs posted on Twitter point to phishing, malware, and scams listed on popular blacklists [41]. They also highlighted that most of the accounts used to spread spam were compromised accounts of legitimate users, rather than fake or dedicated accounts created by spammers. Ghosh et al. characterized social farming on Twitter, and also proposed a methodology to combat link farming [40]. Link farming, as defined by them refers to Twitter accounts linking to each other to promote their contents and be better ranked. One of their major contributions was that they proposed a ranking scheme, where users were penalized for following spammers, they showed that by this ranking scheme, influence of spammers and their followers in the network was lowered considerably. Yang et al. analyzed community or ecosystem of cyber criminals and their supporters on Twitter [108]. They showed how the criminals form a closed small world network. They coined the term *Social Butterflies* who are those accounts that have extraordinarily large numbers of followers and followings. The authors proposed and proved the effectiveness of a Criminal account Inference Algorithm (CIA) to identify new criminals on Twitter from the known set of criminals. CIA works by propagating the malicious score from users to their followers and other social relationships, to predict which of the accounts are malicious. They evaluated the performance of their algorithm on a real world dataset. Yardi et al. applied machine learning techniques to identify spammers [109]. They used features (1) searches for URLs; (2) username pattern matches; and, (3) keyword detection; and obtained 91% accuracy. Some of the other interesting and counter-intuitive findings from their work were, that the spam accounts were not very new as expected, and spammers tweeted only slightly more frequently than legitimate accounts.

Now we discuss, research work done on spam on some other social media like Youtube, a famous video sharing website and Facebook. Benevenuto et al. classified real YouTube users, as spammers, promoters, and legitimates [8]. They used techniques such as supervised machine learning algorithms to detect promoters and spammers; they achieved higher accuracy for detecting promotors; the algorithms were less effective for detecting spammers. Nazir et al. provided insightful characterization of phantom profiles for gaming applications on Facebook [67]. Through this they highlight certain distinctions between the behavior and activity of genuine user profiles from those of phantom profiles. They proposed a classification framework using SVM classifier for detecting phantom profiles of users from real profiles based on certain social network related features. They created an online game called Fighters Club to identify certain users out of the total users of the application as genuine or phantom. Viswanath et al. presented a technique based on Principal Component Analysis (PCA) to detect anomalous behavior on online social media like - like click-spam on Facebook ads [99]. The strength of this technique is that there is no need of a prior labelled dataset to train a model. They showed that normal user behavior can be represented by a low dimensional vector of latent features chosen by PCA which can be used to detect anomalous behavior with high detection rate and low false positives.

### 3.1.3 Trust and Credibility Assessment

In this section, we discuss some of the research work done to assess, characterize, analyze and compute trust and credibility of content on online social media. The first work discussed is Truthy [2], which was developed by Ratkiewicz et al. to study information diffusion on Twitter and compute a trustworthiness score for a public stream of micro-blogging updates related to an event to detect political smears, astroturfing, misinformation, and other forms of social pollution [76]. In their work, they presented certain cases of abusive behavior by Twitter users. Truthy is a live web service built upon the above work. Figure 3.2 provides a screenshot of the same. It works on real-time Twitter data with three months of data history.

Classical approach of machine learning has also been applied by researchers to detect credible and incredible content on OSM. Castillo et al. showed that automated classification techniques can be used to detect news topics from conversational topics and assessed their credibility based on various Twitter features [16]. They achieved a precision and recall of 70-80% using J48 decision tree classification algorithms. They evaluated their results with respect to data annotated by humans as ground truth. The feature sets used in their work included various kinds of features that included message (tweet content), user, topic and propagation based features. They made some interesting observations, such as, tweets which do not include URLs tend to be related to non-credible news; tweets which include negative sentiment terms are related to credible news. The above work

[2] http://truthy.indiana.edu/

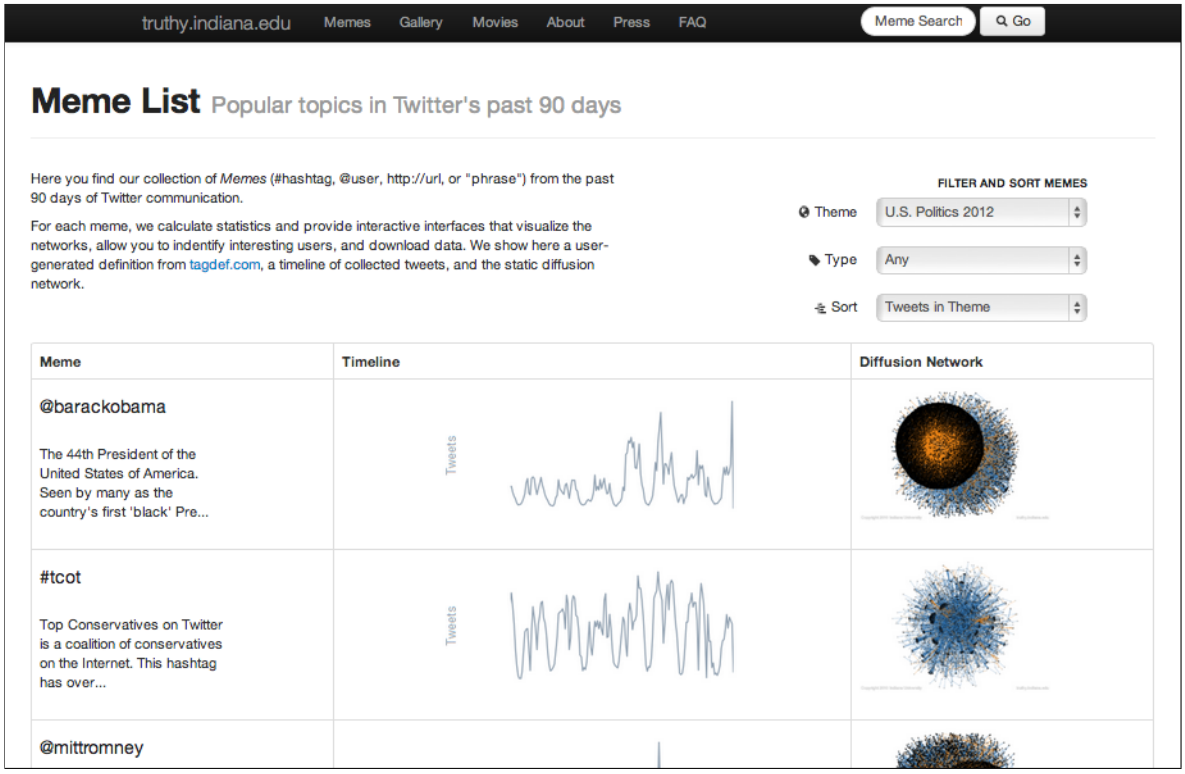



Figure 3.2: A real-time system to study information diffusion on Twitter.

focussed on determining credibility / trustworthiness of the content. Now we discuss research work that has been done focussed on determining the credibility of the users on OSM. Canini et al. analyzed usage of automated ranking strategies to measure credibility of sources of information on Twitter for any given topic [13]. The authors define a credible information source as one which has trust and domain expertise associated with it. They observed that content and network structure act as prominent features for effective credibility based ranking of users of Twitter.

Some researchers focussed their study of trustworthy or credible information during particular events which had high impact real world. Xia et al. used a supervised method of Bayesian Network is used to predict the credibility of tweets in emergency situations [103]. They worked on tweets generated during the England riots of 2011. They proposed and evaluated a two step methodology: in the first step they used a modified sequential K-means algorithm to detect an emergency situation; in the second step a Bayesian Network structure learning algorithm was used to judge the information credibility. Their evaluation of the algorithms showed an improvement over the state of the art techniques. Donovan et al focussed their work on finding indicators of credibility during different situations (8 separate event tweets) were considered. Their results showed that the best indicators of credibility were URLs, mentions, retweets and tweet length [69]. Also, they observed that the presence and effectiveness of these features increased a lot during emergency events.

A different methodology, than the above papers was followed by Morris et al., who conducted a survey to understand users perceptions regarding credibility of content on Twitter [66]. They asked about 200 participants to mark what they consider are indicators of credibility of content and users on Twitter. They found that the prominent features based on which users judge credibility are features visible at a glance, for example, username and picture of a user. By their experiments they showed that users are poor judges of credibility based only on content and are often biased by other information like the username. Also, they highlighted that there exists a disparity between features users consider relevant to credibility and those used by search engines. Another approach to detect users with high value users of credibility and trustworthiness was taken by Ghosh et al., they identified the topic based experts on Twitter [39]. Their techniques rely on the wisdom of the Twitter crowds i.e. they used the Twitter Lists feature to identify experts in various topics.

### 3.1.4 Rumor Detection

Mendoza et al. used the data from 2010 earthquake in Chile to explore the attributes of rumors and true news [63]. Their results showed that propagation of tweets related to rumors versus true news differed and could be used to develop automated classification solutions to identify correct information. Also the tweets related to rumors contained more questions versus news tweets spreading correct news. Qazvinian et al. used three types of features (content-based, network-based, and microblog-specific memes) for identifying rumors on micro-blogging websites [74]. They used statistical models and maximized a linear function of log-likelihood ratios to retrieve tweets containing rumors. Further in their work, they identified users called 'disinformers' who believe the rumor to be true and help in spreading it further. Friggeri at al. analyzed the propagation of thousands of rumor cascades on Facebook [37]. They used Snopes.com website to obtain false and true rumor topics and then searched the Facebook feed for the shares and posts regarding the same. One of the conclusions from their work was that rumors on Facebook evolve over time and their popularity varies with each variant.

### 3.1.5 Hate and Inflammatory Content Identification

Over recent years OSM have also been used to spread hate or inflammatory content. Such content if propagated during some existing real life volatile situations can have major adverse implications. There have been few researchers who have analyzed the hate content on YouTube and Twitter OSM. Sureka et al. used semi-automated techniques to discover content on Youtube that spread hate [92]. They discovered hate videos, users and hidden virtual communities using data-mining and social network analysis techniques. The precision they achieved was 88%, for detecting users that spread hate using bootstrapping techniques. Xiang applied machine learning and topic modeling techniques to detect offensive content on Twitter [104]. They achieved a true positive rate of

approx. 75% outperforming the keyword matching techniques. The authors used a seed lexicon of offensive words, and then applied LDA models for topic discovery. They obtained many useful insights, like many words in a particular topic category, were not offensive by themselves, but when combined with other words, formed offensive sex related phrases.

## 3.2 Analyzing Twitter Data during Real-World Events

The posts and activity on Twitter, impacts and plays a vital role in various real world events. Role of Twitter has been analyzed by computer scientists, psychologists and sociologists for impact in the real-world. Twitter has progressed from being merely a medium to share users' opinions; to an information sharing and dissemination agent; to propagation and coordination of relief and response efforts. Some of the popular case studies analyzed by computer scientists have been, Twitter activities during elections, natural disasters (like hurricanes, wildfires, floods, etc.), political and social uprisings (like Libya and Egypt crisis) and terrorist attacks (like Mumbai triple bomb blasts). Content and user activity patterns of Twitter during events have been analyzed for both positive and negative aspects. Some of the problems studied that result in bad quality of data, presence of spam and phishing posts, content spreading rumors / fake news, privacy breach of users via the content shared by them and use of Twitter for propagation and instigation of hate among people. Researchers have used machine learning, information retrieval, social network analysis and image and video analysis for the purpose of analyzing and characterizing Twitter usage during real-world events. In this section, we discuss some of the prominent research works done to analysis various events on Twitter.

### 3.2.1 Twitter based Early Warning Systems

Some researchers have shown how OSM can be useful during natural disasters. Palen et al. presented a path breaking vision on how Internet resources (technology and crowd based) can be used for support and assistance during mass emergencies and disasters [72]. They viewed people collectively as an important resource that can play a critical role in crisis. In Figure 3.3 they showcase how and at what levels computational techniques can play a role during emergency events. Some of the main computational techniques, which can be used in effective supporters are visualization, security / trust / reputation deductions, NLP evaluations and Query based monitoring. In a followup work to the above research proposal, Palen et al. studied two real world events, to understand and characterize the wide scale interaction on social networking websites with respect to the events [73]. The two events considered by them were: Northern Illinois University (NIU) shootings of February 14, 2008 and Virginia Tech (VT) tragedy 10 months earlier. They monitored OSM content and conducted focused interviews for the crisis events to analyze the role of these services during

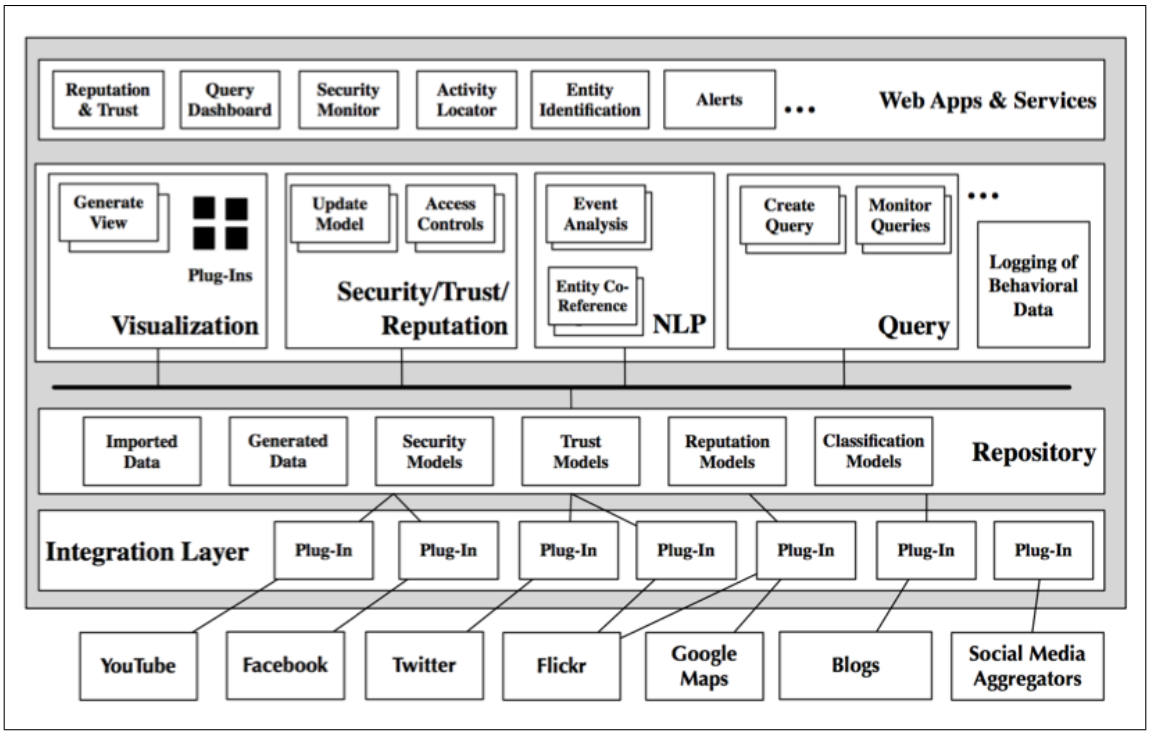



Figure 3.3: The Information Integration Landscape as suggested by Palen et al. [72]

emergencies. The work performed qualitative analysis and highlighted how information retrieval, natural language processing and policy amendments are open research areas to explore the usage of OSM for emergency management, response and support. Reuter et al. conducted a comprehensive literature review followed by interview of EMS (Emergency Management Services) authorities to understand communication needs between authorities and citizens [78]. Their analysis shows that literature on crises informatics currently explores use of OSM to scan and filter information needed during crises. However, they found a need to support bidirectional communication between authorities and citizens to enable relief activities. They concluded with a need of creating such a system and who all are potential actors in such a system. Figure 3.4 shows a screenshot the Emergent system built by them, which detects true, false and unverified topics

During the earthquakes in Japan in 2008-2009, it was observed that the tweets about the earthquake travelled faster than the tremors of the earthquake. It prompted researchers, to exploit such OSM activity to build early warning systems. Sakaki et al. used tweets as social sensors to detect earthquake events. They developed a probabilistic spatio-temporal model for predicting the center and trajectory of an event using Kalman and particle filtering techniques. Based upon the above models, they created an earthquake reporting application for Japan, which detected the earthquake occurrences based on tweets and sent users alert emails [84]. Sakaki et al. in a different research work, analyzed tweet trend to extract the events that happen during a crisis from the Twitter log

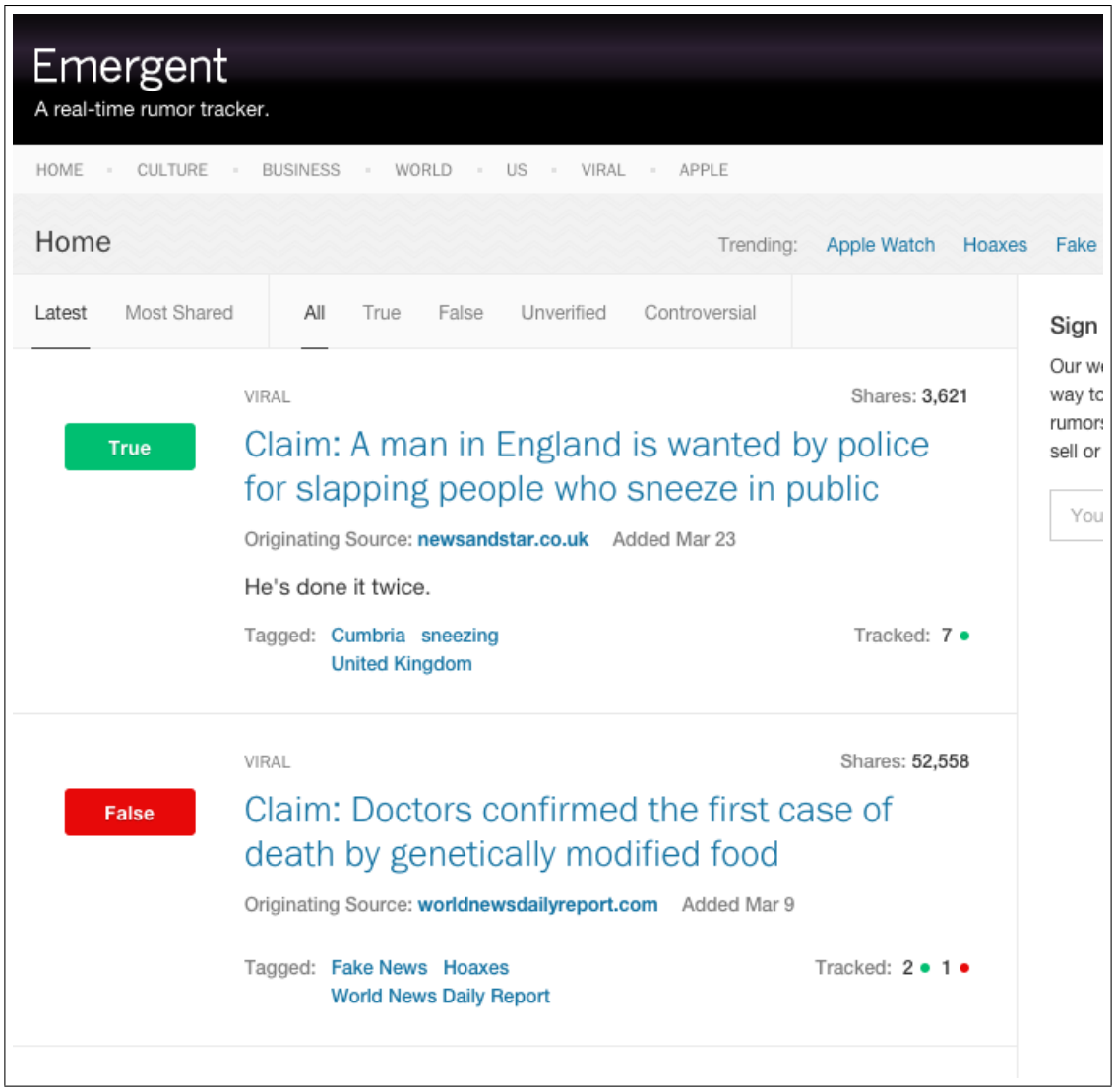
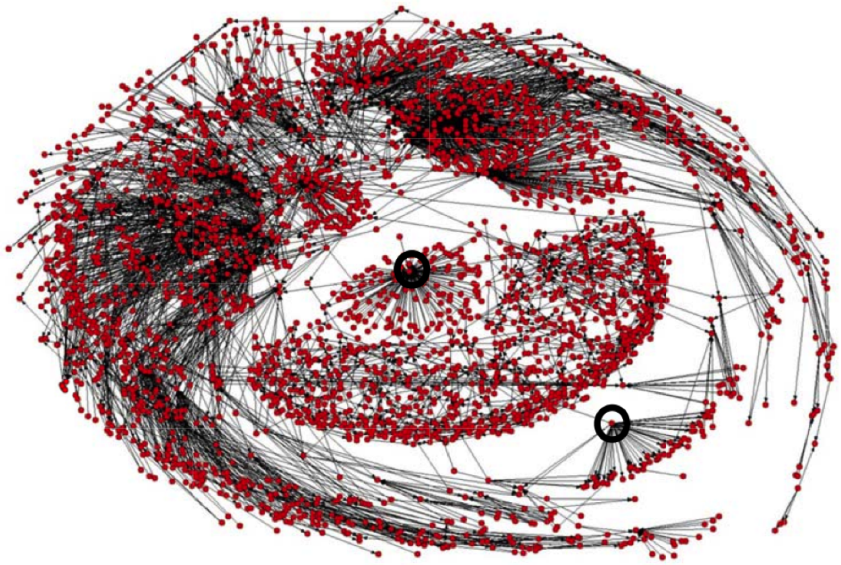



Figure 3.4: Screenshot of the Emergent system, which detects true, false and unverified topics [78].

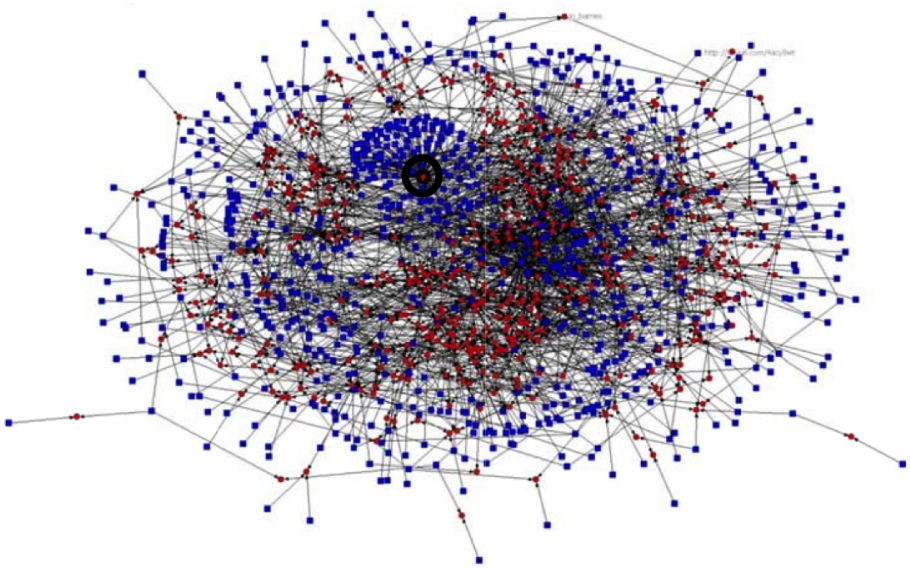
of user activity analyzed Japanese tweets on all earthquakes during 2010-2011 [85]. Some of the prominent results obtained by them via statistical analysis, like tweet frequencies of feature phones and smart-phones were dominant just after the earthquake, although those of PCs was dominant in less-damaged areas.

Cheong et al. performed social network analysis on Twitter data during Australian floods of 2011 to identify active players and their effectiveness in disseminating critical information [35]. They identified the most prominent users among Australian floods to be: local authorities (Queensland Police Services), political personalities (Premier, Prime Minister, Opposition Leader, Member of Parliament), social media volunteers, traditional media reporters, and people from not-for-profit, humanitarian, and community associations, see Figure 3.5(a). In addition to the users-users graph

formed above, they also performed social network analysis on users-resources graphs. Figure 3.5(b) shows the corresponding graph. Main resources shared were; webpages related to people with disability, animal rescue, donations, legal help, damaged produce, fraud investigations, fund raising, counseling, temporary homes, volunteering and situation based links from YouTube.



(a)



(b)

Figure 3.5: Social network analysis performed by Cheong et al. on Twitter dataset during 2011 Australian floods [35].

### 3.2.2 Extracting Situational Awareness from Twitter

Work has ben done to extract situational awareness information from the vast amount of data posted on Twitter during real-world events. Situational awareness is defined as the perception of environmental elements with respect to certain factors like time and/or space, the comprehension of their meaning, and the projection of their status after some variable has changed, such as time, or some other variable, as a predetermined event. [3] Vieweg et al. analyzed the Twitter logs for the Oklahoma Grassfires (April 2009) and the Red River Floods (March and April 2009) for presence of situational awareness content. An automated framework to enhance situational awareness during emergency situations was developed by Vieweg et al. They extracted geo-location and location-referencing information from users' tweets; which helped in increasing situation awareness during emergency events [98]. Verma et al. used natural language techniques to build an automated classifier to detect messages on Twitter that may contribute to situational awareness [97]. Another closely related work was done by Oh et al., where they analyzed Twitter stream during the 2008 Mumbai terrorist attacks [70]. Their analysis showed how information available on online social media during the attacks aided the terrorists in their decision making by increasing their *social awareness*. Corvey et al. analyzed one of the important aspects of applying computational techniques and algorithms to social media data to obtain useful information for social media content, i.e. linguistic and behavioral annotations [25]. They considered datasets for four events: Hurricane Gustav (2008), the 2009 Oklahoma Fires, the 2009 and 2010 Red River Floods, and the 2010 Haiti Earthquake. One important conclusion obtained by them was that during emergency situations, users use a specific vocabulary to convey tactical information on Twitter, as indicated by the accuracy achieved using bag-of-words model for situational awareness tweets classification.

Researchers have highlighted that useful and actionable information can be extracted by mining Twitter data and activity during crisis events. Longueville et al. analyzed Twitter feeds during forest Marseille fire event in France. They showed information from location based social networks can be used to acquire spatial temporal data that can be analyzed to provide useful localized information about the event [26]. A team at National ICT Australia Ltd. (NICTA) has been working on developing a focused search engine for Twitter and Facebook that can be used in humanitarian crisis situation [30]. Hughes et al. in their work compared the properties of tweets and users during an emergency to normal situations [3]. They performed empirical and statistical analysis on their data collected during disaster events and showed an increase in the use of URLs in tweets and a decrease in @-mentions during emergency situations. Their results highlighted the role of Twitter as a medium to spread information from external sources (URLs) rather than internal content (@-mentions).

A recent research performed an ethnographic study of Facebook to highlight its role in a society

[3] http://en.wikipedia.org/wiki/Situation_awareness

dealing with war [11]. The authors interviewed 45 Iraqi citizens, and got survey responses from 218 individuals, who were part of the current Gulf War since March 2003. The study showed how Facebook helped people to directly recover from the war effects, and indirectly re-constructing the social scaffolding of people and re-inventing the society. Some interesting observations that came from their study were: use of Facebook to give condolences to deceased soldiers and peoples families; how girls in Iraq have Facebook profiles, but do not put their pictures on their profiles as their society is very traditional; and use of Facebook pages for progressive and relief efforts in a war zone, like creation of a page to recruit teachers and educated scholars for jobs in Iraq.

## 3.3 User Modeling on Online Social Media

We present some of the research work done in applying user modeling techniques to analyze behavior of users on social networks. Yin et al. modeled user behavior using two factors: the topics related to users' intrinsic interests and the topics related to temporal context [110]. They designed a latent class statistical mixture model, called Dynamic Temporal Context-Aware Mixture model (DTCAM). They evaluated their system on four large-scale social media datasets. The authors demonstrated how user modeling techniques can be effectively used to improve the performance of recommender systems for social networks. Xu et al. proposed a mixed latent topic model to combine various factors to model users' posting behavior on Twitter [106]. The authors assumed that a user's behavior is influenced by three factors: breaking news, posts from social friends and user's interest. They evaluated and showed that their model outperforms other user models in handling the perplexity of held-out content and the quality of generated latent topics. Abel et al. developed a user modeling framework for news recommendations on Twitter using more than 2 million tweets [1]. The authors proposed different strategies for creating hashtag-based, entity-based or topic-based user profiles using semantic enrichment and temporal factors. Their results showed that consideration of temporal profile patterns can improve recommendation quality.

## 3.4 Summary

The research work presented in this chapter highlight the need to build real-time, scalable and customized solutions and algorithms for extracting trustworthy information from Twitter during real life events. We discussed how traditional techniques of machine learning, information retrieval, graph theory, and natural language processing have been applied to Twitter data and problems. Though, we also saw, that the performance of the these traditional techniques has been average, and not very promising. The reasons being the dynamics and scale of Twitter is very different from the traditional application domains. Some of the computational problems discussed by us

were filtering out spam, phishing content, detecting rumors, extracting trustworthy and credible information. In this survey report, we presented the various research work done in the field of computer science to extract quality information from Twitter during various types of real world events like floods, elections, forest fires, bomb blasts, etc. The research on Twitter is still at the stage of researchers collecting larger volumes of social media data, and characterizing or analyzing the nature of issues. Solutions proposed are naive, and often challenged and counter attacked by malicious entities. Very few papers performed in-depth analysis of misinformation of content posted on micro-blogging websites during events. There is a need to understand what types of misinformation can be propagated, who are users who initiate or propagate misinformation and what is the impact of misinformation spread during events. Answering these questions can aid in developing metrics and formulating algorithms to filter credible / trustworthy information from Twitter.

For other online entities like websites, there have been in depth and focussed research on what contributes to trust and credibility on them. Fogg et al. analyzed a large number of websites, and how users perceive credibility of websites and formulated guidelines for organizations or individuals on how to make credible websites [32–34]. Such focussed and targeted research, also needs to be done for Twitter specific problems, like to understand user's perceptions while using Twitter, and to propose guidelines to users on how to make their content on Twitter more trustworthy and credible. Correspondingly, there is also a need for educating and spreading awareness regarding threats of misinformation spread on Twitter. Similar work, was done for email phishing scams by Kumaraguru et al. [57, 58].

The research work discussed is this chapter, showcased that there has been considerable focus on analyzing the trustworthiness of content generated on Twitter during real world events, yet there is dire need to build scalable, generalizable and real-time solutions for the same.

# Chapter 4

# Characterizing and Detecting Fake Content

This chapter is joint work with Prof. Anupam Joshi, Professor, University of Maryland Baltimore County (USA) and Hemank Lamba. Hemank was a former student of IIIT-Delhi and currently pursuing his PhD in the School of Computer Science at Carnegie Mellon University. An earlier version of the content in this chapter was published in Second Workshop on Privacy and Security in Online Social Media, Proceedings of the 22nd international conference on World Wide Web companion (WWW), ACM, 2013 [6]. Work done in this chapter featured in an article by the Wired magazine, Fast Company and online blog by Dr. Patrick Meier. [1] [2] [3]

The first section presents the motivation and contributions of this chapter. Section 4.2 explains methodology that we used in collecting data, annotation and classifying the tweets. Section 4.3 describes the results from our analysis . Section 4.4 discusses the implications of our research.

## 4.1 Fake Content on Twitter during Crisis Events

Over the past few years there has been increase in the usage of Online Social Media (OSM) services as a medium for people to share, coordinate and spread information about events while they are going on. Though a large volume of content is posted on OSM, not all of the information is of good quality with respect to the event, like it may be fake, incorrect or noisy. Extracting good quality information is one of the biggest challenges in utilizing information from OSM. Over last few years,

[1] http://www.wired.co.uk/magazine/archive/2014/01/how-to/spot-a-fake-image-on-twitter
[2] http://irevolution.net/2013/07/01/automatically-identifying-fake-images/
[3] http://www.fastcoexist.com/3016541/never-get-fooled-by-a-fake-tweet-again-with-this-new-technology

people have highlighted how OSM can be used to help in extracting useful information about real life events. But, on the other hand, there have been many instances which have highlighted the negative effects on content on online social media on real-world events. The information shared and accessed on social media such as Twitter, is in real-time, the impact of any malicious intended activity, like spreading fake images and rumors needs to be detected and curbed from spreading immediately. Such false and incorrect information can lead to chaos and panic among people on the ground. Since detecting whether images posted are fake or not, using traditional image analysis methods, can be highly time and resource consuming, we explore the option of using Twitter specific features, like the tweet text, meta-data and user details in differentiating tweets with fake images from tweets with real images.

***Hurricane Sandy***: Hurricane Sandy caused mass destruction and turmoil in and around USA from October 22nd to October 31st, 2012. According to NBC News, the death toll in the U.S. was 109, including at least 40 in New York City. NBC also reported that damages from Hurricane Sandy exceeded $50 billion. Online social media such as Twitter and Facebook were widely used by people to keep abreast about latest updates of the storm [38]. OSM was also widely exploited by malicious entities during Sandy, to spread rumors and fake pictures in real-time [54, 107]. Such fake images and news became extremely viral on OSM and caused commotion among the people affected by the hurricane. Hence, it was an ideal event, to analyze the spread and impact of fake and incorrect information on social media. Figure 4.1 shows some of the fake images that were spread during Hurricane Sandy, which we also found in data collected by us.

There is a need to build automated solutions that can help people judge the authenticity of content appearing on OSM in real-time. The aim of this work is to characterize and identify the propagation of fake pictures on Twitter. These fake images, created panic and chaos among the people. The affect of the spreading such false information can be multifold in case of crisis situations; hence, we analyzed the propagation of fake images URLs during Hurricane Sandy. The power and impact of online social media in shaping real world events has been widely studied by researchers across the globe. To the best of our knowledge this was the first study to analyze the diffusion and spread of fake pictures on OSM. The main contributions of this work are:

- We performed in-depth characterization of tweets sharing fake images on Twitter during Hurricane Sandy. We found that the tweets containing the fake images were mostly retweets (86%), i.e. very few users posted original tweets with fake images. Also, we found that social network of a user on Twitter had little impact on making these fake images viral, there was just 11% overlap between the retweet and follower graphs of tweets containing fake images.
- We used classification algorithms to distinguish between tweets containing fake and real im-

(a)

(b)

(c)

Figure 4.1: Some of the fake pictures of Hurricane Sandy that were shared on Twitter. (a) A tweet with a fake picture depicting a McDonald's being flooded (b) Faked image of stormy New York skyline (c) Picture of shark in New Jersey.

ages. We primarily used two kinds of features: user level and tweet level features. Best accuracy of 97% was achieved by the decision tree classifier using tweet based features.

## 4.2 Methodology

In this section, we discuss our research methodology in detail. First we describe the methodology of collecting data from Twitter, followed by the various analytical techniques applied in this chapter.

### 4.2.1 Data Collection

We collected data from Twitter using the *Streaming API*. [4] This API enables researchers to extract tweets in real-time, based on certain query parameters like words in the tweet, time of posting of tweet, etc. We queried the Twitter *Trends API* after every hour for the current trending topics, [5] and collect tweets corresponding to these topics as query search words for the *Streaming API*.

From our data, we filtered tweets posted during the Hurricane Sandy, i.e. from Oct 20th to Nov 1st, 2012. [6] We used the keywords'*sandy*' and '*hurricane*' to select the relevant tweets. We got about 1.8 million tweets by 1.2 million unique users on Hurricane Sandy from Oct 20th to Nov 1st, 2012. Table 5.1 gives the descriptive statistics of the tweets and users data collected to the event, and Figure 5.5 shows the spatial distribution of these tweets (approx. 19,000 tweets had geo-location embedded in them).

Table 4.1: Descriptive statistics of the Twitter dataset for Hurricane Sandy.

| | |
|---|---|
| **Total tweets** | 1,782,526 |
| **Total unique users** | 1,174,266 |
| **Tweets with URLs** | 622,860 |

Table 4.1 describes the statistics for data related to tweets containing fake and real image URLs. The following steps explain the methodology for data collection of tweets containing fake and real images of Hurricane Sandy:

- STEP 1: We used the Twitter Streaming API to collect data using keywords 'sandy' and 'hurricane' to select the relevant tweets. We got about 1.8 million tweets by 1.2 million unique users on Hurricane Sandy from Oct 20th to Nov 1st, 2012.
- STEP 2: Next, we collected a list of fake and real pictures (and their corresponding URLs) of Hurricane Sandy using online resources (articles, tweets and blogs). One of the prominent data source we used was the list of fake and real images made public by the Guardian news media company [54]. The list of images (and their corresponding URLs) provided by Guardian,

[4] `https://dev.twitter.com/docs/streaming-api.`

[5] `https://dev.twitter.com/docs/api/1/get/trends`

[6] Hurricane Sandy formed on Oct 22nd, yet we filtered our data from two days before Oct 20th till one day after the hurricane.

classified the most frequently shared images shared during the hurricane as fake or real image URLs. There were many other articles and blogs that covered the real and fake images which were spread on Twitter [22, 101].

- STEP 3: Finally, we searched our tweets database to find tweets containing these fake or real images of the hurricane. We found eight unique fake images of Sandy that were spread on Twitter in our dataset. In total 10,350 tweets contained these fake image URLs.

Each image here corresponds to a unique URL for it, for example *https://twitter.com/HerrBestatter/status/303983092421492736/photo/1* is a URL of a fake image of Hurricane Sandy posted on Twitter. By the process described in Section 1.1 above, we found eight unique fake image URLs present in 10,000 tweets. For example a Twitter user posts a fake photo on Twitter, there are two ways in which other users can spread the same, either they directly retweet the tweet or they copy the URL of the fake image posted by the user and include it in their tweet. So 10,350 tweets either directly retweeted the tweets with fake images or copied the fake image tweet URL in their own tweet.

Figure 4.2: Spatial distribution of total tweets on Hurricane Sandy. We have plotted about 19,000 tweets, which had embedded geo-location data in them.

### 4.2.2 Temporal and Network Analysis

We performed characterization of the tweets containing fake images URLs and their propagation, to understand how they became viral. First we performed temporal analysis on the fake images tweets. We analyzed how many such tweets were shared per hour on Twitter. Also, we analyzed the

Table 4.2: Descriptive statistics of the tweets with fake and real images URLs.

| | |
|---|---|
| **Tweets with fake images** | 10,350 |
| **Users with fake images** | 10,215 |
| **Tweets with real images** | 5,767 |
| **Users with real images** | 5,678 |

sudden peaks (from $x1$ hour to $x1+1$) in the graph more closely. We constructed the retweet graph for the sudden peak in the temporal analysis, to find out what changes in the network topology lead to the viral spread of these images. We obtained certain useful insights, about the nature and spread of fake image URLs on Twitter, which are summarized in the next section

Next, we analyzed what role the social network graph of a user on Twitter plays in propagation of fake URLs. The explicit social network of a user on Twitter, is that of his follower graph. We wanted to analyze what percentage of information diffusion takes place via this follower network graph of a user. The details of the algorithm used to compute are summarized in Algorithm 2. In the function, *Create_Graph_Followers*, we crawled the follower network of all the unique users that had tweeted the fake images, using the REST API of Twitter. The network created had 10,779,122 edges and 10,215 nodes. In *Create_Graph_Retweets*, we created a retweet network, where an edge between two nodes exists if one user had retweeted the other's tweet. A hashmap, $H[1..n]$, is created to compute the overlap between the follower and retweets graphs.

**Algorithm 1** Compute_Overlap

```
1: Create_Graph_Retweets()
2: Create_Graph_Followers()
3: for each edge in the retweet network do
4:     num_retweet_edges++
5:     Insert edge into hashmap, H[1..n]
6: end for
7: for each edge in the follower network do
8:     Insert each edge in hashmap, H[1..n]
9:     if collision then
10:        intersections++
11:    end if
12: end for
13: %overlap = (intersections/num_retweet_edges) * 100
```

### 4.2.3 Building Algorithms for Fake Image Detection using Meta-data

We analyzed the effectiveness of machine learning algorithms in detecting tweets containing fake image URLs versus tweets containing real images of Sandy. We performed two-class classification

using Naive Bayes and J48 Decision Tree classifiers. We had a dataset of 10,350 tweets containing fake image URLs and 5,767 tweets containing real images URLs. To avoid any bias, due to unequal size of any of the classes, we randomly selected 5,767 tweets from the fake images tweets, and then applied classification.

We used two kinds of features, for the classification algorithm. Table 4.3 summarizes the features computed by us for each tweet and the user of the tweet.

- **Source or user level features** [F1]: The attributes of the user who posted the tweet. We consider properties such as number of friends, followers and status messages of the user as part of this set.
- **Content or tweet level features** [F2]: The 140 characters posted by users contain data (e.g. words, URLs, hashtags) and meta-data (e.g. is tweet a reply or a retweet) related to it.

## 4.3 Results

In this section, we summarize the results obtained for the characterization and classification analysis performed.

### 4.3.1 Characterization Results

We found that out of the 10,350 tweets containing fake images, about 86% were retweets, i.e., only for about 14% tweets, the users composed new tweet content with fake image URLs embedded. For the temporal analysis, we plotted the per hour tweeting activity of the fake images. From Figure 5.4, we see that the fake images spread spikes at 12 hours after their introduction in Twitter. We analyzed the spread of these picture URLs one hour before and after the spike. We construct the reply and retweet graph for the tweets sharing these fake picture URLs on October 29th, at 21 hours and 22 hours, as shown in Figure 4.5. We see that there are only a few users with very high degree, that is, only a few users results in majority of the retweets. We confirmed this statistically, Figure 4.4 (CDF) shows that top 30 users (0.3% of the users) resulted in 90% of retweets of the fake images. Combining results from both the graphs, we conclude that though the fake URLs were present in the Twitter network for almost 12 hours before they became viral, also the sudden spike in their propagation via retweets happened only because of a few users.

Next, we determine the role of Twitter network graph on the retweets propagation of the fake image tweets. We ran the *Compute_overlap* algorithm discussed above. We found the number of overlapping edges as 1,215, which leads to a percentage overlap of 11% between the retweet and follower graphs. Table 4.4 summarizes the results of the *Compute_overlap* algorithm. This indicates

| **User Features [F1]** |
|---|
| Number of Friends |
| Number of Followers |
| Follower-Friend Ratio |
| Number of times listed |
| User has a URL |
| User is a verified user |
| Age of user account |
| **Tweet Features [F2]** |
| Length of Tweet |
| Number of Words |
| Contains Question Mark? |
| Contains Exclamation Mark? |
| Number of Question Marks |
| Number of Exclamation Marks |
| Contains Happy Emoticon |
| Contains Sad Emoticon |
| Contains First Order Pronoun |
| Contains Second Order Pronoun |
| Contains Third Order Pronoun |
| Number of uppercase characters |
| Number of negative sentiment words |
| Number of positive sentiment words |
| Number of mentions |
| Number of hashtags |
| Number of URLs |
| Retweet count |

Table 4.3: User and tweet based features used for classification of fake and real images of Sandy.

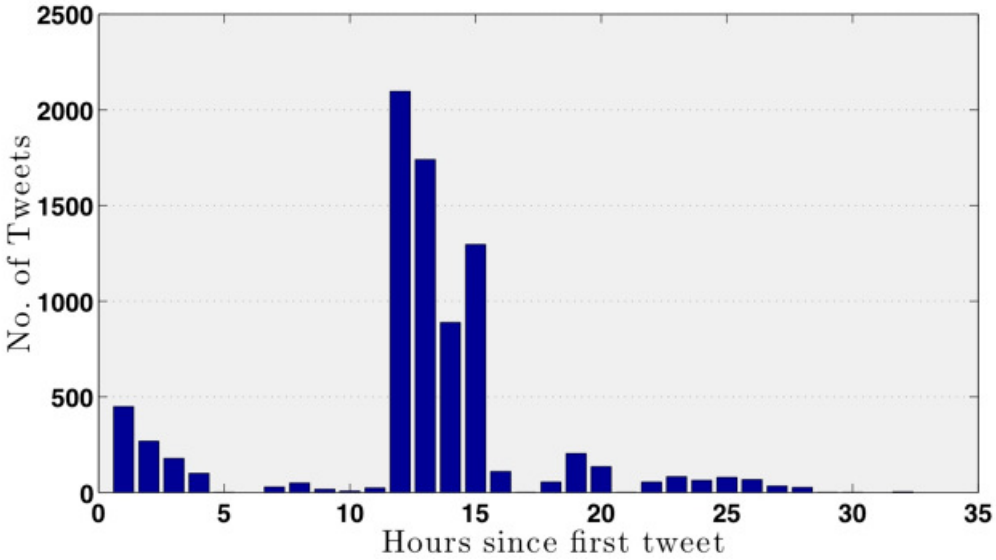



Figure 4.3: Details of data collected for the ftweets with fake images. Temporal distribution of tweets, hour wise, starting from the first hour that a fake image tweet was posted.

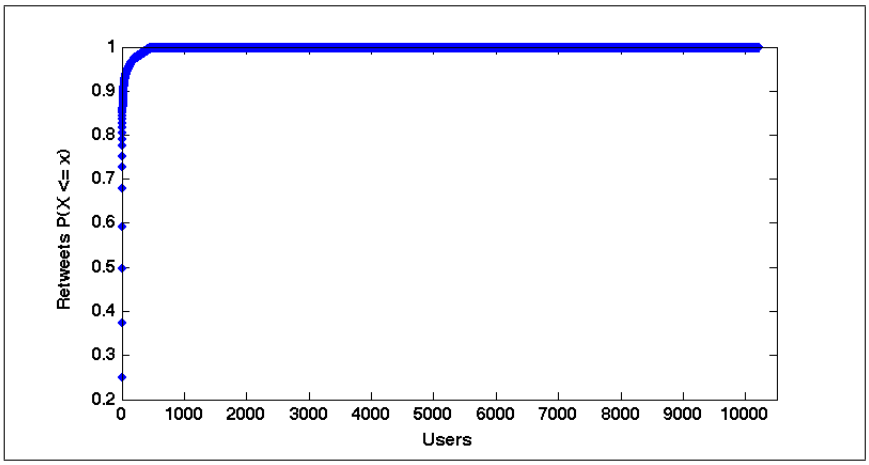
that there was a very limited retweet activity which originated because of the people in a user's follower graph. In case of crisis events, people retweet content from other users that they see in Twitter search results, irrespective of whether they follow that user or not.

| | |
|---|---|
| Total edges in the retweet network | 10,508 |
| Total edges in the follower-followee network | 10,799,122 |
| Total edges that exist in both retweet network and the follower-followee network | 1,215 |
| %age overlap | **11%** |

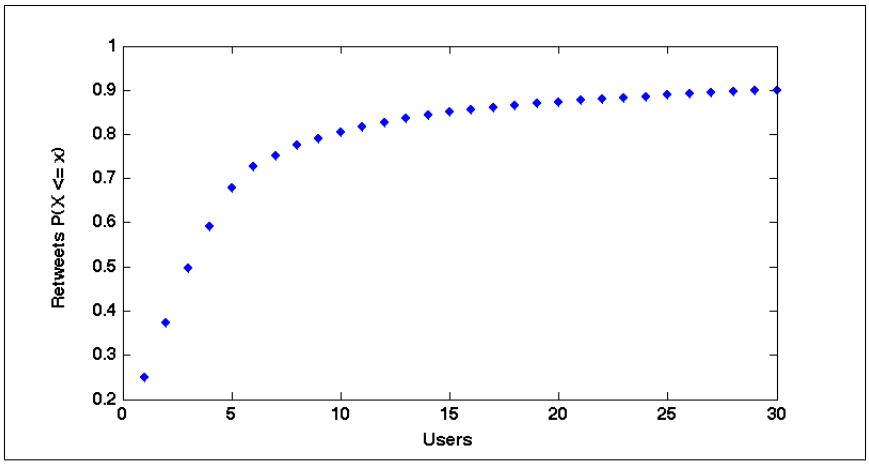
Table 4.4: Results of the Algorithm *Compute_overlap*. We found only 11% overlap between the follower and retweet graphs for the tweets containing fake images.

### 4.3.2 Classification Results

In the above section, we characterized the properties and behavior associated with spread of false information, in form of fake images on Twitter. The next important step is to explore features and algorithms that can effectively help us is identifying the fake content in real-time. We performed 10-fold cross validation while applying classification models. We applied two standard algorithms used for classification: Naive Bayes and Decision Tree (J48). As described before, we took 5,767 tweets for both fake and real image containing tweets. For each data point, we created user and tweet level feature vectors. Table 4.5 summarizes the results from the classification experiment.




(a) All users

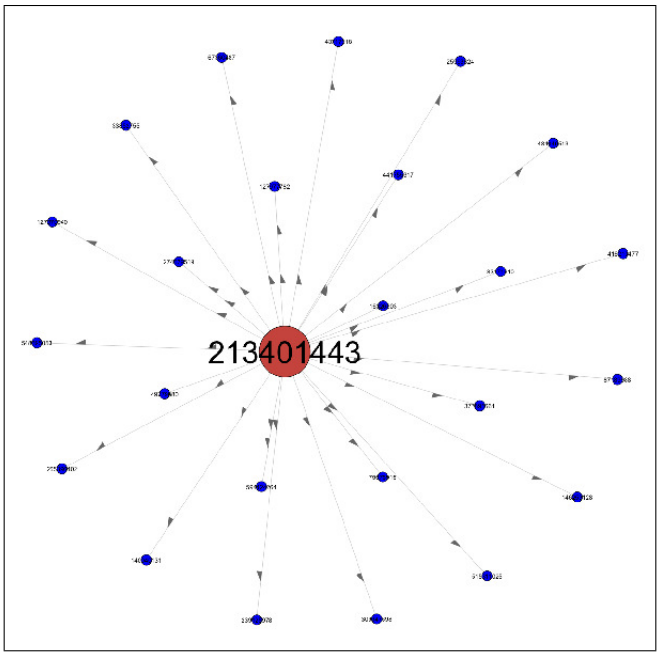



(b) Top 30 users

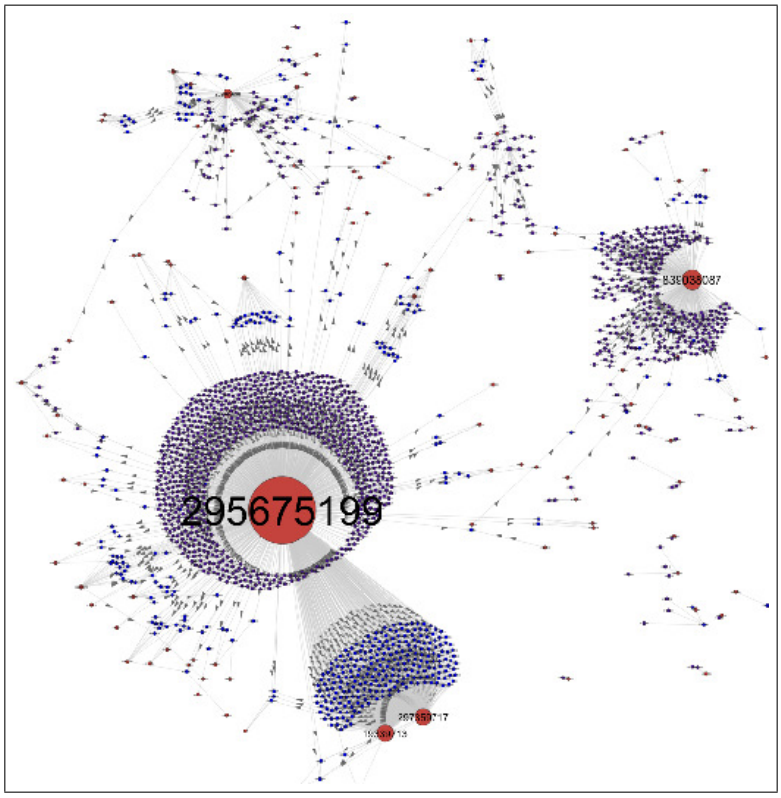
Figure 4.4: CDF of retweets of the fake image tweets by the users. It shows that top 30 users (0.3% of the users) resulted in 90% of retweets of the fake images

We achieve a good accuracy of above 90% for both classifiers, though Decision Tree out performs the Naives Bayes classifier. We can also see that, user based features, provide very poor accuracy in distinguishing fake image URLs, while tweet based features performed very well. We performed an analysis of the top tweet features for the classification and found that the most important five features were: length of the tweet, contains question mark, contains positive words, contains exclamation mark and tweet contains pronouns.

We would also like to mention that high accuracy results obtained by us, may be attributed to the similar nature of many tweets (since a lot of tweets are retweets of other tweets in our dataset). We can conclude that, content and property analysis of tweets can help us in identifying real image URLs being shared on Twitter with a high accuracy.




(a)




(b)

Figure 4.5: Spread of tweets with fake images (retweet and reply graph), the number on the node is user profile ID on Twitter. The figure shows that the fake images became viral very fast, within an hour there was a tremendous growth in the number of people tweeting them. (a) Oct. 29, 2100 GMT (b) Oct. 29, 2200 GMT.

| | F1 | F2 | F1+F2 |
|---|---|---|---|
| **Naive Bayes** | 56.32% | 91.97% | 91.52% |
| **Decision Tree** | 53.24% | **97.65%** | 96.65 % |

Table 4.5: Classification results for tweets containing fake image and real images. Our results showed that, tweet based features are more effective in distinguishing the two classes.

## 4.4 Discussion

Online social media has the capability of playing the role of, either a life saver or that of a daemon during the times of crisis. In this research work, we highlighted one of the malicious intended usage of Twitter during a real-world event, i.e. spreading fake images. We analyzed the activity on the online social networking website Twitter, during Hurricane Sandy (2012) that spread fake images. We identified 10,350 unique tweets containing fake images that were circulated on Twitter, during Hurricane Sandy. We performed a characterization analysis, to understand the temporal, social reputation and influence patterns of the spread of these fake images. We found that 86% tweets spreading the fake images were retweets, hence very few were original tweets by users. Also, our results showed that top 30 users (0.3% of the users) resulted in 90% of retweets of the fake image. Hence, we can concluded that only a handful of users contributed to majority of the damage, via the retweeting activity on the Twitter. We analyzed the role of Twitter social graph in propagating the fake images. We crawled the network data for Twitter, i.e., the follower relationships of the users and applied our algorithm to compute the overlap with fake images retweet propagation. We found only a 11% overlap between the retweet and follower graphs for the users who tweeted fake images of Sandy. This result highlights the fact that, at the time of crisis, users retweet information from other users irrespective of the fact whether they follow them or not. Next, we used classification models, to identify fake images from real images of Hurricane Sandy. Best results were obtained from Decision Tree classifier, we got 97% accuracy in predicting fake images from real. Tweet based features are very effective in distinguishing fake images tweets from real, while the performance of user based features was very poor. Our research work provided insights into the behavioral pattern of the spread of fake image tweets. Also our results provided a proof of concept that, automated techniques can be used in identifying real images from fake images posted on Twitter.

# Chapter 5

# Analyzing and Predicting Propagation of Misinformation

This chapter is joint work with Hemank Lamba, a former student of IIIT-Delhi and currently pursuing his PhD in the School of Computer Science at Carnegie Mellon University. An earlier version of the content in this chapter was published in Proceedings of Eighth IEEE APWG eCrime Research Summit (eCRS) 2013 [5]. Work done in this chapter featured in an article by the Fast Company, [1] and online blog by Dr. Patrick Meier. [2]

This chapter is organized as follows: the next section discusses the motivation of studying malicious and rumor content on Twitter during real-world events. Section 5.2 describes the methodology and data description for this work. Section 5.3 summarizes the temporal, spatial, impact analysis for the propagation of fake content during Boston Marathon blasts. Section 5.4 presents the network, text and user attributes from suspended user profiles created on Twitter. Section 5.5 presents the discussion and key take aways from this work.

## 5.1 Propagating Misinformation on Twitter

Various forms of e-crimes like spam, phishing, rumors, fake information and identity theft engulf the social media experience [7,24]. A lot of research work has been done on analyzing various forms of e-crimes on online social media. Similarly, a lot of researchers have analyzed content generated on OSM during real-world events [35, 84, 98], but not a lot of researchers have worked on the

[1] http://www.fastcoexist.com/3016541/never-get-fooled-by-a-fake-tweet-again-with-this-new-technology
[2] http://irevolution.net/2013/10/20/fake-content-on-twitter-boston-marathon/

intersection of these two problems, particularly in context of identifying fake content during real-world events. We collected a large dataset of fake tweets during the Boston blasts, and presented an in-depth characterization and analysis of propagation of rumors on Twitter. The three main research questions we aim to explore in our work are: Firstly, can simple user attributes like number of followers and account being verified, be used to differentiate between fake and true news? Secondly, can we predict how viral some fake content would become in future based upon who are the users currently sharing the content? Thirdly, what kind of interactions occur between the suspended accounts on Twitter, that are created during a crisis event?

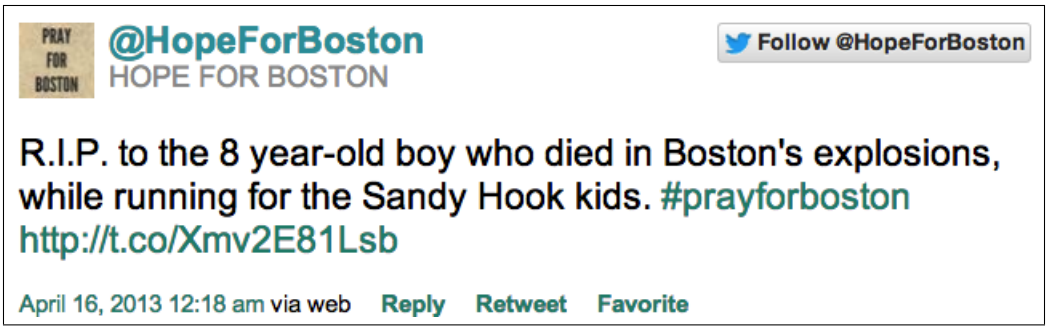
The aim of this chapter is to characterize and propose solutions to counter various forms of malicious activities on Twitter during events such as the Boston blasts. In this work we used data collected during the Boston marathon blasts. We collected about 7.8 million tweets for the Boston marathon blasts using the Twitter APIs. Our data collection was limited from the fact that it was started 45 minutes after the blasts had occurred. To the best of our knowledge this is one of the largest studies, to analyze a dataset of tweets containing fake information / rumors. Also, this work presents the first comprehensive characterization of content posted on Twitter during the Boston blasts, with special focus on fake content propagation. Figure 5.1 presents some such sample tweets of rumors / fake content during Boston blasts.

**Boston Marathon Blasts:** Twin blasts occurred during the Boston Marathon on April 15th, 2013 at 18:50 GMT. Three people were killed and 264 were injured in the incident [94]. Two suspects Tamerlan Tsarnaev and Dzhokhar Tsarnaev carried out the bombings. There was a huge volume of content posted on social media websites, including Twitter, after the blasts. We saw online social media being effectively used by Boston Police to track down the suspects and pass on important information to the public. There was a lot of fake content and rumors posted on Twitter, e.g. fake charities offering to donate money to Boston victims and about children who were running the marathon being dead in the blasts. Figure 5.2 shows a picture clicked during the blasts [93]. Timeline of social media coverage of Boston blasts has been analyzed and visualized by some people [87].

There were two primary kinds of fake content that emerged on Twitter during the Boston marathon blasts. We present analysis about the domain of fake information creation and propagation, along with suspended profiles on Twitter during crisis events. Our main contributions are:

- We characterized the spread of fake content on Twitter using temporal, source and user profile attributes. We found that approx. 75% of fake tweets are propagated via mobile phone devices.
- We applied linear regression model to predict how viral fake content would become in future based on the attributes and impact of users currently propagating the content.

(a)




(b)

Figure 5.1: Two sample tweets containing fake content. (a) A tweet from a fake charity profile. (b) Rumor about a child being killed in the blasts.

- We analyzed the activity and interaction graphs for the suspended user profiles created during Boston blasts. We identified that malicious users exploit a real-world event to indulge in e-crimes like impersonation and propaganda.

## 5.2 Methodology

In this section, we discuss the data collection and annotation methodology in detail. Figure 5.3 presents the architecture diagram of the methodology followed in this work. For the Boston marathon blasts, we identified a set of malicious tweets and user profiles. We created the dataset of malicious tweets by manually tagging the most viral tweets posted on Twitter. The most viral tweets were tagged by us in three categories: True / Fake / None. To identify malicious user profiles, we selected the profiles that were created immediately after the Boston blasts and were later suspended by Twitter for violating its terms and conditions.

Figure 5.2: A picture clicked during the Boston marathon blasts.

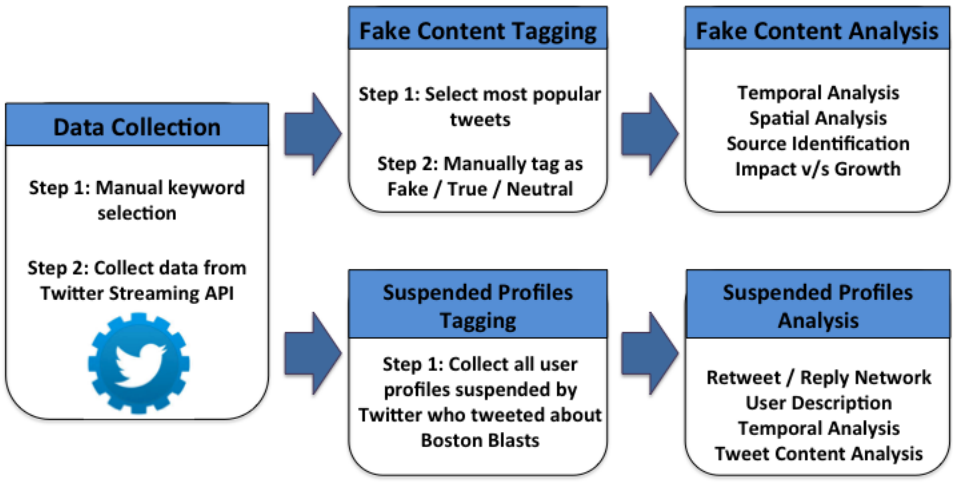



Figure 5.3: Architecture diagram describing the methodology followed in this chapter for analyzing fake content and suspended profiles on Twitter during the Boston marathon blasts.

### 5.2.1 Data Collection

We used the Streaming API of Twitter to collect tweets related to Boston blasts [95]. We used the following keywords to collect data: *Dzhokhar Tsarnaev, #watertown, #manhunt, Sean Collier, #BostonStrong, #bostonbombing, #oneboston, bostonmarathon, #prayforboston, boston marathon,*

*#bostonblasts, boston blasts, boston terrorist, boston explosions, bostonhelp, boston suspect.* [3] We collected about 7.9 million unique tweets by 3.7 million unique users. The descriptive statistics of the data are given in Table 5.1. Our data collection was started about 45 minutes after the blast.

Table 5.1: Descriptive statistics of dataset for Boston blasts, April 15th - 19th, 2013.

| | |
|---|---|
| **Total tweets** | 7,888,374 |
| **Total users** | 3,677,531 |
| **Tweets with URLs** | 3,420,228 |
| **Tweets with Geo-location** | 62,629 |
| **Retweets** | 4,464,201 |
| **Replies** | 260,627 |
| **Time of the blast** | Mon Apr 15 18:50 2013 |
| **Time of first tweet** | Mon Apr 15 18:53:47 2013 |
| **Time of first image of blast** | Mon Apr 15 18:54:06 2013 |
| **Time of last tweet** | Thu Apr 25 01:23:39 2013 |

We were also able to capture most of the tweets for the first 45 mins since they got retweeted later. This is the largest known dataset of Boston marathon blasts. Within 3 minutes of the blasts happening, we got our first tweet; and within 4 minutes of the blasts the first picture of the blast was posted on Twitter, which is captured in our dataset.

Figure 5.4 shows the temporal distribution of the tweets collected from the blasts. We have annotated the figure, with the corresponding real-world happenings to understand when the activity on Twitter peaked. Boston blasts and the manhunt of suspects was an event that generated a lot of posts on Twitter. Many people offered help and coordinated relief measures via Twitter. The Boston police used its official account *@boston_police* to spread the photograph of the suspects and got aid in their manhunt.

In all 0.8% (62,629 / 7,888,374) of total tweets during the Boston blasts, shared geo-location in their tweets. Figure 5.5 shows the distribution of the geo-tagged tweets. For the 500,000 annotated tweets that were retweets of the top twenty most viral tweets, did not have any geo-location information in them.

[3] These phrases were given as a comma-separated list which were used to determine what tweets will be delivered on the stream. A phrase may be one or more terms separated by spaces, and a phrase will match if all of the terms in the phrase are present in the Tweet, regardless of order and ignoring case. By this model, you can think of commas as logical ORs, while spaces are equivalent to logical ANDs (e.g. 'the twitter' is the AND twitter, and 'the,twitter' is the OR twitter).

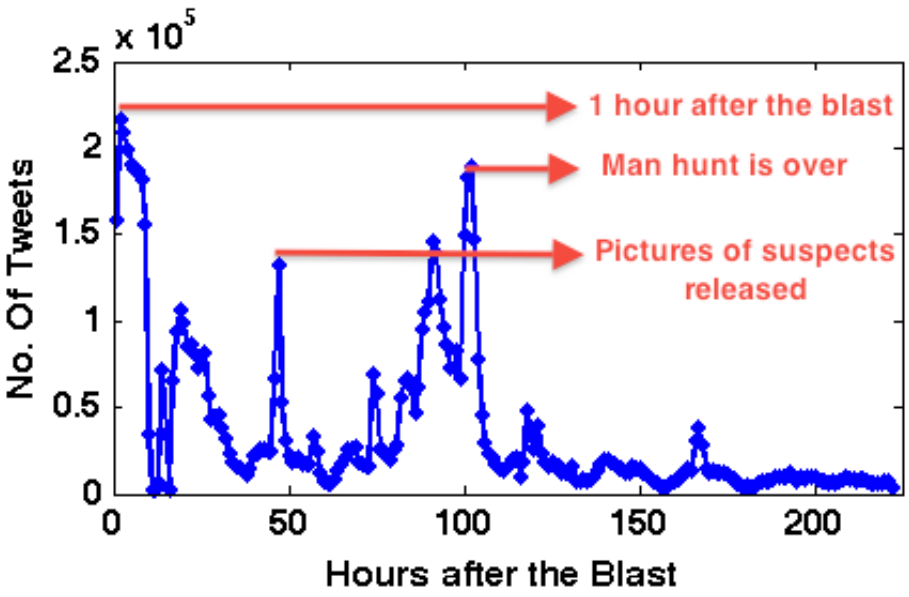



Figure 5.4: Timeline for the tweets collected for the Boston marathon blasts.

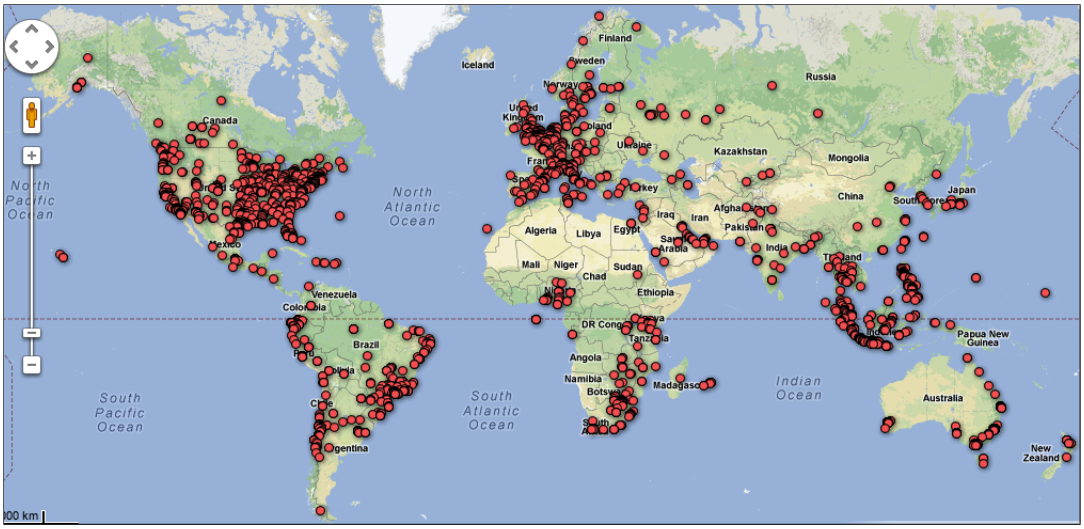



Figure 5.5: Geo-location for the tweets collected for the Boston marathon blasts.

### 5.2.2 Rumors and True Tweets

We selected the top 20 most popular (most retweeted) tweets during the Boston Marathon blasts. In total, the top 20 tweets constituted 453,954 total tweets (6% of all 7.9 million Boston tweets). On Twitter, each retweet of a tweet is a unique tweet in itself (with the same text as the original tweet but a different tweet id). For each retweet, the Twitter API gives a *retweeted_status* field which contains all details of the original tweet. So when we collected Twitter data for any event, we obtain both tweets and retweets using the Twitter API. We then expand our dataset by also

including the original tweets (if not included already) as present in the *retweeted_status*. So the top 20 tweets had 453,954 retweets (each retweet is considered a separate tweet with its unique id, time of post, source and user details) in the 7.9 million total tweets (tweets + retweets) collected by us.

We manually tagged these tweets in three categories: True , Rumor / Fake and NA. NA stands for 'Not Applicable', and it represents tweets which neither give any true or fake information about the event, they were mostly personal opinions and condolences. Table 5.2 shows the tweets, their number of retweets and their corresponding tag. We found that 51% of the tweets were generic comments and opinions of people, with neither true nor fake information. The percentage of fake tweets was much more (29%) than true information (20%).

There were 3,249 users who tweeted both true and fake information, which was 3% of 94,383 unique users who tweeted true information and 2% of 147,169 unique users who tweeted fake information tweets. Although, the time period of fake and NA category tweets were quite overlapping, we found only a overlap of 2,895 users in both the categories. These observations imply, that each set of users who tweeted fake / true / NA category of tweets are unique from each other. Since we considered only retweets and replies to most viral tweets, which were retweeted by the user using retweet button, we can be reassured that all retweets carried the same exact text and hence also belonged to the same category. [4]

## 5.3 Characterizing Fake Content Propagation

In this section, we analyze various aspects about spread of fake information tweets and propagation during the Boston marathon blasts.

### 5.3.1 Temporal Analysis

To analyze the temporal distribution of the tweets posted during the Boston blasts, we calculated the number of tweets posted per hour. Figure 5.6 shows the cumulative growth of the three categories of tweets over five days. We plotted the log values on the Y-axis, to avoid bias from the large amount of total number of tweets. The inset graph shows the growth of the data for the first fifty hours only. For the first 7-8 hours, we observed that only tweets belonging to the NA and fake category were being circulated in the Twitter network. Confirmed and true information by official sources (user profiles of government or news media) is introduced in the Twitter network after a few hours of the blasts. Once introduced, these tweets become viral soon. Atleast for the initial hours after a crisis, we need to distinguish only between the two categories of fake content and NA. For fake category tweets, we see that the first hour has slow growth, but once it becomes viral they

[4] https://support.twitter.com/articles/20169873-retweeting-another-tweet

Table 5.2: Top 20 most popular tweets (retweeted and replied). We tagged each of the tweets to belong to the following three categories: Fake / Rumor, True and Not Applicable (NA).

| **Retweets** | **Tweet Text** | **Category** |
|---|---|---|
| 87,903 | #PrayForBoston | NA |
| 33,661 | R.I.P. to the 8 year-old girl who died in Boston's explosions, while running for the Sandy Hook kids. #prayforboston http://t.co/WhaaTG3nSP | Fake / Rumor |
| 30,735 | Dzhokhar Tsarnaev, I have bad news for you. We never lose at Hide and Seek, just ask Bin Laden and Saddam. Good Luck.Sincerely, America | NA |
| 28,350 | For each RT this gets, $1 will be donated to the victims of the Boston Marathon Explosions. #DonateToBoston | Fake / Rumor |
| 27,163 | #prayforboston | NA |
| 26,954 | Reports of Marathon Runners that crossed finish line and continued to run to Mass General Hospital to give blood to victims #PrayforBoston | Fake / Rumor |
| 26,884 | In our time of rejoicing, let us not forget the families of Martin Richard, Lingzi Lu, Krystle Campbell and Officer Sean Collier. | True |
| 20,183 | I will DONATE $100 for EVERY pass I catch next season to whatever Boston Marathon Relief Fund there is. And $200 for any dropped pass. | True |
| 18,727 | Doctors: bombs contained pellets, shrapnel and nails that hit victims #BostonMarathon @NBC6 | True |
| 17,673 | #prayforBoston | NA |
| 17,560 | For every retweet I will donate 2 to the Boston marathon tragedy! R.I.P! | Fake / Rumor |
| 16,457 | From Sarasota to Boston, our thoughts go to the victims of the marathon bombings. We're saddened by loss of life and injuries to so many.... | NA |
| 13,788 | So far this week- #prayfortexas - #prayforboston - two 14 year olds killed a homeless man as a dare- bomb threats It's only Thursday | True |
| 13,610 | Jhar #manhunt @J_tsar. Look at this from a follower. Look at the time if the tweet http://t.co/xgnAJpeVTr | NA |
| 13,482 | BREAKING: Suspect #1 in Boston Marathon bombing shot and killed by police. Suspect #2 on the run, massive manhunt underway. | True |
| 13,275 | #prayforboston | NA |
| 12,354 | BREAKING: An arrest has been made in the Boston Marathon bombings, CNN reports. | True |
| 12,209 | R.I.P. to the 8 year-old boy who died in Boston's explosions, while running for the Sandy Hook kids. #prayforboston http://t.co/Xmv2E81Lsb | Fake / Rumor |
| 11,950 | For each RETWEET this gets, $1 will be donated to the victims of the Boston Marathon Bombing. | Fake / Rumor |
| 11,036 | #WANTED: Updated photo of 19 year-old Dzhokhar Tsarnaev released. Suspect considered armed & dangerous. http://t.co/pzps8ovJTb | True |

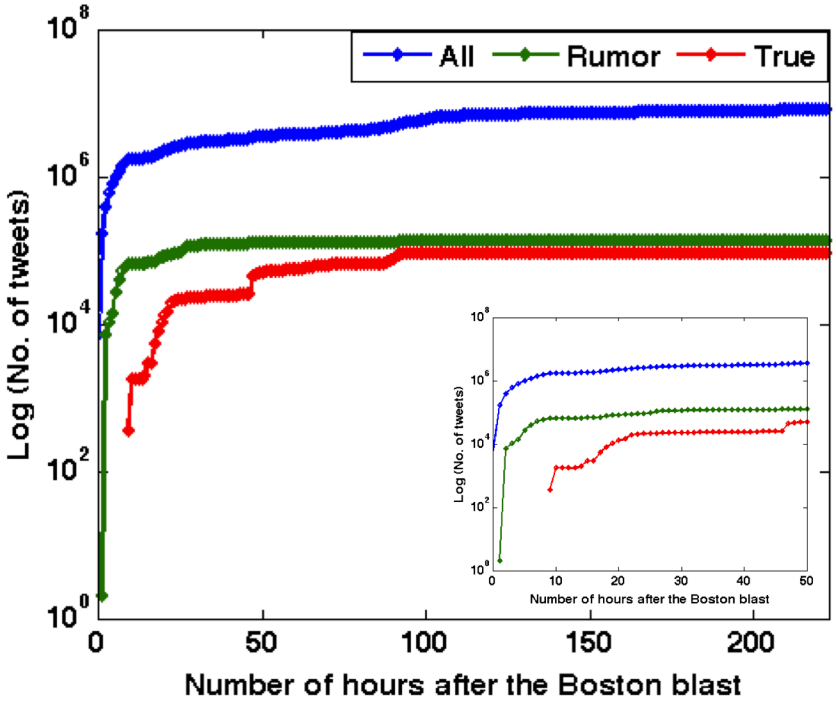



Figure 5.6: The log distribution for the number of the total, fake, true information and NA category tweets. The inset figure presents the results for the first 50 hours after the blast.

have a very steep growth. This may be attributed to the fact that the user profiles (source of a fake tweet) are people with low social status and unconfirmed identity. Hence, initially fake tweet spread is slow, and they become highly viral only after some users with high reach (for e.g. large number of followers) propagate them further.

### 5.3.2 Fake Tweet Seed User Profiles

We analyzed the attributes and activities of the user profiles from where the fake tweets originated. Table 5.3 presents the various profile attributes for the seed users, who initiated the fake content tweet. Of the six fake tweets identified, two set of tweets were started by same user profile. For most of the fake tweets we observe that the seed users are people with very few followers. *Seed 4* is the only user profile with high number of followers. The tweet posted by *seed 4* was *Reports of Marathon Runners that crossed finish line and continued to run to Mass General Hospital to give blood to victims #PrayforBoston.* This tweet even though was false and classified as fake content by the media too [80], was harmless and not deleted by Twitter. For all other sources, except *seed 4* we can say that the originators of the fake content are users with low credibility. We checked for the presence of these seed user profiles on Twitter now; all accounts except *seed 4* have been suspended by Twitter.

Table 5.3: Descriptive statistics of the four user accounts that were the seeds of the six fake content tweets.

| | **Seed 1** | **Seed 2** | **Seed 3** | **Seed 4** |
|---|---|---|---|---|
| **Number of Followers** | 10 | 297 | 249 | 73,657 |
| **Profile Creation Date** | Mar 24 2013 | Apr 15 2013 | Feb 07 2013 | Dec 04 2008 |
| **Number of Statuses** | 2 | 2 | 294 | 7,411 |
| **Number of Fake Tweets** | 2 | 2 | 1 | 1 |
| **Current Status** | Suspended | Suspended | Suspended | Active |

### 5.3.3 Tweet Source Analysis

We studied the source of the tweets that were posted. We analyzed the devices (web / mobile / TweetDeck) through which the tweets were posted. Figure 5.7 shows the results. We found that the

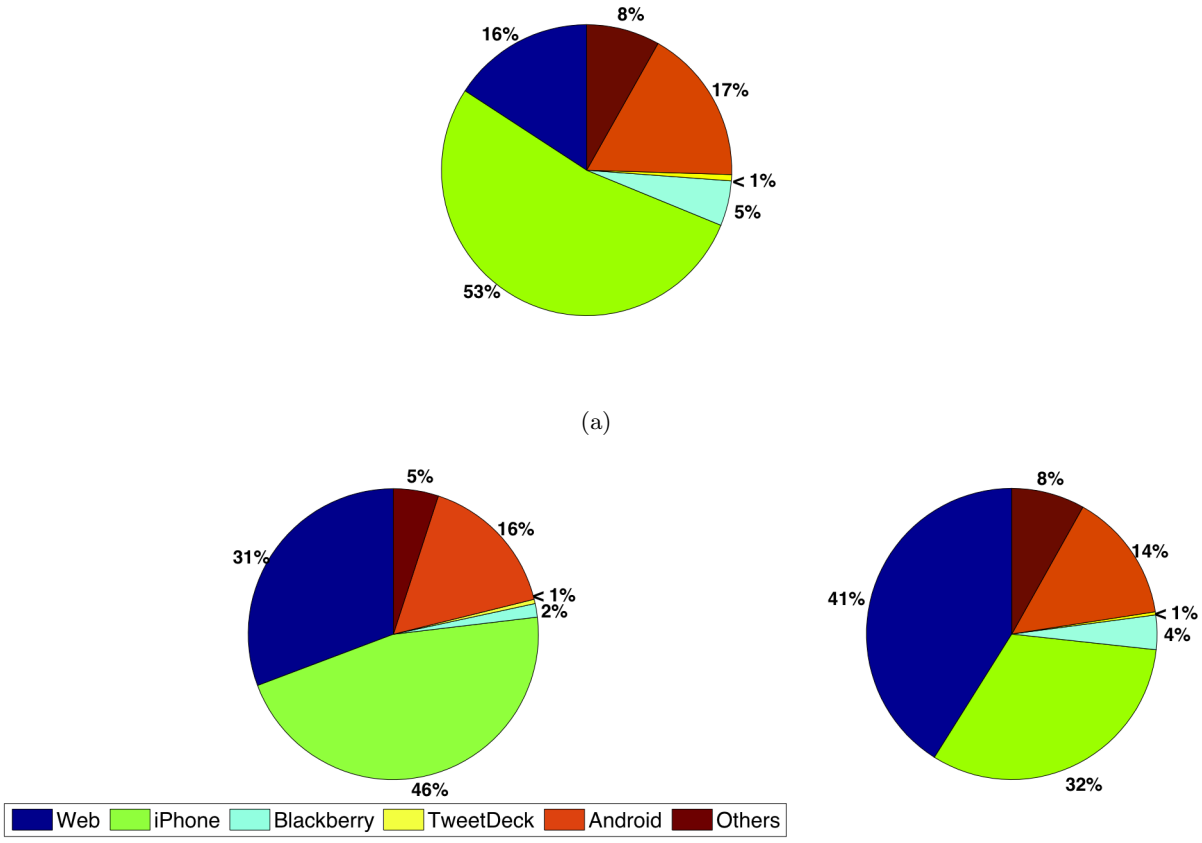



Figure 5.7: Tweet source analysis for the three categories of tweets: (a) Fake (b) True (c) NA. We observed that in case of fake tweets, approx. 75% users use mobile devices: iPhone (53%) + Blackberry (5%) + Android (17%), as compared to 64% for true and 50% for NA category of tweets.

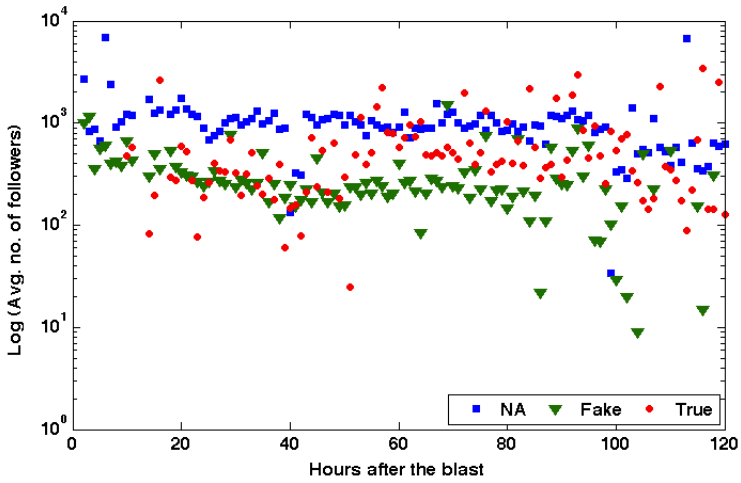
tweets containing information (fake or true) propagated more through mobile devices like iPhone, Android, Blackberry, etc. whereas the general non-informative tweets (NA category) were posted more via web interface. We found that approx. 75% of fake tweets are propagated via mobile phone devices, as compared to 64% true tweets and only 50% generic comments shared via mobile devices. This implies that people are eager to share the informative tweets and also willing to do that while being on the go. For non-informative tweets, people don't feel such urgency and post tweets mostly if they are accessing through a web interface.

### 5.3.4 Role of User Attributes in Fake Content Identification

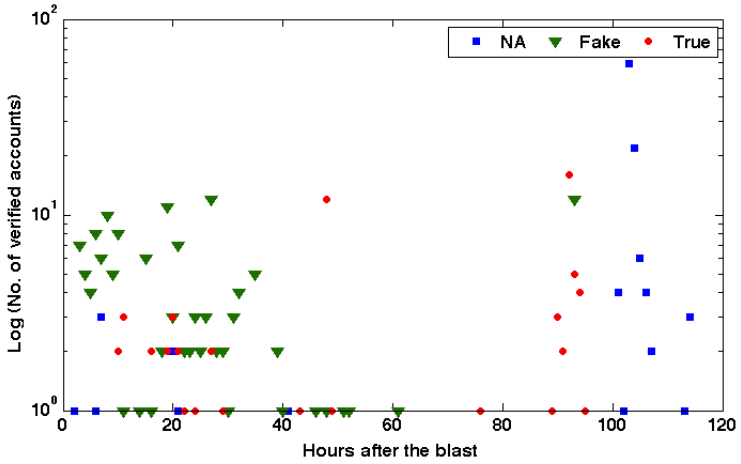
To understand what kind of users aid in propagation of each category of tweets, we analyzed three main attributes of user profiles on Twitter. We calculated the average number of followers of the user accounts and the number of verified accounts that propagated the true, fake and NA tweets. Figure 5.8 summarizes the results for the first 120 hours after the blasts. We observed that the average number of followers is the maximum for NA tweets, followed by true and fake tweets. Even though high number of users tweet generic news, the rumors get more viral. Number of people retweeting fake content tweets drops significantly in the latter hours (80-120 hours), this maybe so, as people start realizing that the information is incorrect. We also observed that a high number of verified accounts propagate fake content, which is quite surprising. We can conclude that determining whether some information is true or fake, based on only factors based on high number of followers and verified accounts is not possible, at least in the initial hours after an event. The high number of verified and large follower base users propagating the fake information, can be considered as the reason for the fake tweets becoming so viral. It becomes difficult for the users to differentiate which sources to trust and which not. In the next section, we validate this hypothesis, by exploring if these user attributes can be used to estimate how viral a rumor / fake information tweet would become in future.

### 5.3.5 Role of User Attributes in Propagating Fake Content

We aim to understand, if user attributes can be used to estimate how viral fake content would become in future. Knowledge about how viral and damaging fake content can be in future can help us be prepared in handling its consequences. In addition to basic user attributes like number of followers, friends, verified accounts, etc. we also define and compute an overall *impact* metric; to measure impact of users who propagate a tweet in making it viral. We used user profile attributes to define a metric which calculates the overall *impact* of a user. We take the *impact* of user $u_i$, as a linear combination of the following metrics:




(a)




(b)

Figure 5.8: Distribution of (a) average number of followers of the user accounts and (b) number of verified accounts in the annotated data.

- **Social Reputation**: We measure social reputation as a function of the number of followers and the number of times the user has been listed. Number of followers denote the popularity of the users and number of times listed indicate how many people classified him in one of the custom lists.

$$SocialReputation(SR(u_i)) \tag{5.1}$$

$$= \frac{log(n(followers))}{Max(log(n(followers)))} + \frac{n(listed)}{Max(n(listed))}$$

- **Global Engagement**: It is measured based on how often does the user engage with posting activity on Twitter by tweeting, replying and retweeting. We take it as the ratio of the number of statuses the person has put to the time (in hours) since his account creation.

$$GlobalEngagement(GE(u_i)) = \frac{n(status)}{age} \tag{5.2}$$

- **Topic Engagement**: This metric measures how well a user is engaged in the current topic of tweets. For our context, the topic is the event under consideration. We measure this by number of tweets the user has posted on the particular topic.

$$TopicEngagement(TE(u_i)) = \frac{n(status_t)}{max(n(status_t))} \tag{5.3}$$

- **Likability**: The Likability of a user is to measure of how much his content is liked by his followers or other users. We take it as the ratio of number of times his statuses have been made favorite, to that of number of statuses posted.

$$Likability(L(u_i)) = \frac{n(favorited)}{n(status)} \tag{5.4}$$

- **Credibility**: Credibility $C(u_i)$ of a tweet is based on, if the user who posted it is verified or not. We take it to be 0, if the account is not verified by Twitter, else we take it as 1.

We define *impact* for a user, $u_i$, as a linear combination of the above mentioned components.

$$Impact(u_i) = SR(u_i) + GE(u_i) + TE(u_i) + L(u_i) + C(u_i) \tag{5.5}$$

**Regression Analysis**: We predict the attributes of propagation for a true or fake tweet in future based on its current attributes. We used linear regression for this purpose. Our basic hypothesis is that $Impact$ of all active users in previous time segment can be used to predict how many new users will get activated in the next time segment. For calculating the regression between attributes and growth of fake tweets, we consider time quantum of 30 minutes each. For a particular time quantum, all users will have a similar contribution towards the cumulative $Impact$, so we weigh the cumulative impact according to the Poisson distribution.

$$CummulativeImpact(t) = \sum_{i=1}^{t-1} Impact(t-i) \times \exp^{(t-i)} \tag{5.6}$$

We estimate the number of people that are going to be activated in the next time segment using Linear Regression as follows:

$$N_{Active}(t+1) = \alpha \times CummulativeImpact(t) + \beta \tag{5.7}$$

For evaluation of linear regression, we used $R^2$ measure. We used $R^2$ metric to measure the confidence with which the model accounts for variability in the output data. Results of the model were compared with individual features as well and are presented in Figure 5.9. On an average for impact metric we achieve approx. 0.7 value of $R^2$. These results show us that it is possible to predict how viral a fake information tweet would become in future based on the attributes of the users currently propagating the fake content.

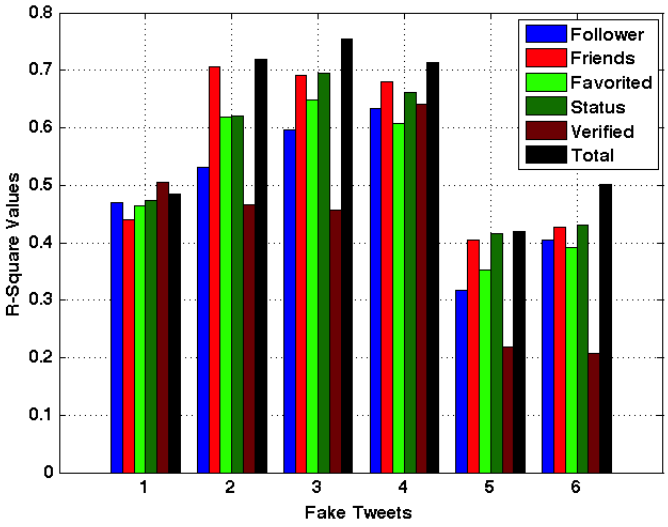



Figure 5.9: Regression results of the overall impact of the users in previous time quantum. These results show us that it is possible to predict how viral the fake content would become in future based on the attributes of the users currently propagating the fake content.

## 5.4 Suspended Profiles Analysis

Thousands of new accounts on Twitter get created everyday, many of these accounts are often malicious and spam accounts. In this section, we aim to identify the characteristics and activities of malicious new accounts created during the Boston marathon blasts. We identified 31,919 new Twitter accounts that were created during the Boston blasts tragedy (Apr. 15th - Apr. 20th), that also tweeted atleast one tweet about the event. Out of these 19% (6,073 user profiles) were deleted or suspended by Twitter, when we checked two months after the blasts. Some of these accounts were quite influential during the event. We also tried to measure how effective were these

accounts during the Boston marathon events. Figure 5.10 shows the number of suspended profiles created in the hours after the blast. We observed that there are a lot of malicious profiles created just after the event occurs. Such profiles and accounts aim to exploit the high volume of content and interest of users in the event to spread spam, phishing and rumors. We constructed a network graph $G = (V, E)$ for the interaction between these newly created malicious profiles. Where each node in $V$ represents a suspended user profile, and an edge between two suspended nodes represents a retweet, reply or mention by them.

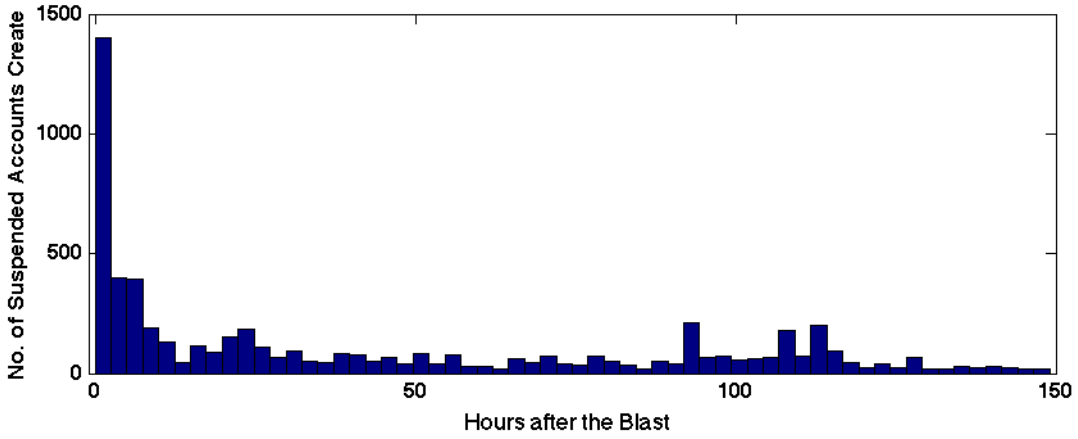
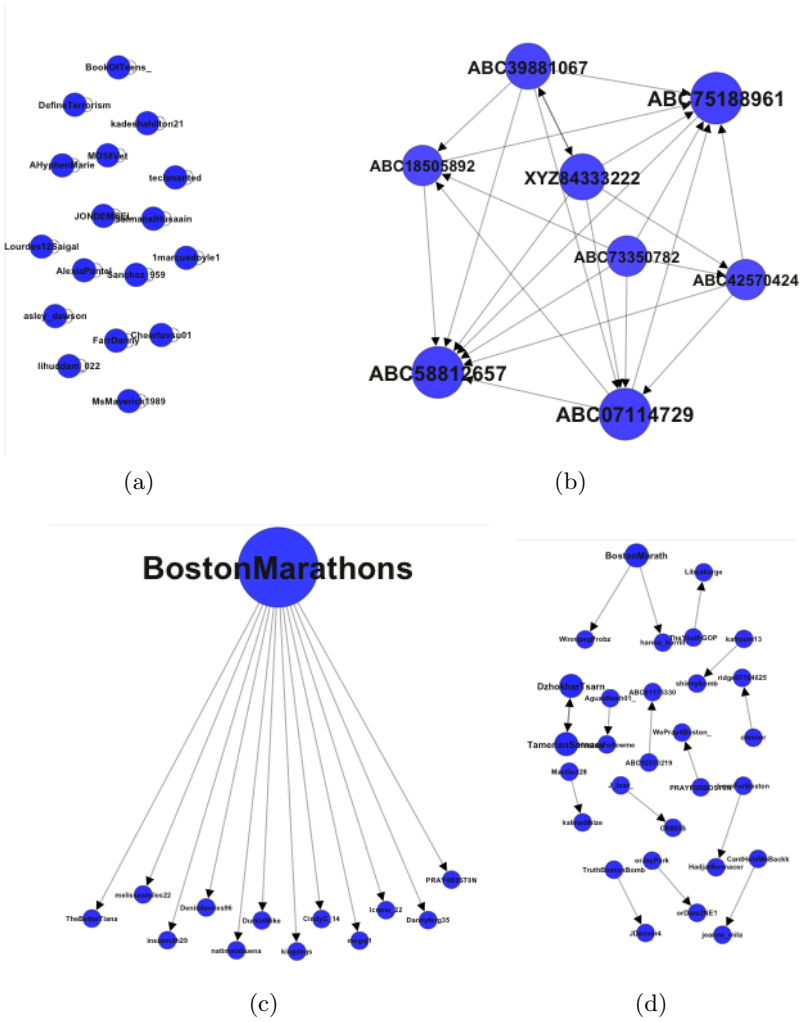



Figure 5.10: Number of suspended profiles created in the hours after the blast. We observed that there are a lot of malicious profiles created just after the event occurs.

Figure 5.11 shows the network obtained (some of the usernames are anonymized). We have removed all nodes with degree of zero, we found 69 nodes out of 6,000 suspended profiles had an edge to another node in the graph. Though the number of edges may look small, but we observed some interesting characteristics of the resulting network graph formed. We found four types of interactions amongst these accounts (left to right in Figure 5.11):

- *Single Links:* We saw the content posted by a suspended user profile is propagated by one or two other suspended Twitter profiles. Some of these links are also bi-directional, indicating a mutual agreement between the nodes. This technique of creating multiple spam accounts to promote common content is commonly used by spammers on Twitter [40].

- *Closed Community:* We observed a community of users who retweet and mention each other, and form a closed community, as indicated by high closeness centrality values for the nodes. All these nodes have similar usernames too, all usernames have the same prefix and only numbers in the suffixes are different. This indicates that either these profiles were created by same or similar minded people for posting common propaganda posts. We then analyzed the text posted by these users. These twelve accounts were all tweeting the same propaganda and hate filled tweet. Since, Twitter does not allow users to post multiple posts with same




(a) (b) (c) (d)

Figure 5.11: Network of suspended accounts (retweets / replies / mentions) created during the Boston blasts. We observed four different forms of interactions amongst the suspended profiles (left to right): Single Links, Closed Community, Star Topology and Self Loops.

content, another strategy applied by these accounts is tweeting the same text as above, but changing one character in each of the tweets. In all we found, 156 tweets by these 12 accounts.

- *Star Topology:* A fake account *@BostonMarathons* was created similar to the original Twitter account *@bostonmarathon*, resulting in users getting confused between the two, leading to a lot of content being circulated by the fake BostonMarathons profile. Impersonation or creating fake online profiles is a crime that results in identity theft and is punishable by law in many countries.

- *Self Loops:* We saw that some of the profiles mentioned themselves in their tweets, resulting in self loops in the graph. This may be done by the users to promote the content posted by

themselves.

We saw that a large number of malicious accounts were created during crisis events. Next, amongst the suspended user profiles we searched for profile specifically created for exploiting the event. Some of them created related hoax profiles by using usernames similar to original accounts. We searched for the presence of the term *boston* in the name and username of the six thousand suspended profiles. We found 83 profiles which satisfied this criteria. Figure 5.12 shows the tag-cloud of the user description of these profiles. We found most of these profiles exploited the sympathy of people by using words such as *prayforboston, prayers, victims*. The malicious intent of such user profiles is clear, as they attempt to create hoax accounts, as indicated by usage of words such as *official account*. The account *BostonMarathons* was also one such account which tried to impersonate the real *bostonmarathon* account.

(a)

(b)

Figure 5.12: Tag cloud of (a) user description and; (b) tweets posted by the suspended accounts created targeting specifically the Boston event. In the left tag cloud, we see the usage of words such as *official account* indicating profiles were created impersonating real profiles.

## 5.5 Discussion

Rumors or fake or incorrect information spread via online social media, have resulted in chaos and damage to people in real-world. Specially, during crisis events like earthquakes, bomb blasts and political uprisings, rumors can be very harmful. Malicious entities exploit the vulnerable emotions of people during crisis to make their rumors viral. Online social media, in particular, Twitter, is a

mass media with reach to millions of users across the globe. Over recent years, misinformation on Twitter had resulted in damages ranging from financial to human lives. Detection and curbing of fake information on social media, is a relatively new and unexplored domain.

Our aim in this chapter was to characterize and provide useful insights into the domain of fake content propagation on Twitter. We collected about 7.9 million tweets for the Boston marathon blasts using the Twitter APIs. Our data collection was limited from the fact that it was started 45 minutes after the blasts had occurred. We analyzed source users of the fake tweets, spatial and temporal details of the fake tweets. We attempted to find out reasons that govern how viral (and in turn harmful) a fake information tweet becomes. We explored, using simple metrics, can we predict how the fake tweet would propagate in future. Another kind of fake content that is present on Twitter are the fake / spam user profiles. We analyzed six thousand malicious profiles that were created on Twitter right after the Boston blasts and were later suspended by Twitter.

Some of the major challenges in real time rumor detection and control on online social media are the following:

- *Volume of Content*: Most of the popular online social services have users of the order of hundreds of millions. A huge amount of content is generated every second, minute and hour of the day. Any algorithms or solutions build to detect rumors on OSM should be scalable enough to process content and user data up to the order of millions and billions.
- *Short Impact Time*: Impact of malicious activities in online social media, such as, spread of spam, phishing or fake content, causes vast amount of damage within hours of being introduced. Hence, solutions and algorithms built need to be able to solve and detect such content in real-time. Post-analysis may be able to capture concerned criminals, but would not be able to contain the damage.
- *Anonymity*: The Internet, by itself is an anonymous medium. Lack of tracing technologies, distributed framework and private regulations worsens the condition. People can create one or multiple profiles on OSM, with little or no verification required. Hence, validation and authenticity of any information source on OSM is extremely challenging.

We observed some differences in temporal propagation of rumors and true news. Since fake content tweets are often started by people with low number of followers on Twitter, they have a slow growth in the beginning, and then become viral as they are retweeted by higher influence people. While true news were mostly generated by accounts such as *boston_police* which already had 276,838 number of followers and hence, the true news have a more steady growth from the beginning. Spatial analysis did not give us much insights since the annotated tweets for the category of Fake / True / NA (Not Applicable) categories did not have much tweets with geo-tagged information. This was

surprising, we found only 1 tweet out of approx. 500,000 tweets containing geo-location data. We found that approx. 75% of fake tweets are propagated via mobile phone devices. Next, we explored, if attributes of user profiles who propagate fake information can be used to answer the following two questions: Firstly, we identified fake tweets in the Twitter network using user profile attributes; and secondly, we used the attributes of user profiles to predict which rumor / fake tweet would become viral in the future on Twitter. For the first question, we observed that the user attributes are quite similar for fake, true news and NA category tweets, like a lot of verified and high number of followers accounts also retweet rumor tweets. Hence, user attributes by themselves were not very useful in identifying the fake tweets. For the second question on use of user attributes for predicting how viral fake content will get in future gave us more positive results. We used logistic regression to estimate the popularity of a fake tweet in future (30 mins to 1 hour in advance), based on the overall impact and other characteristics of the users who are currently propagating the fake tweets.

We observed that a great number of malicious profiles were created just after the Boston blasts, which were later suspended by Twitter. Out of the six thousand such profiles created (who tweeted about the event), we found about 83 profiles which further targeted the blasts, by creating usernames having the word 'boston' in it. On manual observation, we found many of these accounts to be fake accounts impersonating the real accounts (e.g. fake username: *bostonmarathons*, real username: *bostonmarathon*). The network analysis of the interactions among these suspended profiles revealed interesting techniques applied by these users to spread their propaganda or spam; for e.g. we found a community of users who had similar usernames, and posted the same hatred targeting tweets multiple times. They also retweeted and replied to each other's tweets, to increase their visibility.

# Chapter 6

# Analyzing Role of Twitter during India Centric Events

The goal of this chapter is to analyze the role of Twitter during India centric events and extract information that maybe useful for emergency responders, authorities and citizens. The roadmap of this chapter is as follows: Section 6.1 explains the need for studying role of Twitter in India centric events. Section 6.2 discusses analysis and results obtained from analyzing Twitter data for one of the events in detail: Mumbai blasts of 2011. The last section introduces a system Twit-Digest developed by us to aid emergency responders and authorities to analyze and visualize Twitter data. The first version of system Twit-Digest was presented and demoed at IBM Collaborative Academia Research Exchange (IBM-ICARE), 2012 [9]. It was also presented as an invited poster at International Conference on Management of Data (COMAD), 2012. The second version of Twit-Digest was presented at 2nd Security and Privacy Symposium IIT, Kanpur, 2014 [43].

## 6.1 Analyzing Twitter Data during India Centric Events

Presence and role of OSM in developing countries such as India has increased multifold over last few years. Today government departments, journalists, news media, celebrities and citizens, all use OSM services actively. The adaptation of OSM in India for use of governing, policing and emergency management is still at its nascent stage and not fully understood by the concerned user groups. The official handle for the Prime Minister and President of India was only created in 2012 and 2014 respectively. [1] One of the first incidents of Indian government blocking Twitter handles

[1]Twitter handle @pmoindia was created on 23rd Jan, 2012, and @rashtrapatibhvn was created on 1st July, 2014.

at Internet Service Providers' (ISP) level occurred during the Assam riots of 2012 [6]. Even though the blocking of Twitter account was only partial, since it was only blocked by ISPs and not Twitter.

As part of this thesis, we also performed analysis on the role of Twitter during some events in India. Table 6.1 lists most of the prominent events analyzed. We studied the activity patterns and attributes of use of Twitter during these events which would aid authorities and emergency responders. In this chapter, we present in detail a case study of triple bomb blasts in Mumbai, 2011. This event was one of the first events in India during which Twitter played a pivotal role. It was effectively used for both good (relief and aid coordination) and bad reasons (spreading misinformation). We performed similar analysis on Twitter data collected for all events listed in Table 6.1.

Table 6.1: Some prominent India centric events analyzed by us.

| **Event** | **Date** |
|---|---|
| Triple Mumbai Blasts | July 2011 |
| Lokpal Bill Agitation | August 2011 |
| Assam Riots | April 2012 |
| Delhi Gang Rape Protests | December 2012 |
| Earthquake in North India | April 2013 |
| Uttrakhand Floods | June 2013 |

## 6.2 Case Study: Mumbai Blasts, 2011

The city of Mumbai (India) was made target of triple bomb blasts by terrorists on July 13th, 2011. Seventeen people were killed and 130 were injured in these blasts. All three blasts took place within a time period of 10 minutes (the fist took place at 13:24 hours at Zaveri bazaar, second at 13:25 hours at Opera House and third at 13:35 hours at Dadar). [2] What followed the blasts, apart from the loss of lives, injuries and loss of assets, was a complete state of panic and chaos in the entire city of Mumbai. There was commotion on roads, telephone lines were jammed and people were left stranded and helpless. At such times, Internet and social media came to the rescue of many. Soon after the explosions, the activity on online social websites increased manifold, users on Facebook and Twitter posted updates and information about their whereabouts, well being and offering help to those in need [5, 29]. Figure 6.1 is the first tweet captured by us on Mumbai blasts, within 5 minutes of the last blast. Social media platform, emerged as a huge data source of information about what followed as an after math of the events. A Twitter user, Nitin Sagar [3] created a spreadsheet on Google to coordinate relief operation among people. Within hours hundreds of

[2] All time stamps mentioned in this chapter are present in GMT.
[3] @nitinsgr on Twitter

people registered on the sheet via Twitter. People asked for or offered help on that spread-sheet for many hours. Similarly, a Twitter user created a disaster tracker map that enabled users to crowd-source information for crisis management from mediums like Twitter. [4]

Figure 6.1: The first tweets captured by us about the attacks after the Mumbai blasts; captured 5 minutes after the last bomb attack.

### 6.2.1 Data Collection

We collected data on the Mumbai terror attacks from July 13th till 17th July, as long as *#mumbaiblast* remained as a trending topic in India, the first tweet on the Mumbai blast that we captured in our dataset was after 5 minutes of the last blast. Castillo et al. also used a similar technique, of collecting tweet data using current trending topics, in their research work assessing credibility of news topics on Twitter [16]. We queried Twitter *trends API* after every 3 hours for new trending topics, and collected tweets corresponding to them as query search parameters for the *streaming API* of Twitter. Few of the trending topics related to Mumbai blasts were *#mumbai, #mumbaiblasts, dadar, kasab, opera house, andheri, #needhelp, #here2help*. As the majority of tweets were posted just after the incident, the dataset we will consider for this chapter is for the tweets which were posted on Twitter for the Mumbai blast on 13th, 14th and 15th July, 2011. In total, we collected 2,711,061 tweets by 1,172,766 unique users. Out of these tweets related to Mumbai blasts were 100,075 tweets, tweeted by 51,633 unique users during these three days. Words related to Mumbai blasts were also trending for some duration on 16th and 17th July, but the data posted during these days was quite less and adding that would have lead to no difference in results. Table 6.2 gives the descriptive statistics of these tweets. On 13th July, there were 53,809 tweets posted on Twitter related to Mumbai blasts. On 14th and 15th July, there were 45,162 and 973 tweets collected by us respectively.

[4] http://maps.myindia.bz/mumbai/

Table 6.2: Descriptive statistics for the Twitter dataset for 13th, 14th and 15th July, 2011.

| | |
|---|---|
| **Total tweets** | 2,711,061 |
| **Total unique users** | 1,172,766 |
| **Total tweets for Mumbai Blasts** | 100,075 |
| **Total unique users for Mumbai Blasts** | 51,633 |
| **First tweet in our dataset at** | Wed Jul 13 13:40:44 2011 |
| **Last tweet in our dataset at** | Fri Jul 15 21:41:21 2011 |
| **Unique URLs shared** | 19,173 |
| **Number of singleton tweets** | 68,739 |
| **Number of retweets** | 31,336 |

### 6.2.2 Analysis

We first analyzed the content of the tweets like the words, URLs and @-mentions in a tweet. Next we explored the relationships between tweets, users and their attributes like number of followers, friends, etc.

#### Content Analysis

We analyzed properties of tweet content posted during the Mumbai blasts. We analyzed the @-mentions and vocabulary of words in the tweets in our dataset and compared them to earlier work done on the subject.

**URLs and @-mentions** The posts on Twitter can contain URLs along with text in the tweets. Users share URLs of news articles, images, etc. In the 100,075 tweets we collected, for the analysis that follows we will consider only 68,739 tweets (which are not retweets). We observed that nearly 40.28% of tweets (27,693) contained URL links and 28.16% tweets (19,363) contained @-mentions in them. In the work done by Hughes et al., they showed that during a crisis situation the number of tweets containing URLs increase than in normal situations (40-50% tweets contained URLs in case of emergency situation, while 25% in normal situations) [3]. The number of tweets containing @-mentions decreased to 6-8% in emergency scenarios compared to normal 22%. The percentage of tweets containing URLs agrees with previous literature, but the percentage of tweets containing @-mention is not in line with the previous literature. We found 28.16% tweets excluding retweets that contained @-mentions in them, which is higher than what Huges et al. found for normal circumstances. Our results highlight the fact that a higher number of people share external resources via URLs and reply or mention other users in their tweets.

#### Vocabulary Analysis

We performed a vocabulary analysis of the tweets by studying the frequency of words in tweets

day-wise for the three days under consideration. A tag-cloud of such terms indicates the thoughts and sentiments of people on each day. Figure 6.2 shows the tag cloud of top 50 terms of used in all tweets, after removing stop-words, trending topics like *#mumbaiblast* and slangs. We saw some characteristic differences in the terms in tweet on all three days.

(a)

(b)

(c)

Figure 6.2: Tag clouds showing the vocabulary analysis for the tweets, that is 50 most frequently occurring terms in the tweets on that day. (a) July 13th, 2011, 13:25-23:59 (b) July 14th, 2011, 00:00-23:59 (c) July 15th, 2011, 00:00-23:59

Words like *injured, attack, help, here2help and hospital* were among the most prominent words on day 1. There were tweets of mainly two topics popular on day 1, first which propagated news about the attack and number of people injured and the second which were from users who were trying to coordinate help to those affected during the blasts. The tag cloud of day 3, presented a different composition of topics than day 1, which users tweeted regarding Mumbai blasts. Words like *Ajmal Kasab, hang, govt, protect, request* were most frequently used. For example, two of the most popular tweets on day 3 were:

- *#DearPM we don't ask u to hang Ajmal Kasab. We just request u to Protect us in same way as u protect him. Sincerely, Afraid Indians.*
- *Dear #RahulGandhi : 99% of the countries in the world would have hanged or executed Afzal & Kasab by now. But alas.. we live in the other 1%*

The tweets mentioned above from day 3 reflected the sentiments of people about the actions taken by the government. Day 2 saw an intermix of topics from both day 1 and day 3.

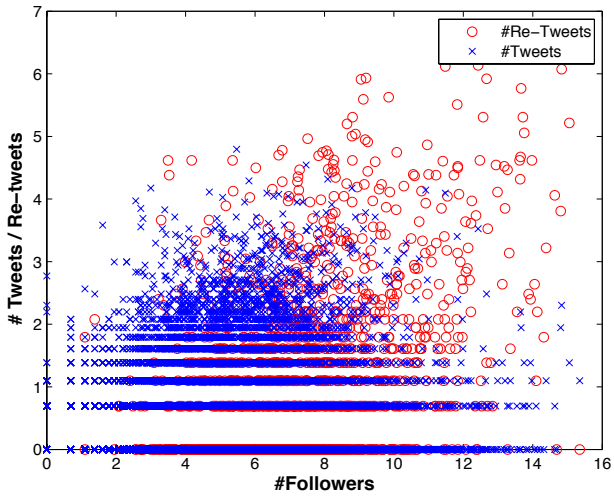
**Information Sources during the Mumbai Blasts**

In this section, we present the results of analyzing the source or users who posted content on Twitter. We used attributes such as number of retweets, friends and followers of a user to act as indicators to the truthfulness / reliability of tweet posts done by her. Secondly, we used @-mentions in Twitter; @-mentions may be used by a user in her tweets if its a retweet, a reply to someone's tweet or marking certain user in one's tweets.

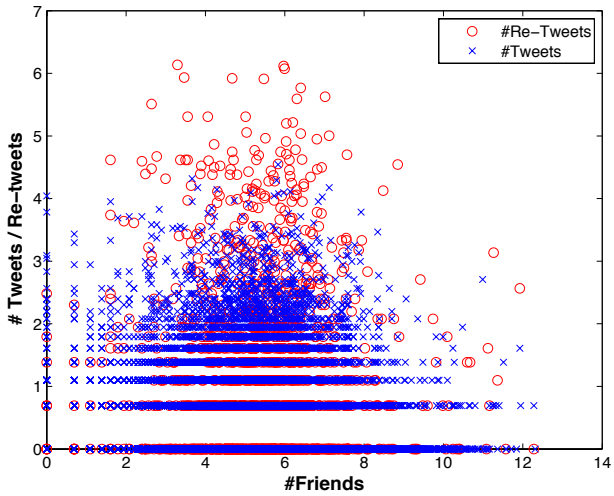
At the time of crisis, people look out for information from reliable sources like government agency owned accounts, news media and celebrities. Most of such accounts have a huge follower base and low number of friends. We can safely assume that, information coming from such sources to be more reliable and credible or atleast not be spam to a topic. Hence, we will now explore the relationship between the number of followers and the tweet activity by these users and the retweet counts of their tweets. Figure 6.3 shows the relationship between tweets / retweets and the number of followers and friends a user has. Since, the number of followers varied from 1 to 6 million, the number of tweets by each user had a maximum value of 121, and the number of retweet count varied from 0 to 1,600, we took log of values while plotting the graph, to improve the visualization.

We observed from Figure 6.3(a) that the number of tweets by people with large followers (who are generally like celebrities, media companies and government owned account) are quite less. Though the number of posts were less by such users with large number of followers, these posts registered high number of retweets.

Next, we found which users on Twitter post maximum status updates, or who are the top Twitter




(a) Follwers v/s Tweets / retweets




(b) Friends v/s Tweets / retweets

Figure 6.3: The left graph shows that the number of tweets by authority users (i.e. users with a large number of followers) are very less. The number of retweets for the tweets done by them are quite high.

users. We considered only the new tweets being generated, excluding retweets. We can observe the relationship between the top users and their number of followers and friends in Figure 6.4. In an earlier work done by Mendoza et al., they showed a very steady decay in number of tweets as the number of followers decrease. In our results, though we find that the average number of followers of top 50 users is above 9,000, but after that there is a sharp decline in number of posts. Thus, between top 100 and 350 users, we see the average number of followers is quite low.

Though we must point out, that most of earlier work done was on developed nations like US, France and Japan, hence the differences maybe attributed to the fact that government and media in developing nations like India, have only recently started becoming active on social media like Twitter. So there are only handful of such authority users in India, who proactively tweet about or during incidents. In proportion, the activity of general users has increased manifolds.

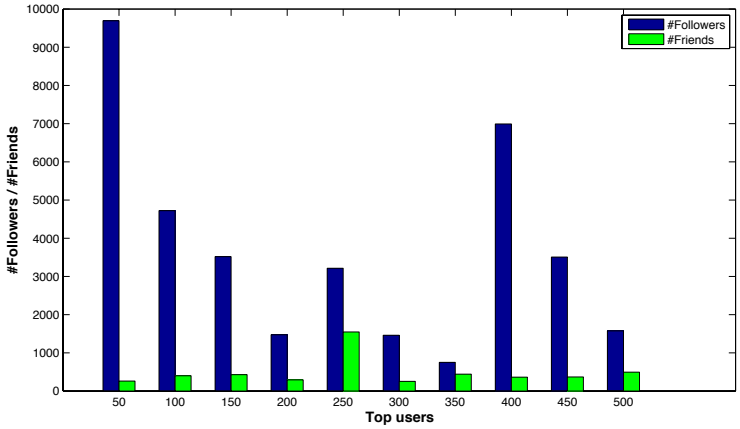



Figure 6.4: The variation in average number of friends and followers for the top 500 users (maximum number of tweets) during the Mumbai Blasts. We observe that the average number of followers of top 500 users of Twitter during the Mumbai blasts are very low. For users among the top 100 to top 350 the number of followers are less than 5,000.

### Rumors on Twitter During the Mumbai Blasts

At the time of such crisis, malicious elements take advantage of such situations in various ways. Like they take advantage of the emotional and sentimental vulnerabilities of people at such times to exploit them. They also instigate hatred and anger among people. Certain incidents happened during the Mumbai attacks, where there were false news, fake cries of help and negative sentiments that were propagated. We now discuss one such rumor in detail.

Rumor about terrorist Ajmal Kasab's birthday: A false news was reported on Twitter on Wednesday, that the Mumbai blasts have taken place on the eve of Ajmal Kasab's birthday. It was being portrayed that this event has been organized by Kasab's sympathizers. A sample tweet said: *Three blasts in Mumbai. It happens to be a birthday of the captured terrorist Kasab or was it planned to be on his birthday? #Mumbaiblasts.* There were 446 tweets and 223 retweets captured by us on this topic. We applied a key word based search among the tweets, to retrieve all tweets mentioning *Kasab's birthday* and its variants, that were posted on Wednesday, 13th July on Twitter. The first tweet regarding this on 13th July came at 14:17. Figure 6.5 depicts the number of tweets

and retweets activity during the first 24 hours of the rumor being spread (13:00 Wed - 14:00 Thu). People retweeted the news without knowing whether it is true or not, there was a high activity in tweets and retweets as soon as the rumor was started using the *#mumbaiblast* hash-tag.

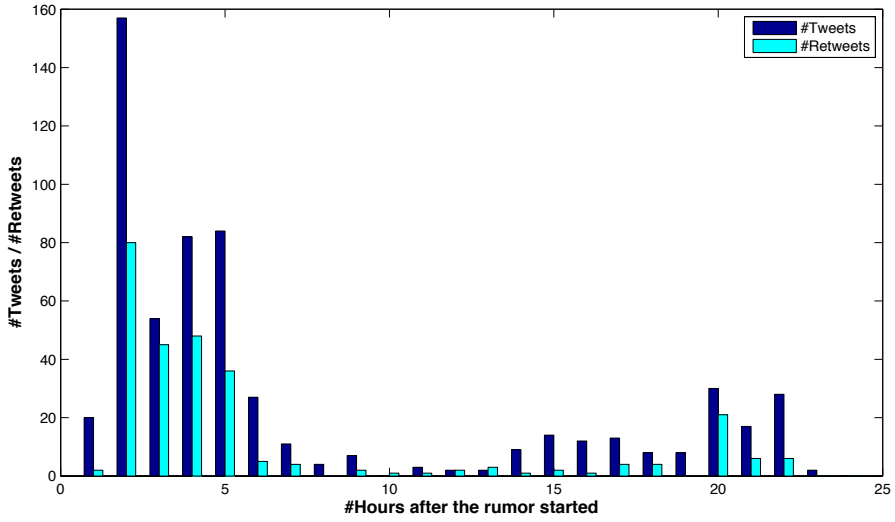



Figure 6.5: The activity regarding the rumors of 13th July (13:00 to 23:59 hours) being Kasab's birthday through the network of Twitter.

Some more rumors which propagated on Twitter during the Mumbai blasts:

- Rumors were spread about a fourth blast. Some tweets even specified locations of 4th blast as Lemington Street, Colaba and Charni. Around 500+ tweets and retweets were posted about this.
  - Sample tweet: *Fourth blast in mumbai!!! At lemington road! Let da focus be on innocent ppl*
- False blood required news spread Twitter. They were initiated by a user, @KapoorChetan and around 2,000 tweets and retweets were made regarding this by Twitter users. It led a lot of chaos, as many people reported at specified hospitals to donate blood. They were turned back by the hospital authorities as there blood-banks were sufficiently filled [89].
  - Sample tweet: *RT idurgesh: #needhelp #MumbaiBlasts RT KapoorChetan: Bombay Hospital Blood Bank is in need of B+, B-, O+ blood groups. Please donate*

**Twitter Monitoring using Social Network Analysis**

We proposed that social network analysis of Twitter users network and resource network can be helpful for authorities to analyze the community of users / topics during a crisis situation. In users network, we considered a graph, with users as nodes and links are present between two users by the presence of @-mentions in tweets (i.e. retweeting, replies and mentions) [14]. Table 6.3 gives the top 10 users for the above centrality measures in the user network for tweets on Mumbai blasts. Degree centrality of a node is the normalized sum of its row and column degrees. Those who are ranked high on this metrics have more connections to others in the same network.

Table 6.3: Top 10 users in the user network for tweets on Mumbai blasts by betweenness, closeness and degree centrality. Users are represented as nodes and the links are present between two users by the presence of @-mentions in tweets (i.e retweeting, replies and mentions).

| **Degree Centrality** | **Closeness Centrality** | **Betweenness Centrality** |
|---|---|---|
| Breaking News | Natasha Monteiro | calamur |
| CNN-IBN News | Z RAVITEJA REDDY | Bombay Addict |
| Bombay Addict | ur's TEJ | Netra Parikh |
| Kiran Manral | Santosh Kumar | Venkat Ananth |
| barkha dutt | Goutham Ram | Lydia Polgreen |
| Mid Day Infomedia | nitin c | Kiran Manral |
| NDTV | ShoutDeck | Anand Giridharadas |
| Sarita Tanwar | Shanti | kaveri ahuja |
| Anupam Kher | 9-win | Shailesh Tewari |
| Times of India | Ramana Achini | shiv singh |

Users with high betweenness centrality are users who are connectors between many other users. Closeness is the average closeness of a users (no. of links) to the other users in a network. Thus, calculating these measures helped in identifying the most influential users in a network. From the table we observe that, degree centrality represents mostly celebrities and media accounts. Since, a lot of users retweeted and replied to tweets posted by them, their degree centrality is higher. Advanced measures of closeness and betweenness helps us in identifying, other users, who play the role of connectors (betweenness) in the network. Social network analysis of the Twitter graph provide us with users and resources that play the most central role, and can be monitored by authorities or news media to represent the network.

## 6.3 Twit-Digest: An Online Solution for Analyzing and Visualizing Twitter

We introduce a real-time system Twit-Digest developed by us to aid emergency responders and authorities to analyze and visualize Twitter data. Twit-Digest is an online service, developed to provide an interface over the live Twitter stream. In addition to providing tweets from Twitter based on a search query requested by the user, Twit-Digest also provides a variety of analysis of the Twitter data. Analyses are aimed at providing an analyst with quick inferences and big-picture about the activity around the search query. The tool can be effectively used by security analysts, organizations, and professional agencies that want to monitor content of their interest on Twitter. One of the major contributions of this work, are the credibility and spam detection of all the tweets displayed to the analysts who use the Twit-Digest system. The algorithms for credibility and spam assessments compute the score for the data in real-time and presents the results to the user.

### 6.3.1 Twit-Digest Functionality

Twit-Digest provides analytical results for the Twitter data at two levels: First at the tweet level, and second at the topic level.

#### Tweet Level Functionality

For each tweet, Twit-Digest analyzes the content and user attributes of a tweet to assess the credibility score and identify if tweet is spam or not.

- **Credibility Score**: For each of the tweet extracted from Twitter containing the query word, Twit-Digest provides the credibility score. We compute the credibility of a tweet based on features such as the tweet content and user who tweeted it. This score calculation is done based on the algorithm developed and evaluated by us [47]. We adopted a supervised machine learning and relevance feedback approach, to rank tweets according to their credibility score. The value obtained from the above algorithm is mapped to a score of 1-5 and visually presented to user as shown in Figure 1. Less the number of stars imply less credibility.
- **Spam / Phishing Detection**: The tool validates whether a tweet is spam or phishing using two different techniques. Firstly, this is done with the help of GoogleSafeBrowsing and PhishTank APIs. [5] [6] Secondly, we use some of the results obtained from previous research works regarding spam and phishing detection on Twitter, for example, presence of more than

[5] http://www.phishtank.com/developer_info.php
[6] https://developers.google.com/safe-browsing/

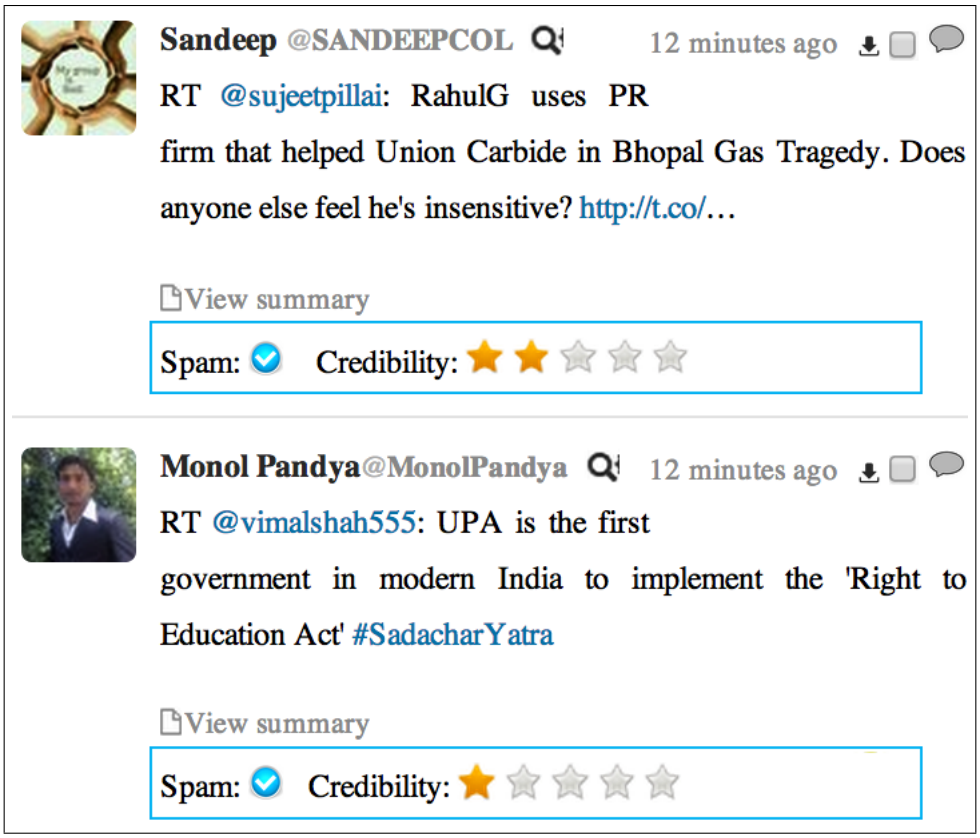



Figure 6.6: For each tweet credibility score is computed and tweets are marked as spam or not spam.

three hashtags [4, 8, 20]. Based on the above results, each Tweet is marked with *Green Tick* if it is not a Spam or a *Red Cross*, which indicates Spam / Phishing.

### Topic Specific Analysis

For the topic or keyword provided by a user, twit-Digest displays analysis of all tweets obtained for that topic or keyword.

- **Geographical Analysis**: Location of tweets and users are marked on a map to show the impact of an event on a particular country or continent using GoogleMaps API. On Twitter there are two primary ways in which it can be shared: firstly, as an user profile attribute called location; and secondly, as a geographic coordinate (latitude and longitude) associated with geographically tagged tweet of a user. A sample geo-location output by Twit-Digest is shown in Figure 2.
- **Network Analysis**: In user based network analysis, Twit-Digest presents a graph, with users as nodes and presence of retweets or @-mentions in tweets (i.e. retweeting, replies and

Figure 6.7: Sample geographical distribution of the tweets and users corresponding to a query.

mentions) as edges between two user nodes. Such graphs visualization helps in identifying the influencers and communities of users who are posting messages about the topic and driving the discussions on the Twitter. For social movements and campaigns organized via Twitter, we often observed very dense graphs, with a few nodes driving the discussion, and they are retweeted or replied by all other nodes. Figure 6.8 shows the snapshot of the network graph from Twit-Digest.

- **Top Words Analysis**: Since many a times, the user may not be aware of the exact query she wants to analyze or all the components of a news event, we provide a query expansion feature in Twit-Digest. When a user searches for a word, we extract all the hashtags from the tweets containing that query word, and present to the user with a tag-cloud of 50 top hashtags related to her based on the TF-IDF score [91]. The size of the word in the tag-cloud indicates its TF-IDF score. Figure 4 represents the tag-cloud for a query.

- **Statistical and Popular Link Analysis**: We also present the user with cumulative graphs representing various statistical analysis, such as: piechart to represent percentage of tweets versus retweets, scatter chart to show the relation between number of retweets and follower count, column chart between users and their follower count and lastly, bar chart to show frequency of various URLs present in tweets. In the popular link analysis tab, we display the link and preview of top URLs (based on frequency of occurrence) posted about the event and also classify them as images, videos and news links.

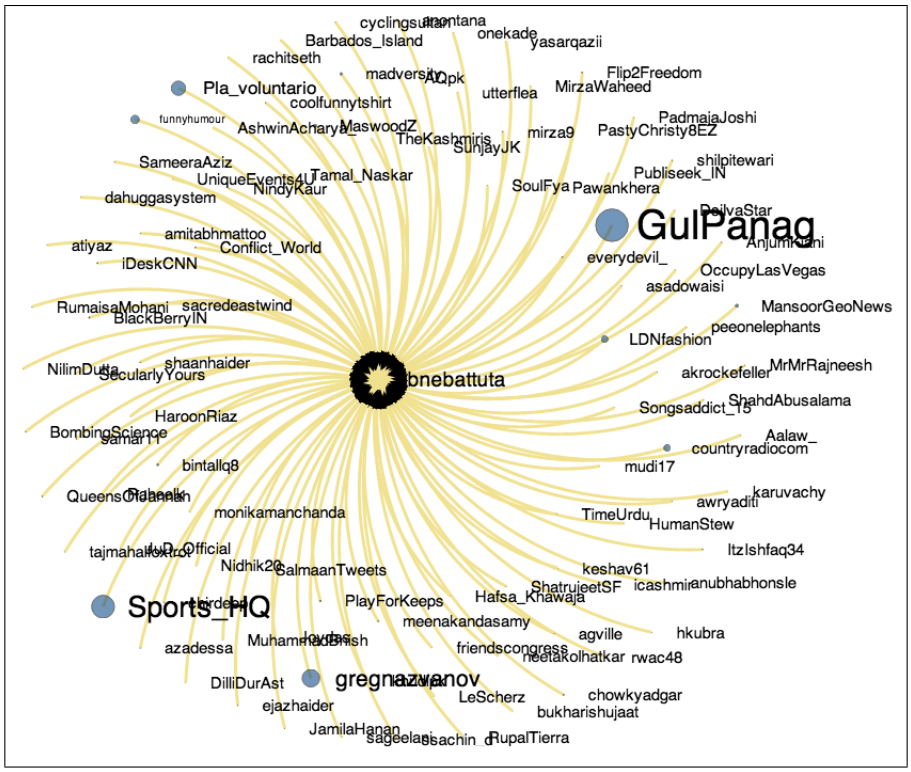



Figure 6.8: The network graph, with users as nodes and edges representing the retweets and replies by one user to another.

Figure 6.9: The tag-cloud of hashtags present in the tweets containing the query word.

## 6.4 Discussion

In this chapter we analyzed the role of Twitter during events in India. We identified and analyzed how information from Twitter can be used by emergency responders and authorities during a real-word event. The analysis of Twitter data for the Mumbai bomb blasts was the first attempt to do an India-centric study of the response and user activity on Twitter in an emergency situation. Understanding the dynamics and behaviors of online social media during a crisis situation, would help in developing technology to monitor and manage social media on the Internet, and prevent any

damage being inflicted by them in such situations. Knowledge about the impressions and opinions of public helps in making policy-decisions. We performed content based analysis on the tweets collected during and after the Mumbai blasts. Analysis of the vocabulary of tweets day-wise gave us an insights about what people talked about during the three days and how the topics evolved. We observed that the number of tweets generated by users with a large number of followers (who are generally like government owned accounts, celebrities, media companies) were very few. Majority of content generated at the time of crisis was from unknown users. It was also observed that, though the number of posts were less by users with large number of followers, these posts registered high numbers of retweets. We also extracted some false news and rumors that were spread during these blasts on Twitter. With the lack of any source on Twitter to verify these news as true or untrue, many of them were tweeted and retweeted hundreds of times. The situation could be improved if there were more credible sources like government authorities posting updates about the blasts.

We performed similar characterization analysis for several other events in India. We developed a real-time tool system Twit-Digest to be used by authorities, emergency responders and citizens to collect, analyze and visualize Twitter data. The system Twit-Digest is installed and being used till date by some government agencies in India. We performed multiple iterations of user feedback given by the agencies which helped improve our system further.

# Chapter 7

# Identifying and Characterizing User Communities

This chapter is joint work with Prof. Anupam Joshi. An earlier version of the content in this chapter was published in Proceedings of workshop on Data-driven user behavioral modelling and mining from social media, in CIKM companion ACM, 2012 [1]. Work done in this chapter featured in an article in an online blog by Dr. Patrick Meier. [1]

In next section we present the theory and methodology followed in this chapter. Section 7.3 gives the data description considered for the analysis. Section 7.4 describes the evaluation and results. The last section presents the discussion and conclusion of this chapter.

## 7.1 Identifying Communities of Users on Twitter during Crisis Events

Users who post messages about events, have varied opinions, sentiments, and information content. The biggest challenge in analyzing data from any social media is the volume of content being generated. The motivation for our work is that, high volume of content and users on such social networks, makes manual monitoring of each message or user impossible. We propose a framework to cluster together users who are similar and then extract views of prominent people in each of these communities formed by identifying the top central users in each of them.[2] Community detection helps us in monitoring the broad sub-topics and sub-communities of users that are formed during an event. We propose that identifying top central users is useful as they can represent the entire

[1] http://irevolution.net/2013/01/21/sna-for-digital-response/

[2] The words cluster, community and groups are used interchangeably in this paper.

community's views, hence we need to monitor and analyze the opinions and sentiments of only these few user profiles rather than all the users in a community. To identify the communities of users, and top users in each of these community, we utilized techniques of clustering and social network analysis.

As case studies for the generic framework proposed in this paper, we analyzed three major crisis events of 2011: the hurricane Irene, the riots in United Kingdom and the earthquake in Virginia. Hurricane Irene alone, caused a damage of about 7-10 billion+ USD and claimed 56 lives [23]. All these three real-life events, saw a vast number of posts on online social media during the event. For instance, we collected over 1.1 million tweets related to England riots alone. The methodology and framework evaluated in the work is generic and may be applied to any event.

The aim is to identify different user communities, and characterize them by the properties of top central users in them. Communities in a network are groups of nodes such that, there are maximum number of edges inside the clusters than between clusters [88]. The main contributions of this paper are:

- We defined a novel similarity metric to compute similarities between users based on their links, content posted and meta data.
- We applied spectral clustering to obtain communities of users formed during three different crisis events.
- We showed most central people (by degree centrality) represent what users in the cluster are talking about with 81% accuracy on average.

## 7.2 Theory and Methodology

Most of the previous work have focussed on detecting and analyzing communities based on explicit social links like friends, followers or retweets [100, 102]. We followed the approach proposed by Java et al. for clustering blog data by using spectral methods [53]. Java et al. adapted the N-Cut spectral clustering algorithm for image segmentation to online social media [88]. Adapting the approach proposed by them for blogs to Twitter, we formulated the communities based on different metrics like content, link and meta-data of user on Twitter. Figure 9.2 represents the architecture of the methodology followed in this paper. After we extracted the communities of users, we applied centrality metrics of social network analysis in order to obtain top central users in each community. The identification of top users in each community would help us in characterizing and analyzing, the topics and opinions of each community.

We performed clustering of users based upon the tweets posted, the link structure and the meta-data of the users. A *user*user* similarity matrix $U$ is constructed, where for all users $u_i$ and $u_j$,

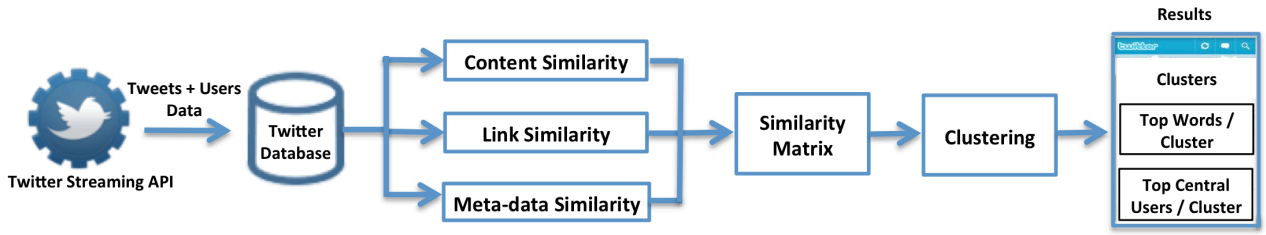



Figure 7.1: The architecture diagram for the methodology followed.

the value $a_{i,j}$ is the sum of the following:

- **Content similarity** $C(i,j)$: We calculated the similarity of content between two users by computing number of common words, hashtags and URLs in all the tweets by $u_i$ and $u_j$ related to the particular event.[3]
- **Link similarity** $L(i,j)$: The similarity between two users, $u_i$ and $u_j$ with respect to links is computed based on how many times two users retweet, mention or reply to each other tweets or a common third person's tweet.
- **Meta-data similarity** $M(i,j)$: Twitter profile of a user contains an optional field called location. We computed similarity between two users based on the value in their location field. The field is a text box, the user may leave it blank, or fill it with a valid / invalid location. We used Yahoo PlaceFinder API[4] to check whether the given text corresponds to a valid location. From the values returned by the API, we checked if the users have similarity at the level of country, state or city.

For each of the similarity metrics above, we normalized the score of each similarity using the maximum value of similarity score for each of the feature and then computed their sum (Equation 7.1).

$$U(i,j) = C(i,j) + L(i,j) + M(i,j) \tag{7.1}$$

After data collection from Twitter API, and construction of $U(i,j)$, the next step was to perform spectral clustering on this matrix. We performed the spectral clustering using the *specc* function of $R$ statistical analysis package [56], by specifying the number of clusters as three. [5] The users were divided in different communities, which is followed by identifying top users in each of the community. We used the degree centrality metric of social network analysis to identify the top

[3]Before finding the common words among the tweets, we removed all the stop words from the tweet text.
[4]http://developer.yahoo.com/geo/placefinder/
[5]The number of clusters was selected as three intuitively, for our preliminary study. In future, we would like to apply standard techniques for selection of number of clusters.

users for the communities. We then evaluated that with how much accuracy, the top central users represent the opinions and sentiments of the entire cluster.

## 7.3 Data Description

Three events analyzed in this paper are England riots, hurricane Irene and earthquake in Virginia. All events occurred during the months of July and August, 2011. Riots in United Kingdom caused 5 deaths, in addition to 16 civilian and 186 police injuries. We selected tweets about the UK riots, based on the keywords related to them that emerged as trending topics on Twitter from 6th August to 11th August [47]. Our second event under consideration is the earthquake of magnitude 5.8 that hit the Piedmont region of the U.S. state of Virginia. The third event is the hurricane Irene, which caused 55 deaths and a damage of US $10.1 billion. We used the Streaming API from Twitter to collect the tweets and user information corresponding to these three events. Table 7.1 presents the tweets and users of the three events analyzed in this chapter.

Table 7.1: Data statistics for the three crisis events.

| **Event** | **Tweets** | **Users** |
|---|---|---|
| England Riots | 1,165,628 | 546,966 |
| Hurricane Irene | 90,237 | 55,718 |
| Earthquake in Virginia | 277,604 | 219,621 |

## 7.4 Results and Evaluation

We performed two kinds of evaluation for the analysis. First, we measured the quality of communities formed using the modularity metric. Second, we evaluated if the top users identified using degree centrality metric represent the opinions and topic of the entire community.

### 7.4.1 Clustering Evaluation

For evaluating the quality of clusters obtained by spectral clustering, we computed the modularity score for each of the three events. Modularity score was proposed by Newman et al. as an index to measure how good a division of the network is in communities [68]. This score represents the fraction of edges that lie inside the communities minus the factor of expected value if the edges were placed at random, while maintaining the degree of nodes. Newman et al. also showed that

value of modularity above 0.3 implies a good division of graph into communities.[6] We obtained the modularity score greater than 0.5 for all events (Hurricane Irene = 0.67, England Riots = 0.51 and Virginia Earthquake = 0.53), hence concluded that the division of graph into communities is of good quality.

### 7.4.2 Top Users Evaluation

Next, we evaluated if the top central nodes in a community based on the similarity metrics represent the opinion and topics of the entire community. We aim to show that the top users based on similarity represent what the entire community shares, and hence to analyze and characterize the network, we need to monitor and investigate these few users rather than all the users in the community. We used Mean Average Precision (MAP) evaluation metric to test the same. MAP for a set of queries is the mean of the average precision scores for each query. We extracted the top $N$ words, according to the TF-IDF [7] score, for all users in each of the community and then only by the top central users in each of the community. Hence, for computing MAP scores, the top $N$ words by all users in the community form the ground truth and we compute the MAP for each community for the top $N$ words by the central users. Figure 7.2 shows the results obtained, by MAP, we obtained 81% accuracy (on average; max: 96%) for all three events. We concluded that the top central people represent what the entire community is saying with a high accuracy of 81%, which implies monitoring and analyzing these users can provide us with the opinions and topic of the entire community.

Next, we checked if the above result holds true only in case of tweets corresponding to events, or, in general with any set of tweets. To do this, we selected a random sample of messages from Twitter corresponding to the same time frame as the three events and having number of tweets equal to the mean average of the number of tweets in the three events. Figure 7.2 showed that the performance of MAP drastically reduces for the random sample of tweets, we get a MAP score of 32% on average. The conclusion that top users based on degree centrality represent the opinions and topic of the community holds true in case of events, since the tweets related to an event contain higher similarity in various properties like content of tweets and location of users.

### 7.4.3 Community Characterization

We characterized the communities obtained for Hurricane Irene in depth to understand how the approach evaluated in this research can be useful in real-life. Table 7.2 shows the top twenty words

[6]The maximum value of modularity for a network is 1 and minimum is 0, a high value of modularity score indicates, good quality division of the network, i.e. more number of edges in the network are within the clusters than between clusters.

[7]TF-IDF stands for Term Frequency-Inverse Document Frequency. It is a statistical measure which used to evaluate how important a word is to a document in a collection of documents.

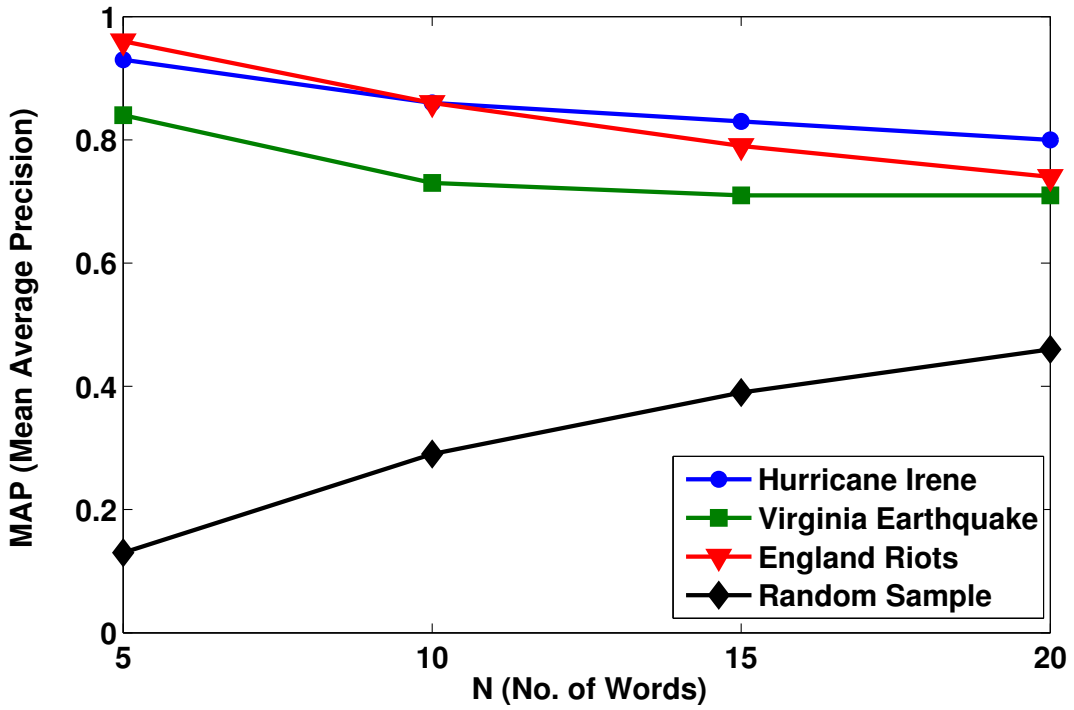



Figure 7.2: The graph shows the Mean Average Precision score for the top central users' detection for the three crisis events. We observe that while for the three events, MAP values are high; the corresponding values for a random sample are much lower. The graph is truncated to show the main effect. The values plateau after the points shown in the graph.

in the three clusters corresponding to the communities for all three events. We observed that while community 2 and 3 contain words related to information about the hurricane in different forms, community 1 consists of tweets by spammers on Twitter. It is a common practice by spammers to use trending hashtags to promote spam content, hence community 1 contains the words that were trending with respect to the hurricane, and rest of the words refer to spam content (marked in red) like *photos and videos of Angelina Jolie*. Using the technique of spectral clustering followed by top users detection using degree centrality measures, we clustered together information sharing users and separate them from the spammers who exploit the trending topic terms to spread their content. By looking at the top words in each community, community 2 and 3 look very similar, hence we extracted the corresponding top central users in cluster 2 and 3. Table 7.3 gives the details of the top users in community 2 and 3. By analyzing the screen name, user description and tweets by the top central users, we found that: community 2 is formed mostly of official news media, emergency and alert profiles (for e.g. *ShareWith911, metofficestorm, RES911CU, Nednewsnow*), while community 3 is formed with common or general users of Twitter (for e.g. *KerrieRamagano3, YaekoAvilez3837, AureliaHickman1*), none of whom had even provided any

Table 7.2: Top 20 words from each community (cluster) for the three crisis events.

| Events | Community 1 | Community 2 | Community 3 |
|---|---|---|---|
| **Hurricane Irene** | hurricane, irene, http, jada, song, brenda, butistillloveu, photos, icanhonestlysay, check, neverapologizefor, video, scandal, angelina, jolie, storm, nick, college, bahamas, puerto | irene, hurricane, http, coast, category, bahamas, east, puerto, storm, rico, winds, advisory, issued, forecast, heads, track, path, update, moving, number | hurricane, irene, http, puerto, coast, storm, rico, news, track, winds, bahamas, category, east, forecast, advisory, florida, update, issued, tropical, latest |
| **England Riots** | londonriots, http, ukriots, police, london, riots, people, hackney, rioters, news, fire, dont, riot, cameron, riotcleanup, birmingham, looting, croydon, night, stop | londonriots, http, ukriots, police, london, riots, people, rioters, cameron, hackney, news, fire, dont, riot, looting, prayforlondon, birmingham, manchesterriots, stop, riotcleanup | londonriots, http, ukriots, police, riots, london, people, rioters, hackney, cameron, riot, birmingham, news, fire, riotcleanup, dont, looting, make, manchesterriots, stop, rioting |
| **Virginia Earthquake** | earthquake, http, today, quake, east, coast, earth, richmond, washington, august, virginia, news, campbell, glen, monument, heard, norwegian, alaska, new york, video | earthquake, http, east, coast, virginia, washington, felt, magnitude, news, monument, feel, damage, today, nuclear, quake, epicenter, people, school, video | earthquake, http, coast, east, washington, felt, virginia, news, damage, today, magnitude, monument, feel, quake, people, video, area, breaking, shit, nuclear |

description on their Twitter profiles. Thus, even though community 2 and 3 had almost similar top words, our characterization showed that tweets in community 2 are formed from more news media and official user profiles than community 3 and thus can be easily verified and trusted. Similar to hurricane Irene analysis, when we analyzed communities for Virginia earthquake in Table 7.2, we observed while the first cluster contains more words related to the location of the earthquake (*richmond, virginia, new york, alaska, norwegian*), community 2 and 3 emphasize on the damage and people affected (*damage, people, feel, school, nuclear*).

Table 7.3: Screen name and user description of top users in Community 2 and 3.

| **C2:Username** | **C2:Description** | **C3:Username** | **C3:Description** |
|---|---|---|---|
| top_tw_science | none | YaekoAvilez3837 | none |
| top_trend_world | none | MarxInbody5668 | none |
| Neednewsnow | none | KerrieRamagano3 | none |
| iBreakings | none | AureliaHickman1 | none |
| metofficestorm | Official Met Office page for latest information on tropical storms, hurricanes, typhoons and cyclones | LenitaBross5906 | none |
| RES911CU | RES911CU Bringing you the latest news/weather from around the world & country using over 1000 news sources | BeckieSteakley4 | none |
| qianafgtt | none | KirbyPietig6221 | none |
| Georgeannla47 | none | GailSterner5774 | none |
| moneyworlds | none | LeroyRasheed | none |
| LinJosh2011 | none | ClorindaKan1926 | none |
| top_trend_uk | none | StaceeRosebure2 | none |
| ShareWith911 | Emergency Information Sharing Network connecting 911, First Responders and Citizens before, during and after emergencies of all shapes and sizes. | VaniaAsmus | none |

## 7.5 Discussion

In this work, we identified and evaluated communities of users on Twitter during three different crisis events. We applied spectral clustering technique to cluster users based on a similarity metric, taking into account the content, link and meta-data similarities of the users. We then used the degree centrality measure from social network analysis to extract top central users from each of the communities formed. We showed that the top central users represent the topics and opinions of the entire community with an average 81% accuracy. The results obtained with the case study of

three events show the potential of applying spectral clustering and centrality measures, to identify community of users and the prominent top users in each community, during crisis events.

# Chapter 8

# Credibility Ranking of Tweets during Real-world Events

An earlier version of the content in this chapter was published in Workshop on Privacy and Security in Online Social Media, proceedings of the 21st International Conference on World Wide Web companion (WWW), ACM 2012 [2].

The rest of this chapter is organized as follows: Next section discusses the motivation of using supervised learning techniques for credibility evaluation on data generated during real-world events. Section 8.3 explains the methodology that we used in data collection, events selection, and coding scheme used for annotating the tweets. Section 8.4 describes the approach of supervised machine learning and relevance feedback to rank tweets. Section 8.5 presents the experimental results for the framework proposed on the data for real-world events. Section 8.6 summarizes the results from our analysis and highlights the implications of our results.

## 8.1 Using Supervised Learning Methods to Assess Credibility of Content

Though a large volume of content is posted on Twitter, not all information is trustworthy or useful in providing information about the event. The credibility and quality of information often plays a critical role during high impact real-world events, as fake news and rumors also propagate along with genuine news [63]. Researchers have shown that role of Twitter during mass convergence and emergency events differs considerably than more general Twitter activity [51]. They showed that tweets during such events lead to more information broadcasting and brokerage. This was the motivation for us to specifically consider some major events during 2011 for our analysis. Each of

the events that we analyzed had thousands of tweets (minimum number of tweets for each event was 25,000 and the maximum number of tweets for one event was 542,685 posted about them from all around the globe.

Figure 8.1 gives some sample tweets from our dataset for the event Hurricane Irene. All three tweets contain the words matching to the event and were posted while Hurricane Irene was the trending. [1] The tweet in 8.1(a) provides correct and credible information about the event. The tweet in 8.1(b) is related, but contains no information about the event, it expresses personal opinion of the user. Even though the tweet in 8.1(c) contains related words, it includes a URL to an advertisement to sell a product, so it is a spam tweet with respect to the event. In this work, information refers to the situational awareness information, that is information that leads to gain in the knowledge or update about details of the event, like the location, people affected, causes, etc. [70].

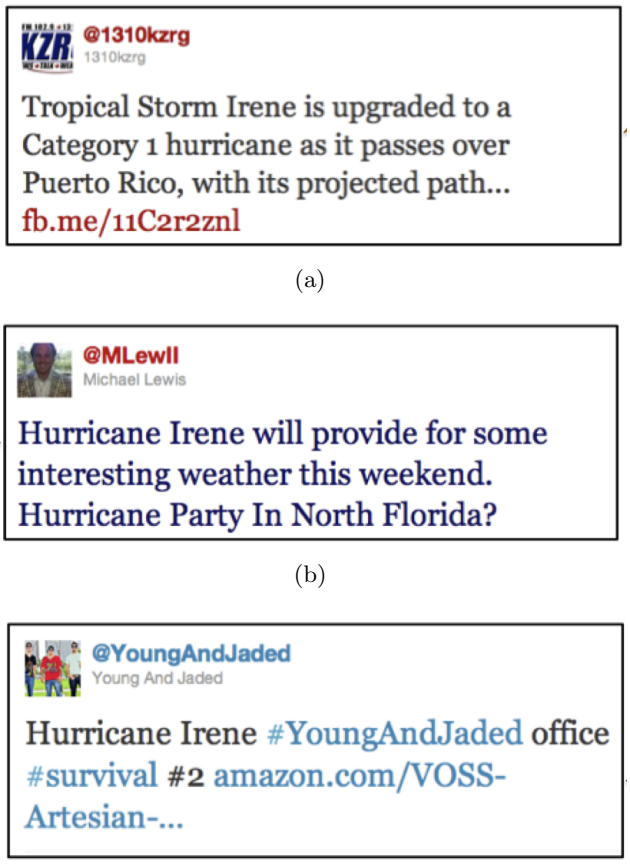



Figure 8.1: Sample tweets in our dataset for the event Hurricane Irene. (a) The tweet provides correct and credible information about the event. (b) The tweet is related to the event, but contains no information about the event. (c) Tweet contains related words, it includes a URL to an advertisement to sell a product.

[1] Trending topics on Twitter are the current most talked about words or phrases on Twitter.

We envision, understanding the credible (incredible) information on Twitter to be useful for devising strategies to mitigate the widespread of incredible information (like fake news or rumors). To the best of our knowledge, this was the first work to use automated ranking techniques to assess credibility at the most atomic level of information on Twitter, i.e. at the tweet level.

The main contributions of this chapter are:

- We found that on average, 30% content about an event, provides situational awareness information about the event, while 14% was spam. In all, 17% of the content was found to be credible information by users.
- We performed linear logistic regression analysis on various Twitter based (content and user) features. Prominent content based features were, number of unique characters, swear words, pronouns and emoticons in a tweet, and user based features like the number of followers and length of username.
- We showed that automated algorithms using supervised machine learning and relevance feedback approach based on Twitter features can be effectively used in assessing credibility of tweets.

## 8.2 Background on Relevance Ranking for Tweets

Ranking techniques have been used widely to rank URLs, content and users on various Web 2.0 platforms. Page et al. developed a PageRank algorithm for webpages on the Internet, they used the number of out-links and in-links of a webpage to calculate its relative relevance to a query [71]. Duan et al. in their paper proposed a supervised learning approach for ranking tweets based on certain query inputs [28]. They used content and non-content features (like authority of users) to rank tweets according to their relevance to a topic. Their work used Rank-SVM technique and extracted the best features, that resulted in good ranking performance. The three prominent features were: whether a tweet contains URL, the length of tweet (number of characters), and authority of user account. Chen et al. built a tool called *zerozero88*, [2] which recommends URLs that a particular Twitter user might find interesting [19]. They showed, how topic relevance and social voting parameters help in effective recommendations. Dong et al. worked on using inputs from Twitter to improve recency and relevance ranking for search engines using Gradient Boosted Decision Tree (GBDT) algorithm [27]. They showed how in addition to existing features used to rank URLs on web, information from Twitter can be used to enhance the ranking of URLs on the Web.

[2] http://zerozero88.com/

So far, the work done to assess credibility on Twitter, have explored credibility with respect to trending topics and users. Our work differs from that done by Castillo et. al [16] – their analysis was based on credibility of a trending topic (all tweets belonging to a topic were marked as credible or incredible) on Twitter, while we focus on assessing credibility at the level of tweets. This difference in approaches lends a significant impact in case of Twitter, since a topic (e.g. earthquake at a particular location) maybe credible, yet the tweets in that topic maybe of credible or incredible (e.g. Richter scale of the earthquake) in nature. Hence, credibility of a topic may not be a good indicator to judge the credibility of the content of the tweet. In this chapter, we used automated ranking techniques to assess credibility at the most atomic level of information on Twitter, i.e. at a tweet level. Using supervised machine learning and relevance feedback approach, we showed that ranking of tweets based on Twitter features (topic and source) can aid in assessing credibility of information in messages posted about an event. We believe, our results can help users in making a decision on the credibility of the tweet.

## 8.3 Methodology

In this section, we discuss the data collection setup, the process of selecting the events, and the coding scheme used to annotate the tweets. Figure 8.2 describes the methodology and analysis performed in the research presented in this research work.

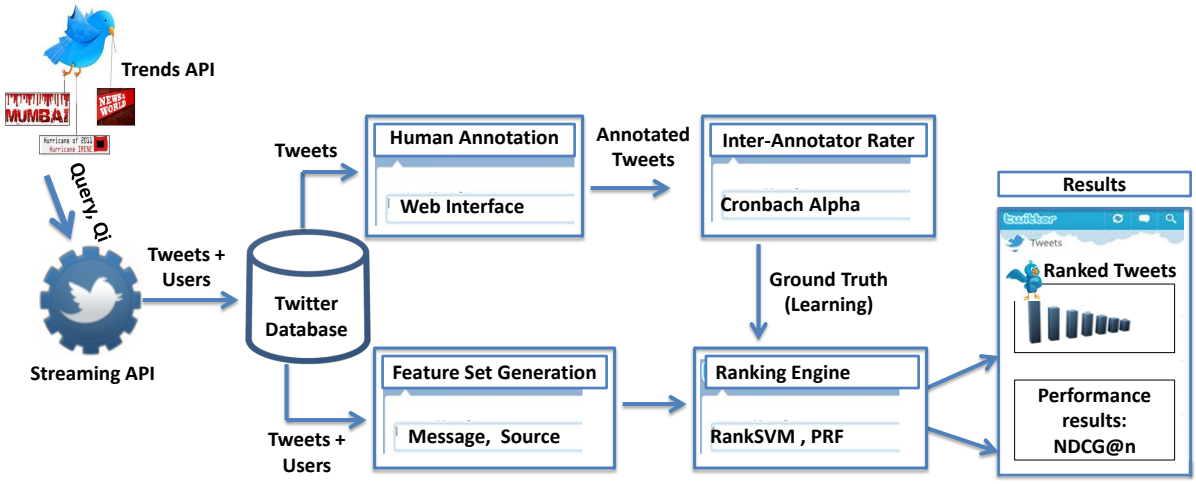



Figure 8.2: Methodology and analysis performed in this chapter.

### 8.3.1 Data Collection

We collected data from Twitter using the *Streaming API*. [3] This API enables researchers to extract tweets in real-time, based on certain query parameters like words in the tweet, time of posting of tweet, etc. To obtain query terms, we used, *Trends API* from Twitter, which returns top 10 trending topics on Twitter. [4] We queried *Trends API* after every 3 hours for the current trending topics, and collected tweets corresponding to these topics as query search words for the *Streaming API*. We collected tweets corresponding to a topic until the time it remained as a trending topic. Castillo et al. also used a similar framework to collect tweets using current trending topics [16]. In our data collection, we considered both worldwide and local trending topics from Twitter, for e.g. United States of America, Canada, India, Brazil, United Kingdom, etc. We collected data of over 35 million tweets by more than 6 million users in the time period $12^{th}$ July, 2011 to $30^{th}$ August, 2011. Table 8.1 gives the descriptive statistics of the data collected.

Table 8.1: Descriptive statistics of the Twitter dataset.

| | |
|---|---|
| **Total tweets** | 35,748,136 |
| **Total unique users** | 6,877,320 |
| **Tweets with URLs** | 4,973,457 |
| **Number of singleton tweets** | 22,481,898 |
| **Number of re-tweets / replies** | 13,266,238 |
| **Trending Topics (unique)** | 3,586 |
| **Start date** | $12^{th}$ July, 2011 |
| **End date** | $30^{th}$ August, 2011 |

### 8.3.2 Events Selection

Using the methodology described in Section 8.3.1, in total 3,586 unique trending topics were obtained. We shortlisted 14 major events that occurred all around the globe between July $12^{th}$ and August $30^{th}$, 2011. Each event had one or more trending topics associated with it, for example trending topics related to the debt and downgrading crisis in the USA were *AAA to AA, S&P,* etc. For each event, we considered tweets containing the words in trending topics to be the set of tweets for that event. Table 8.2 describes the fourteen events that we selected, the number of tweets for each event, the corresponding trending topics for the event, and a short description of the event. We selected events covering various domains of news events like political, financial, natural disasters, terror strikes and entertainment news. To ensure that we select events with high impact and relevance, we applied following minimum criterion for selecting an event for analysis:

[3] https://dev.twitter.com/docs/streaming-api.
[4] https://dev.twitter.com/docs/api/1/get/trends

- An event which had minimum of 25,000 tweets were considered. For example, *Riots in UK* in August, 2011, our system collected 542,685 tweets for the event.
- Topics corresponding to the event which were trending for minimum 24 hours as a country or worldwide trending topic on Twitter. For example, *Hurricane Irene* was a trending topic on Twitter for 150 hours.

### 8.3.3 Annotation Scheme

This section describes how we annotated the set of tweets for each event. We took help from human annotators to obtain the ground truth regarding the presence of credible information in tweets related to a news event. Human annotation for understanding the ground truth is a well-established research methodology [16]. For the fourteen selected events, we picked a random sample of 500 tweets per topic. We restricted our annotation to tweets in English language; we selected tweets by those users who had set English as their language on Twitter. Though, there were some tweets by users which were in languages other than English, we provided the annotators with a *Skip tweet* option to avoid such non-English tweets. For the purpose of annotation, we developed a web interface and we provided each annotator with an authenticating login and password.

To assess the presence of credible information, if any, we asked the human annotators to select one of the following options for each tweet:

- Tweet contains information about the event. Rate the credibility of information present:
  - Definitely Credible
  - Seems Credible
  - Definitely Incredible
  - I can't Decide
- Tweet is related to the news event, but contains no information
- Tweet is not related to news event
- Skip tweet

We provided the annotators with the definition of credibility. Oxford dictionary defines the term credibility as "the quality of being trusted and believed in". [5] In the context of this research, we aim to assess the credibility of the information in the content of a tweet (message) by a user on Twitter.

[5] `http://www.oxforddictionaries.com/definition/english/credibility`

Table 8.2: Fourteen events selected for the time period from 12 July to 30 August, 2011.

| **Events** | **Tweets** | **Trending Topics** | **Description** |
|---|---|---|---|
| UK Riots | 542,685 | #ukriots, #londonriots, #prayforlondon | Riots in United Kingdom caused 5 deaths, 16 civilian and 186 police injuries |
| Libya Crisis | 389,506 | libya, tripoli | Rebels opposing Col. Qaddafi seized Zawiyah |
| Earthquake in Virginia | 277,604 | #earthquake, Earthquake in SF | Earthquake of magnitude 5.8 hit the Piedmont region of the U.S. state of Virginia. |
| US Downgrading | 148,047 | S&P, AAA to AA | Debt crisis in the US, led Standards & Poor to downgrade it from AAA to AA-plus |
| Hurricane Irene | 90,237 | Hurricane Irene, Tropical Storm Irene | Hurricane Irene in US caused 55 deaths and a damage of US $10.1 billion |
| Indiana State Fair Tragedy | 49,924 | Indiana State Fair | Five people died and 40 were injured in a stage accident at the Indiana State Fair. |
| Mumbai Blast, 2011 | 32,156 | #mumbaiblast, Dadar, #needhelp | Three bomb blasts in Mumbai (India) on 13th July, 26 people died and 130 injured |
| JanLokPal Bill Agitation | 182,692 | Anna Hazare, #janlokpal, #anna | An anti-corruption movement against the Government of India. |
| Apple CEO Steve Jobs resigns | 158,816 | Steve Jobs, Tim Cook, Apple CEO | Apple's stock dropped 7% when Steve Jobs resigned as its CEO |
| Google acquires Motorola Mobility | 68,527 | Google, Motorola Mobility | Google buying Motorola Mobility in a $12.5bn cash deal, was a huge acquisition |
| News of the World Scandal | 67,602 | Rupert Murdoch, #murdoch | The News International phone hacking scandal exposed Rupert Murdoch |
| Abercrombie & Fitch stocks drop | 54,763 | Abercrombie & Fitch, A&F | Abercrombie & Fitch stocks drops 9% after a controversy |
| Muppets Bert and Ernie were gay | 52,401 | Bert and Ernie | Rumors circulated that muppet pair Ernie and Bert, are a gay couple |
| New Facebook Messenger | 28,206 | Facebook Messenger | Facebook launched a new messenger for mobile users |

A tweet is said to contain credible information about a news event, if you trust or believe that information in the tweet to be correct / true. We then explained the above mentioned categories

using an illustrative example. For each of the events, we provided a 5-10 line description of the event along with two URL links to news articles on the event featured in premier news websites likes CNN, Guardian and BBC. During our pilot study, we observed options 'Seems Credible' and 'Seems Incredible' were redundant, as both indicated that a tweet seemed both credible and incredible to the user. Hence, for the final annotation, we kept only one of the options. Each tweet for the events (500 tweets per event) was annotated by three different annotators. Figure 8.3 shows the screenshot of annotation procedure.

Figure 8.3: Screenshot of annotation portal. It shows the options presented to the annotators per tweet.

To check the reliability of results obtained via annotation, we computed the Cronbach Alpha score. The overall Cronbach alpha value for inter-annotator agreement for all 7,000 (14 events * 500 tweets) tweets was 0.748. Alpha $> 0.7$ implies a high agreement between annotators [60]. We selected the majority score for a tweet (i.e. value given by at least 2 annotators) as the final scores for each tweet; we discarded all tweets for which all three annotators gave different scores. After removing tweets that had all three annotators giving different ranking score and tweets which annotators decided to skip, in total we obtained 5,578 (around 80% from 7,000) tweets in our final annotated dataset.

## 8.4 Analysis

We proposed an automated ranking scheme to output of tweets ordered according to the credibility of information provided in them. We used a combination of supervised machine learning and relevance feedback approach to rank tweets. We analyzed the effectiveness of Twitter based features (message and source level) to rank tweets according to information quality in the tweet. As a next step, we evaluated an enhancement to the above ranking technique by using pseudo feedback relevance re-ranking scheme. We used SVM ranking algorithm to build a model for credibility of information in tweets. Ranking SVM algorithm is an extension of SVM classifier traditionally used for the classification task [55]. We used the SVM$^{Rank}$ implementation code from Cornell

University. [6] SVM$^{Rank}$ trains a Ranking SVM on the training set, and outputs the learned rule to a model file. Based on the learned model, the algorithm predicts a ranking score that are written to the output file. We performed four-fold cross validation of our results. The ground truth for the task was obtained from the human annotated tweet scores.

### 8.4.1 Types of Features

Two basic characteristics, the features of the message itself, and the properties of the user who posted the message characterize any post or tweet on Twitter. Table 8.3 presents features that are available in the message and the user. We considered following two types of features as input to the ranking algorithm:

- **Content or message level features**: 140 characters posted by users contain data (e.g. words, URLs, hashtags) and meta-data (e.g. is tweet a reply or a retweet) related to it. We do not consider text semantic features here in our analysis.
- **Source or user level features**: The attributes of the user who posted the tweet. We consider properties such as number of friends, followers and status messages of the user as part of this set.

Table 8.3: Content and source based features that were considered for ranking.

| **Message based features** |
|---|
| Length of the tweet, number of words, number of unique characters, number of hashtags, number of retweets, number of swear language words, number of positive sentiment words, number of negative sentiment words, tweet is a retweet, number of special symbols [$, !], number of emoticons [:-), :-(], tweet is a reply, Number of @-mentions, number of retweets, time lapse since the query, has URL, number of URLs, use of URL shortener service |
| **Source based features** |
| Registration age of the user, number of statuses, number of followers, number of friends, is a verified account, length of description, length of screen name, has URL, ratio of followers to followees |

### 8.4.2 Pseudo Relevance Feedback

Pseudo Relevance Feedback (PRF) also known as Blind Relevance Feedback, is a prominent re-ranking technique used in information retrieval tasks to improve the performance of ranking re-

[6] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

sults [12]. The basic idea of PRF is to extract $K$ ranked documents and then re-rank those documents according to a defined score. In our algorithm, we extracted most frequent unigrams from the top $K$ tweets and used the text similarity between the most frequent unigrams and $K$ tweets to re-rank them. The improvement achieved by re-ranking using PRF is highly dependent on the quality of top $K$ results given by the ranking algorithm. We applied PRF to the best set of results obtained by previous analysis, that is the ranking results obtained using both message and source. We calculated the text similarity using the metric BM25 [81], between a tweet $T$ and the query set Q (formed with the most frequent unigrams extracted from top $K$ tweets) for each event. Each word in query set Q was represented by $q_i$. The BM25 metric is given by:

$$BM25(T,Q) = \Sigma_{q_i \in Q} \frac{IDF(q_i).tf(q_i,T).(k_1+1)}{tf(q_i,T) + k_1(1 - b + b\frac{Length(T)}{avg_{length}})} \tag{8.1}$$

where $tf(q_i)$ is the frequency of occurrence of word $q_i$ in Tweet T, $Length(T)$ denotes the length of $T$ and $avg_{length}$ represents average length of tweet in corpus. The variables $k_1$ and $b$ are constants; we take their standard values as $k_1$=1.2 and $b$=0.75 in our case [90]. The value of $IDF(q_i)$, Inverse Document Frequency for a query term $q_i$, is calculated as follows:

$$IDF(q_i) = log\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \tag{8.2}$$

where, $n(q_i)$ represents the number of documents (tweets) containing $q_i$, and $N$ is the total number of documents.

The algorithm (**Algorithm 1**) describes all the above steps of extracting *top k* ranked tweets. Function *ExtractFeatures*(T) computes message and source based features for each tweet $t_i$ from the set of tweets T. The $RankSVM(F,T)$ function, takes the feature set matrix $F$ and the column vector $A$ containing the ground truth annotation value for each of the $n$ tweets. The *SortAsc* and *SortDsc* functions sort the tweets according to their given score value in ascending and descending value respectively. $FreqLUnigrams(\mathrm{T}_K)$ extracts the frequent $L$ word unigrams from the top $K$ tweets. $BM25$ method computes the similarity score between the top $L$ unigrams and each tweet $t_i$ in T.

### 8.4.3 Evaluation Metric

For evaluating the relevance ranking results, we used the standard metric of NDCG (Normalized Discounted Cumulative Gain) [52]. We chose this measure over other information retrieval measures like MAP (Mean Average Precision), as NDCG captures data with multiple grades. Given a rank-ordered vector V of results $< v_1, \ldots, v_m >$ to query q, let label($v_i$) be the judgment of $v_i$

**Algorithm 2** Ranking (T[1..$n$], A[1..$n$])

1: **for** $i$ <- 0 to $n-1$ **do**
2: F$_i$ <- $ExtractFeatures$(T[$i$])
3: **end for**
4: FeatureRank <- $RankSVM(F, A)$
5: T$'$ <- $SortAsc$ (FeatureRank)
6: **for** i ¡- 0 to $k-1$ **do**
7: T$_K$[i] <- T$'$[i]
8: **end for**
9: W$_L$ = $FreqLUnigrams$(T$_K$)
10: PRFRank <- $BM25$ (T$_K$, W$_L$)
11: TweetRank <- $SortDsc$ (PRFRank)
12: return TweetRank[1..$k$]

(4=Credible, 3=Maybe credible, 2= Incredible, 1=Relevant but no information, 0=Spam). The discounted cumulative gain of V at document cut-off value $n$ is:

$$DCG@n = \Sigma_{i=1}^{n} \frac{1}{log_2(1+i)}(2^{label(v_i)} - 1) \tag{8.3}$$

The normalized DCG of V is the DCG of V divided by the DCG of the "ideal" (DCG-maximizing) permutation of V (or 1 if the ideal DCG is 0). The NDCG of the test set is the mean of the NDCG's of the queries in the test set.

## 8.5 Experimental Results

For the human annotated data, on an average, 50% of tweets on an event are composed of tweets which were related to the event but provided no useful information about it. Such tweets generally express the personal opinion or reactions of Twitter users on the event. We also found 13.5% spam tweets in our dataset about the events, i.e. the tweets contained the words belonging to the trending topics but were not related to the event. We found that 30% of tweets contained information about the event, but only 17% of the tweets had information that was credible.

### 8.5.1 Regression Analysis

We performed logistic linear regression analysis, with respect to the features listed in Table 3 to estimate the good predictors for credibility of tweets. To perform the regression analysis, we considered all tweets annotated as definitely and seems credible as the data points for the positive class (dependent variable = 1), and rest all as not credible class of tweets (dependent variable value

= 0). As a result, we obtained the following features as strong indicators (p-value $< 0.001$) of credibility: number of characters present in tweet, number of unique characters present in tweet, presence of swear words, inclusion of pronouns and presence of sad / happy emoticons.

The number of unique characters present in tweet was positively correlated to credibility, this may be due to the fact that tweets with hashtags, @mentions and URLs contain more unique characters. Such tweets are also more informative and linked, and hence credible. Presence of swear words in tweets indicates that it contains the opinion / reaction of the user and would have less chances of providing information about the event. Tweets that contain information or are reporting facts about the event, are impersonal in nature, as a result we get a negative correlation of presence of pronouns in credible tweets. Low number of happy emoticons [:-), :)] and high number of sad emoticons [:-(, :(] act as strong predictors of credibility. Some of the other important features (p-value $< 0.01$) were inclusion of a URL in the tweet, number of followers of the user who tweeted and presence of negative emotion words. Inclusion of URL in a tweet showed a strong positive correlation with credibility, as most URLs refer to pictures, videos, resources related to the event or news articles about the event.

By performing regression analysis for credibility, we were able to identify some prominent features based on the content and user of the tweet. We mostly got all tweet (content) based features to play an important role in determining credibility, but user based features like the number of followers also come out as a strong predictor. Hence, we concluded that a ranking algorithm based on both the content of the tweets, and the user properties, would be effective in determining the credibility of information in the tweets.

### 8.5.2 Evaluation of Ranking

In this section, we evaluate the performance of Rank-SVM and PRF using the NDCG evaluation metric. We measured the ranking output's performance, for the top 25 ranked tweets in each of the analysis presented. For baseline evaluation, we considered time recency (most recent tweet on top), that is the order in which tweets are displayed by Twitter search. On an average for top 25 tweets, we achieved 0.37 NDCG for the recency metric.

#### Content and Source based Analysis

We performed ranking individually on content based features (tweet) and source level features (user). We observed that both set of features perform comparatively. Using the combined set of features (tweet and user) there was a significant statistical improvement observed in the performance of ranking (Paired T-test, t=7.47, p-value $< 0.05$) than the performance of individual feature sets. We conclude that both message and source based features play a role in predicting the credibility

based rank of the tweets. Figure 8.4 (a) shows the cumulative gain using the two feature sets over time recency. These results highlights the point that content based features are as important as the source based features on Twitter with respect to credibility. Hence, it is not only important *who you are* when you tweet, but also the quality of *what you post*.

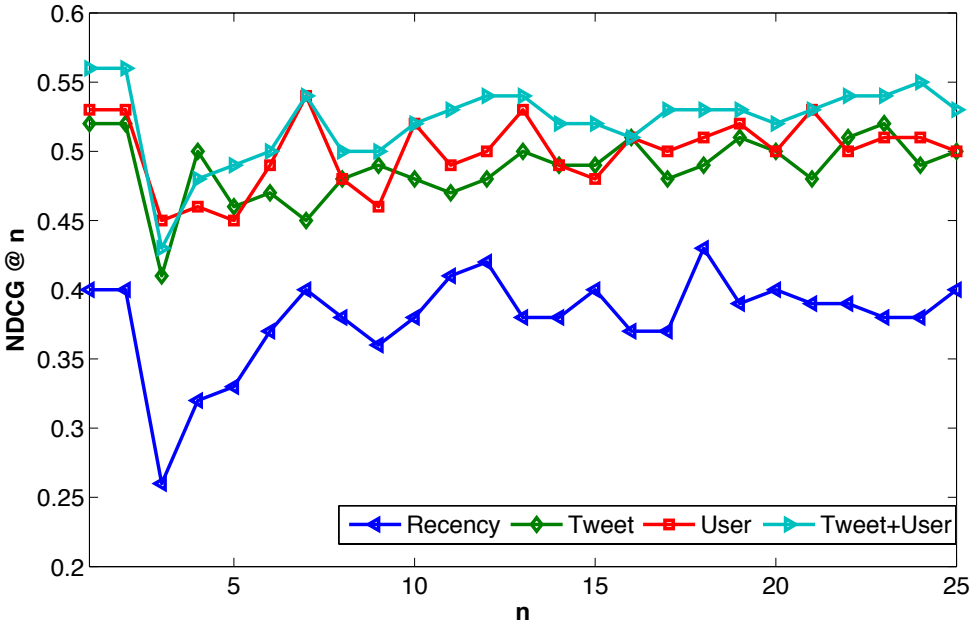
**Relevance Feedback Re-ranking Analysis**

For PRF analysis, we took the top 50 tweets obtained from Rank-SVM (tweet and user features), extracted the ten most frequent unigrams (after removing stop words, user-ids and URLs) from the tweets. We re-ranked the tweets, based on the similarity score based on BM25 metric, for the tweet and the top 10 unigrams. A similarity score based on this metric was computed for all frequent unigrams and the tweet. The top tweets were then re-ranked in the descending order of their similarity score. Figure 8.4 (b) shows that the performance of credibility is enhanced considerably using PRF (average 0.73 NDCG score). Using the context (e.g. frequent n-grams) in Twitter for ranking can be useful in increasing the effectiveness of credibility ranking.

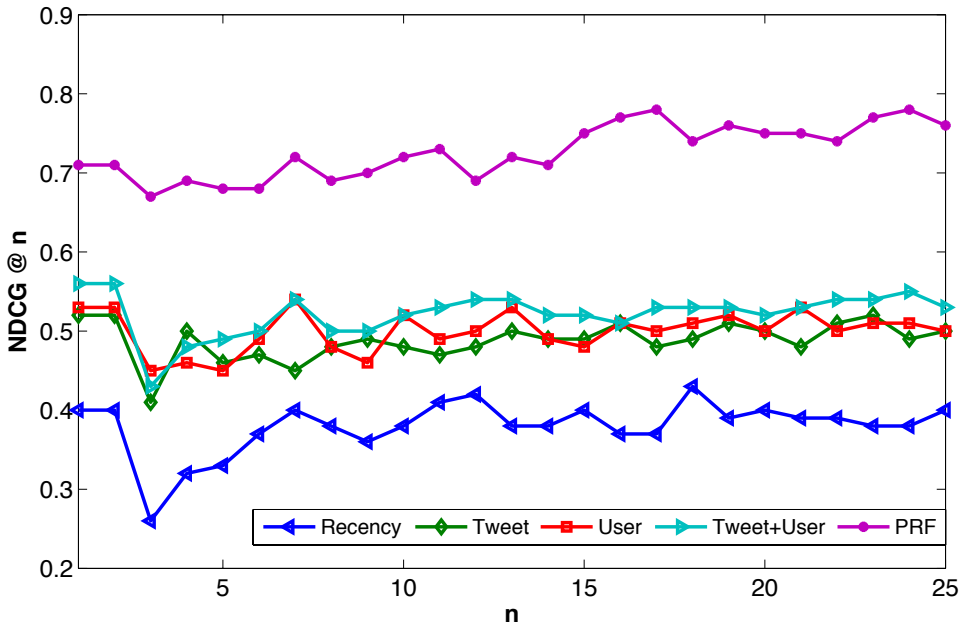
## 8.6 Discussion

During high impact events, there is a sudden rise in activity about the events over the Internet. People log on to social media websites to check for updates about the event and also to share information about the event with others. We considered, Twitter as our medium of information for this work. In recent years, Twitter has emerged as a news and information sharing platform during such events. Though a large volume of content is posted on Twitter, not all of the information is trustworthy or useful in providing information about the event.

Credibility of information on Twitter is a big challenge in its utilization as news and information sharing platform. In particular, credibility of information during high impact events can be important. Researchers have shown that role of Twitter during mass convergence and emergency events differs considerably than regular Twitter activity. In this research work, we considered fourteen high-impact events from all around the globe and analyzed the tweets for these events for the credibility of information in them. By information here, we mean situation awareness information about an event. Situational awareness information is information that leads to gain in the knowledge or update about details of the event, like the location, people affected, causes, etc. We found that on average, 30% content about an event, provides situational awareness information about the event, while 14% was spam. Annotators found that only 17% of the total tweets contained situational awareness information that was credible. We applied linear logistic regression analysis to identify the prominent Twitter features (content and user based) which can help in assessing the credibility. Prominent content based features were number of unique characters, swear words,




(a)




(b)

Figure 8.4: Performance Evaluation of ranking algorithm. (a) Recency results v/s Twitter features (tweet and user based); (b) Improved performance using PRF technique.

pronouns and emoticons in a tweet; and user based features like the number of followers and length of username. We evaluated an algorithm (using RankSVM and relevance feedback approach) to rank the tweets, according to the credibility of information contained in the tweet. We observed that the content based features were as important as the source based features on Twitter with respect to credibility. The results assert the fact that it is not only important *who you are* when you tweet, but also the quality of *what you post*. By applying the relevance feedback techniques based on most frequent n-unigrams, we were able to achieve considerably enhanced performance of ranking results. We show that both context independent features (Twitter based) and context specific features (unigrams) aid in the ranking mechanism. Results show that extraction of credible information from Twitter can be automated the with high confidence.

# Chapter 9

# TweetCred: A tool to Measure Credibility on Twitter in Real-time

> This chapter is a joint work with Dr. Carlos Castillo and Dr. Patrick Meier of Qatar Computing Research Institute (QCRI). An earlier version of the content in this chapter was published in Proceedings of the 6th International Conference on Social Informatics, (SocInfo14), 2014 [4]. The work done in this chapter featured in many news websites and blogs including the Washington Post, [1] the New Yorker [2] and the Daily Dot [3] among others.

This chapter is organized as follows: Section 9.1 discusses the motivation for building a real-time model for credibility assessment. Section 9.2 describes how we collect labeled data to train our system, and Section 9.3 discusses how we apply a learning-to-rank framework for automatically ranking tweets by credibility. Section 9.4 presents the implementation details and a performance evaluation, and Section 9.5 discusses the evaluation from users and their feedback. Finally, in the last section we discuss the results and their implications.

## 9.1 Building Real-time Credibility Model for Twitter

During sudden onset crisis events, the presence of spam, rumors and fake content on Twitter reduces the value of information contained on its messages (or "tweets"). A possible solution to this problem is to use machine learning to automatically evaluate the credibility of a tweet, i.e.

[1] http://wapo.st/1pWE0Wd

[2] http://newyorker.com/online/blogs/elements/2014/05/can-tweetcred-solve-twitters-credibility-problem.html

[3] http://www.dailydot.com/technology/tweetcred-chrome-extension-addon-plugin/

whether a person would deem the tweet believable or trustworthy. This has been often framed and studied as a supervised classification problem in an off-line (post-hoc) setting.

In this chapter, we present a supervised ranking model for scoring tweets according to their credibility. We introduce TweetCred (available at `http://twitdigest.iiitd.edu.in/TweetCred/`), a novel, practical solution based on ranking techniques to assess *credibility* of content posted on Twitter in real-time. We understand credibility as "the quality of being trusted and believed in", [4] following the definition in the Oxford English Dictionary. A tweet is said to be credible, if a user would trust or believe that the information contained on it is true.

In contrast with previous papers based on off-line classification of content in a post-hoc setting (e.g. [2, 63] and many others), TweetCred uses only the data available on each message, without assuming extensive historical or complete data for a user or an event. Also in contrast with previous work, we evaluate TweetCred with more than 1,808 users who downloaded a browser extension that enhanced their Twitter timeline, as shown in Figure 9.1.

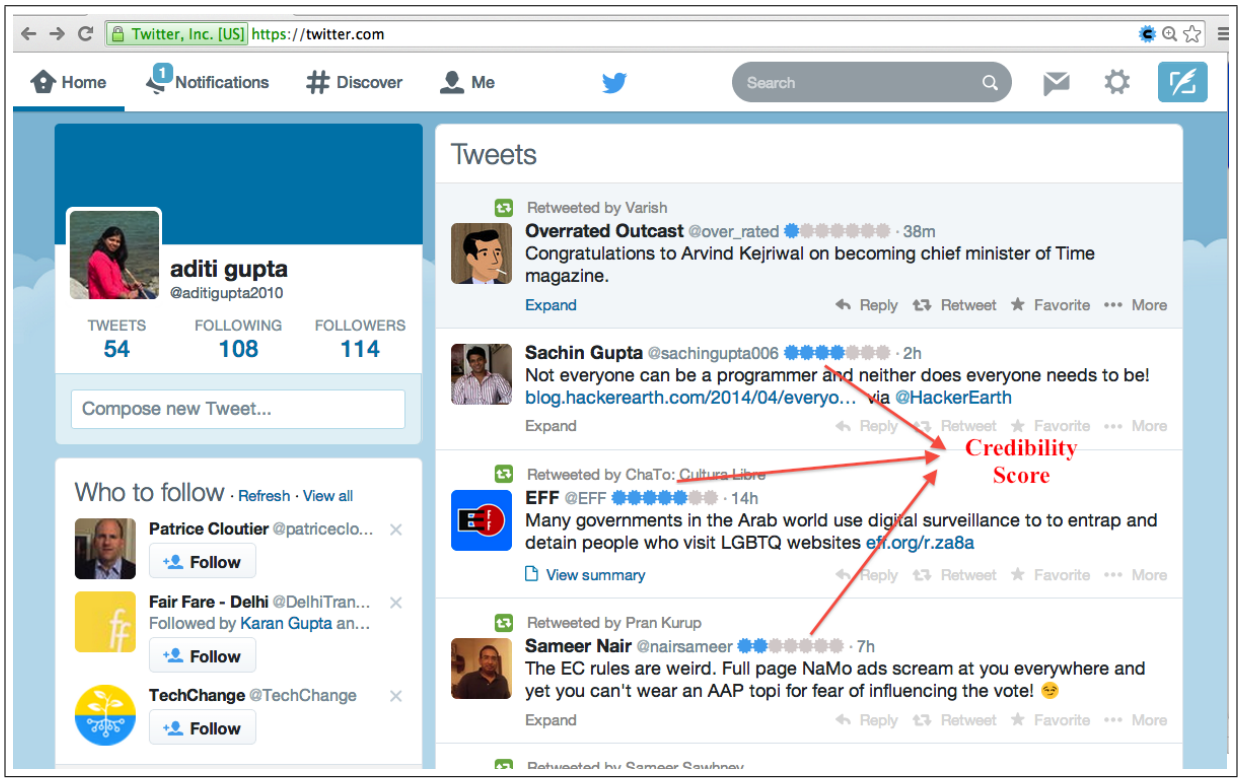



Figure 9.1: Screenshot of timeline of a Twitter user when TweetCred browser extension is installed.

The main contributions of this work are:

- We presented a supervised ranking model using SVM-rank for assessing credibility based,

[4] `http://www.oxforddictionaries.com/definition/english/credibility`

on training data obtained from 6 high impact crisis events of 2013. An extensive set of 45 features was used to determine the credibility score for each tweet.

- We developed and deployed a real time system, TweetCred, in the form of a browser extension, web application, and REST API. The TweetCred extension was installed and used by 1,808 Twitter users, there were 14.2 million requests for computing credibility score of tweets.
- We evaluated the performance of TweetCred in terms of response time, effectiveness and usability. We observe that 80% of the credibility scores are computed and displayed within 6 seconds, and 63% of users either agreed with our automatically-generated scores or disagreed by 1 or 2 points (on a scale from 1 to 7).

To the best our knowledge, the work presented in this chapter is the first research work that describes the creation and deployment of a practical system for credibility on Twitter, including the evaluation of such system with real users. Figure 9.2 depicts the methodology followed in this chapter.

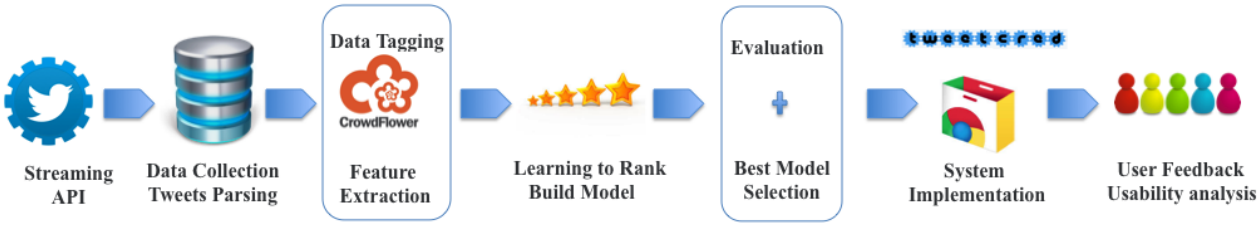



Figure 9.2: Architecture diagram for this research work.

## 9.2 Training Data Collection

TweetCred is based on supervised learning. As such, it requires as input a *training set* of tweets for which a credibility label is known.

To create this training set, we collected data from Twitter using Twitter's streaming API, filtering the results using keywords representing six prominent events in 2013:

- Boston Marathon blasts in the US
- Typhoon Haiyan/Yolanda in the Philippines
- Cyclone Phailin in India
- Shootings in the Washington Navy Yard in the US

- Polar vortex cold wave in North America
- Tornado season in Oklahoma, US

These events affected a large population and generated a high volume of content in Twitter. Table 9.1 describes the characteristics of the data collected around the events we used to build a training set.

Table 9.1: Number of tweets and distinct Twitter users from which data was collected for the purposes of creating a training set. From each event, 500 tweets were labeled.

| Event | Tweets | Users |
|---|---|---|
| Boston Marathon Blasts | 7,888,374 | 3,677,531 |
| Typhoon Haiyan / Yolanda | 671,918 | 368,269 |
| Cyclone Phailin | 76,136 | 34,776 |
| Washington Navy yard shootings | 484,609 | 257,682 |
| Polar vortex cold wave | 143,959 | 116,141 |
| Oklahoma Tornadoes | 809,154 | 542,049 |
| Total | 10,074,150 | 4,996,448 |

In order to create ground truth for building our model for credibility assessment, we obtained labels for around 500 tweets selected uniformly at random from each event. The annotations were obtained through crowdsourcing provider CrowdFlower. [5] We selected only annotators living in the United States, since our focus was to annotate English language tweets. For each tweet, we collected labels from three different annotators, keeping the majority among the options chosen by them.

The annotation proceeded in two steps. In the first step, we asked users if the tweet contained information about the event to which it corresponded, with the following options:

- R1: The tweet contains information about the event.
- R2: The tweet is related to the event, but contains no information.
- R3: The tweet is not related to the event.
- R4: None of the above (skip tweet).

Along with the tweets for each event, we provided a brief description of the event and links from where users could read more about it. In this first step, 45% of the tweets were considered informative (class R1), while 40% were found to be related to the event for which they were extracted, but not informative (class R2), and 15% were considered as unrelated to it (class R3).

[5] `http://www.crowdflower.com/`

In the second step, we selected 45% of tweets that were marked as informative, and annotated them with respect to the credibility of the information conveyed by it. We provided a definition of credibility as said before: "the quality of being trusted and believed in", and example tweets for each option in the annotation. We asked workers to score each tweet according to its credibility with the following options:

- C1: Definitely credible
- C2: Seems credible
- C3: Definitely incredible
- C4: None of the above (skip tweet)

Figure 9.3 shows the snapshot of the annotation task. Among the informative tweets, 52% of tweets were labeled as *definitively credible*, 35% as *seems credible*, and 13% as *definitively incredible*.

Table 9.2 gives the distribution of the annotations. There were about 23% of tweets that contained definitely credible information about an event; and about 6% information that the users definitely did not trust. The annotation results are quite similar to those obtained from our previous work for 2011 events [2].

Table 9.2: Distribution of labels over tweets.

| Label | Percentage | |
|---|---|---|
| | 2013 events | [2] |
| Definitively credible | 23% | 17% |
| Seems credible | 16% | } 13% |
| Definitively incredible | 6% | |
| Not informative | 40% | 56% |
| Not related to the event | 15% | 14% |

## 9.3 Credibility Modeling

Our aim was to develop a model for ranking tweets by credibility. We adopted a supervised learning-to-rank approach. First, we performed feature extraction from the tweets. Second, we compared the speed and accuracy of different machine learning schemes, using the training labels obtained in the previous section.

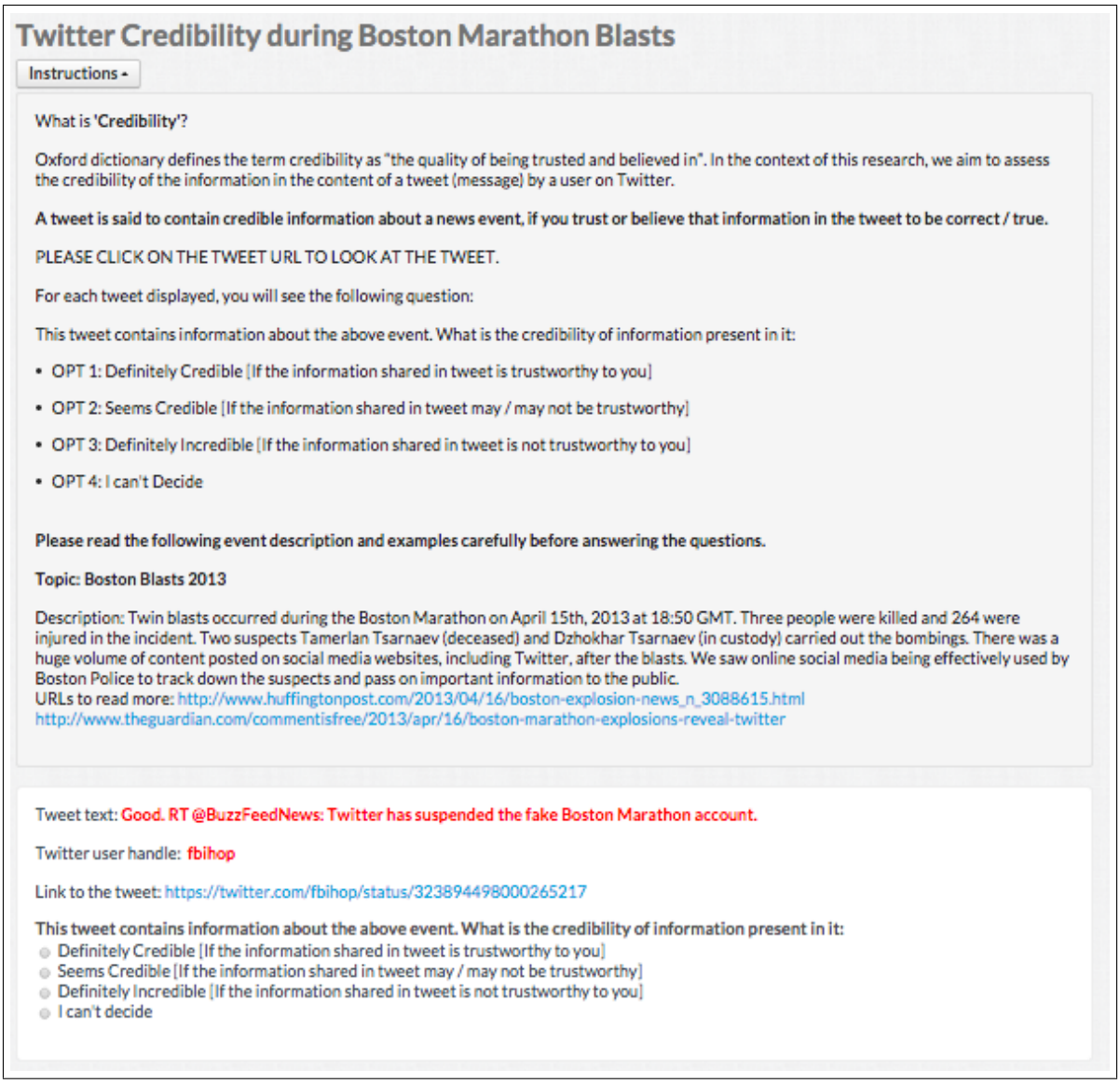



Figure 9.3: Screenshot of the annotation task using CrowdFlower.

### 9.3.1 Feature Extraction

Generating feature vectors from the tweets is a key step that impacts the accuracy of any statistical model built from that data. We used a collection of features from previous work [2, 4, 16, 109], restricting ourselves to those that can be derived from single tweets in real-time.

A tweet as downloaded from Twitter's API contains a series of fields in addition to the text of the message. [6] For instance, it includes meta-data such as posting date, and information about its

[6] `https://dev.twitter.com/docs/api/1.1/get/search/tweets`

Table 9.3: Features used by the credibility model.

| Feature set | Features |
|---|---|
| Tweet meta-data | Number of seconds since the tweet; Source of tweet (mobile / web/ etc); Tweet contains geo-coordinates |
| Tweet content (simple) | Number of characters; Number of words; Number of URLs; Number of hashtags; Number of unique characters; Presence of stock symbol; Presence of happy smiley; Presence of sad smiley; Tweet contains 'via'; Presence of colon symbol |
| Tweet content (linguistic) | Presence of swear words; Presence of negative emotion words; Presence of positive emotion words; Presence of pronouns; Mention of self words in tweet (I; my; mine) |
| Tweet author | Number of followers; friends; time since the user if on Twitter; etc. |
| Tweet network | Number of retweets; Number of mentions; Tweet is a reply; Tweet is a retweet |
| Tweet links | WOT score for the URL; Ratio of likes / dislikes for a YouTube video |

author at the time of posting (e.g. his/her number of followers). For tweets containing URLs, we enriched this data with information from the Web of Trust (WOT) reputation score. [7] The features we used can be divided into several groups, as shown in Table 9.3. In total, we used 45 features.

### 9.3.2 Learning Scheme

We tested and evaluated multiple learning-to-rank algorithms to rank tweets by credibility. We experimented with various methods that are typically used for information retrieval tasks: Coordinate Ascent [64], AdaRank [105], RankBoost [36] and SVM-rank [55]. We used two popular toolkits for ranking, RankLib [8] and SVM-rank. [9]

Coordinate Ascent is a standard technique for multi-variate optimization, which considers one dimension at a time. SVM-rank is a pair-wise ranking technique that uses SVM (Support Vector Machines). It changes the input data, provided as a ranked list, into a set of ordered pairs, the (binary) class label for every pair is the order in which the elements of the pair should be ranked. AdaRank trains the model by minimizing a loss function directly defined on the performance measures. It applies a boosting technique in ranking methods. RankBoost is a boosting algorithm based on the AdaRank algorithm; it also runs for many iterations or rounds and uses boosting

[7] The WOT reputation system computes website reputations using ratings received from users and information from third-party sources. `https://www.mywot.com/`

[8] `http://sourceforge.net/p/lemur/wiki/RankLib/`

[9] `http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html`

techniques to combine weak rankings.

### 9.3.3 Evaluation metric

Two important factors for our system are correctness and response time, hence, we compared the methods based on two evaluation metrics, NDCG (Normalized Discounted Cumulative Gain) and running time. NDCG is useful to evaluate data having multiple grades, as is the case in our setting. Given a query $q$ and its rank-ordered vector V of results $\langle v_1, \ldots, v_m \rangle$, let label($v_i$) be the judgment of $v_i$. The discounted cumulative gain of V at document cut-off value $n$ is:

$$DCG@n = \Sigma_{i=1}^{n} \frac{1}{log_2(1+i)} (2^{label(v_i)} - 1) \ .$$

The normalized DCG of V is the DCG of V divided by the DCG of the "ideal" (DCG-maximizing) permutation of V (or 1 if the ideal DCG is 0). The NDCG of the test set is the mean of the NDCGs of the queries in the test set.

To map the training labels from Section 9.2 to numeric values, we used the following transformation: 5=Informative and definitively credible (class R1.C1), 4=Informative and seems credible (R1.C2), 3=Informative and definitively incredible (R1.C3), 2=Not informative (R2), 1=Not related (R3). From the perspective of quality of content in a tweet, a tweet that is not credible, but has some information about the event, is considered better than a non-informative tweet.

### 9.3.4 Evaluation

We evaluated the different ranking schemes using 4-fold cross validation on the training data. Table 9.4 shows the results. We observe that AdaRank and Coordinate Ascent perform best in terms of $NDCG@n$ among all the algorithms; SVM-rank is a close second in performance. We performed Kruskal-Wallis test to compute the statistical significance of NDCG values of various ranking methods. We observed a p-value of 0.5 indicating that there is no statistical difference in performance of different methods.

The gap is less as we go deeper into the result list, which is relevant given that Twitter's user interface allow users to do "infinite scrolling" on their timeline, looking at potentially hundreds of tweets.

The table also presents the learning (training) and ranking (testing) times for each of the methods. The ranking time of all methods was less than one second, but the learning time for SVM-rank was, as expected, much shorter than for any of the other methods. Given that in future versions of TweetCred we intend to re-train the system using feedback from users, and hence need short training times, we implemented our system using SVM-rank.

Table 9.4: Evaluating ranking algorithms in terms of Normalized Discounted Cumulative Gain (NDCG) and execution times. Boldface values in each row indicate best results.

| | AdaRank | Coord. Ascent | RankBoost | SVM-rank |
|---|---|---|---|---|
| NDCG@25 | **0.6773** | 0.5358 | 0.6736 | 0.3951 |
| NDCG@50 | **0.6861** | 0.5194 | 0.6825 | 0.4919 |
| NDCG@75 | 0.6949 | **0.7521** | 0.6890 | 0.6188 |
| NDCG@100 | 0.6669 | **0.7607** | 0.6826 | 0.7219 |
| Time (training) | 35-40 secs | 1 min | 35-40 secs | 9-10 secs |
| Time (testing) | <1 sec | <1 sec | <1 sec | <1 sec |

For the above ranking task, we have considered only data collected for the six events of 2013 for this research work. We then analyzed if we can consider the data annotated in our 2012 study for fourteen events [2]. For checking the same, we trained the ranking model using SVM-rank on 2011 events data and tested on 2013 events data. Table 9.5 shows the results of this experiment. We observe that for the given feature vectors, the SVM-rank gives good results when trained and tested on the same year dataset, when trained on 2011 and tested on 2013 dataset, we observe there is a drastic drop in the accuracy. This can be attributed to various factors like evolution of Twitter and its usage during large scale events over past few years.

Table 9.5: Performance of SVM-rank algorithm in credibility ranking of tweets using 2011 and 2013 data. We observe a significant drop in NDCG metric values when training on data from one year and testing on data from a different year.

| Training | NDCG @25 | NDCG @50 | NDCG @100 | Testing |
|---|---|---|---|---|
| 2011 Dataset | 0.4765 | 0.5966 | 0.7359 | 2011 Dataset |
| 2013 Dataset | 0.3951 | 0.4919 | 0.7219 | 2013 Dataset |
| 2011 Dataset | 0.3743 | 0.3693 | 0.3783 | 2013 Dataset |

Table 9.6 shows the top 10 features of the models for credibility ranking built for 2011 events [2] and 2013 events [this chapter]. For both sets, we observe that both tweet- (e.g. number of characters in a tweet, presence of URL in tweet) and user-based (e.g. ratio of friends / followers, user location) features are important. The fact that many of the top features are different for both set of events, explains why the 2011 data should not be used to predict real-time credibility in current scenario. It also highlights that there is temporal evolution in the landscape of credibility prediction models. Hence, for all our analysis and implementation described in next few sections we used only the data for 2013 events.

Table 9.6: Top 10 features obtained using SVM-rank for ranking tweets according to their credibility. We observe that many of the top features are different for both scenarios.

| 2013 Dataset | 2011Dataset |
|---|---|
| Tweet contains *via* | Presence of $ symbol |
| No. of characters in tweet | Tweet contains URL |
| Unique characters in tweet | User has location in profile |
| No. of words in tweet | User has URL in profile |
| User has location in profile | No. of characters in tweet |
| Number of retweets | No. of words in tweet |
| Age of tweet | Unique characters in tweet |
| Tweet contains URL | Friends / Followers |
| Statuses / Followers | Favorites / Statuses |
| Friends / Followers | User is verified |

## 9.4 Implementation and Performance Evaluation

In order to encourage many users to interact with TweetCred, we provided it in a way that was easy to use, a browser extension. We also provided access to TweetCred as a web-based application and as an API, but the browser extension was much more commonly used.

### 9.4.1 Implementation

The implementation includes a back-end and a front-end which interact over RESTful HTTP APIs.

### 9.4.2 Back-end

Figure 9.4 shows the basic architecture of the system. The flow of information in TweetCred is as follows: A user logs on to his/her Twitter account on `http://twitter.com/`, once the tweets starts loading on the webpage, the browser extension passes the IDs of tweets displayed on the page to our server on which the credibility score computation module is done.

We do not collect the tweet or user information from the web page, we pass only the tweet IDs to web server. The reason is that what the server needs to compute credibility is more than what is shown through Twitter's interface.

From the server a request is made to Twitter's API to fetch the data about an individual tweet. Once the complete data for the tweet is obtained, the feature vectors are generated for the tweet, and then the credibility score is computed using the prediction model of SVM-rank. This score is re-scaled to a value in the range from 1 to 7 using the distribution of values in our training data. Next, this score is sent back to the user's browser. Credibility scores are cached for 15 minutes, meaning that if a user requests the score of a tweet whose score was requested less than 15 minutes

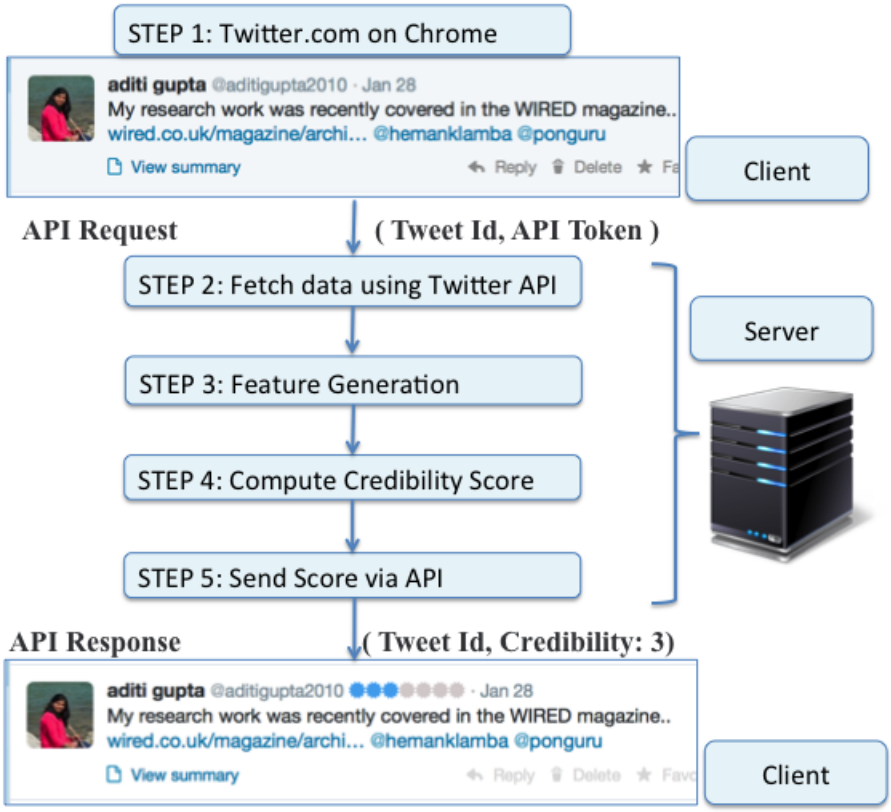



Figure 9.4: Data flow steps of the TweetCred extension and API.

ago by the system for any user, the previously-computed score is re-used. After this period of time, cached credibility scores are discarded and computed again if needed, to account for changes in tweet or user features such as the number of followers, retweets, favorites and replies.

All feature extraction and credibility computation scripts were written in *Python* with *MySQL* as a database back-end. The RESTful APIs were implemented using *PHP*. The hardware for the backend was a mid-range server (Intel Xeon E5-2640 2.50GHz, 8GB RDIMM).

Figure 9.5 shows the interface webpage to access the TweetCred API. [10] The input parameter is the 'tweet id' and the API returns the corresponding 'credibility score' for the tweet. The users require an API token to access this service, which we provide to them on request. We have received requests for TweetCred API tokens from more than 10 research / research groups across the globe.

[10]Link to TweetCred API: `http://twitdigest.iiitd.edu.in/twit-digest-plugin/api/index.php/`

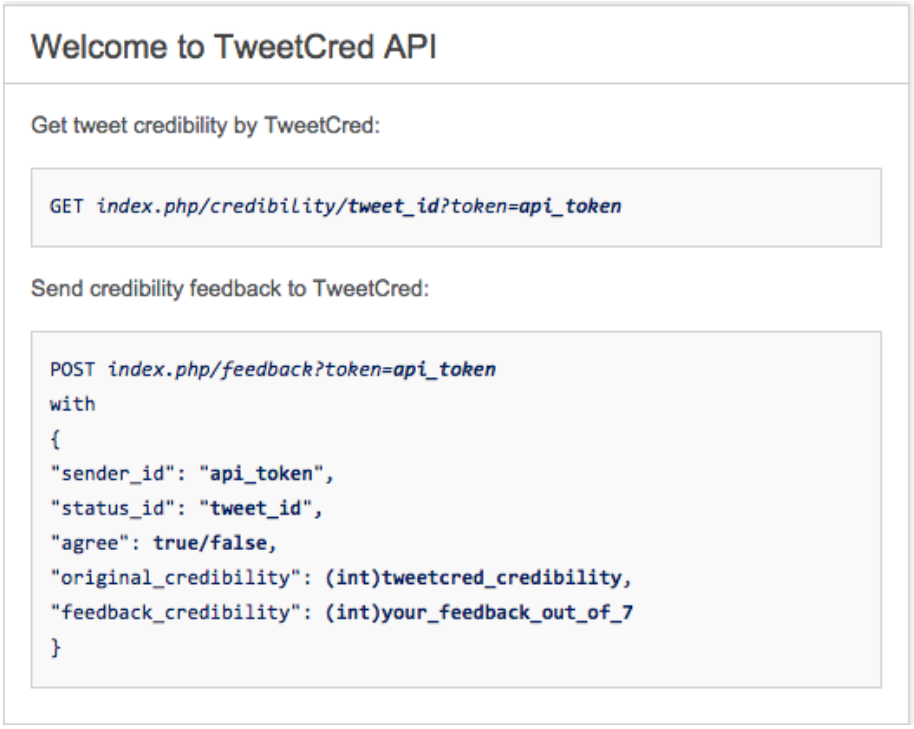



Figure 9.5: The interface webpage to access the TweetCred API.

### 9.4.3 Front-end

The Chrome browser currently enjoys the largest user base by far among various web browsers, [11] and hence was our target for the first version of the browser extension. In order to minimize computation load on the web browser, heavy computations were offloaded to the web server, hence the browser extension had a minimalistic memory and CPU footprint. This design ensures that the system would not result in any performance bottleneck on client's web browser.

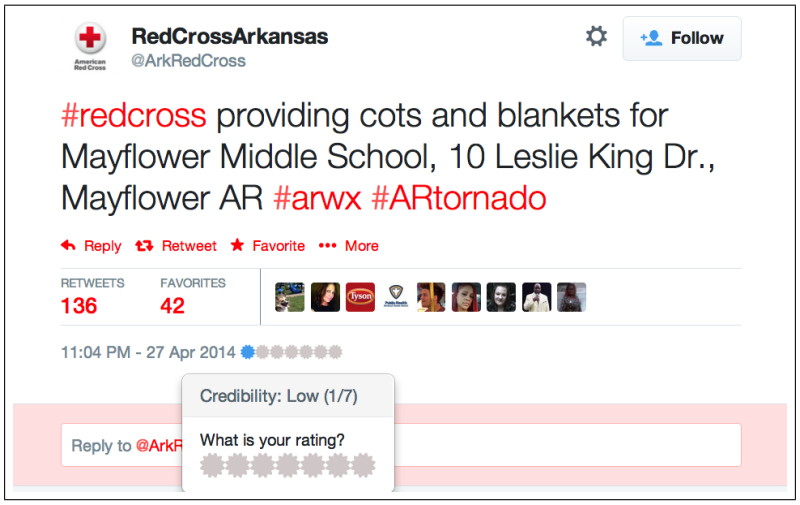
In an initial pilot study conducted for TweetCred with 10 computer science students that are avid Twitter users, we used the *Likert Scale* of score 1–5 for showing credibility for a tweet [96]. We collected their feedback on the credibility score displayed to them via personal interviews. The users found it difficult to differentiate between a high credibility score of 4 and a low credibility score of 2, as the difference in values seemed too small. Eight out of the ten participants felt that the scale of rankings should be slightly larger. They were more comfortable with a scale of 1–7 ranking, which we adopted.

TweetCred displays this score next to a tweet in a user's timeline, as shown in Figure 9.1. Additionally, the user interface includes a feedback mechanism. When end users are shown the credibility

[11] As of August 2014, Chrome has 59% of market share, more than doubling the 25% of the second place, Firefox `http://www.w3schools.com/browsers/browsers_stats.asp`

score for a tweet, they are given the option to provide feedback to the system, indicating if they agree or disagree with the credibility score for each tweet. Figures 9.6(a) shows the two options given to the user upon hovering over the displayed credibility score. In case the user disagrees with the credibility rating, s/he is asked to provide what s/he considers should be the credibility rating, as shown in Figure 9.6(b). The feedback provided by the user is sent over a separate REST API endpoint and recorded in our database.

(a)




(b)

Figure 9.6: Users can provide feedback to the system. Figure (a) shows how users can push the agree ("thumbs up") button to agree with a rating, the case for the disagree ("thumbs down") button is analogous. Figure (b) shows how users can provide their own credibility rating for a tweet.

### 9.4.4 Response Time

We analyzed the response time of the browser extension, measured as the elapsed time from the moment in which a request is sent to our system to the moment in which the resulting credibility score is returned by the server to the extension. Figure 9.7 shows the CDF of response times for 14.2 million API requests received. From the figure, we can observe that for 80% of the users the response time was less than 6 seconds, while for 98% of the users the response time was under 10 seconds. The response time is dominated by the requests done to Twitter's API to obtain the details for a tweet.

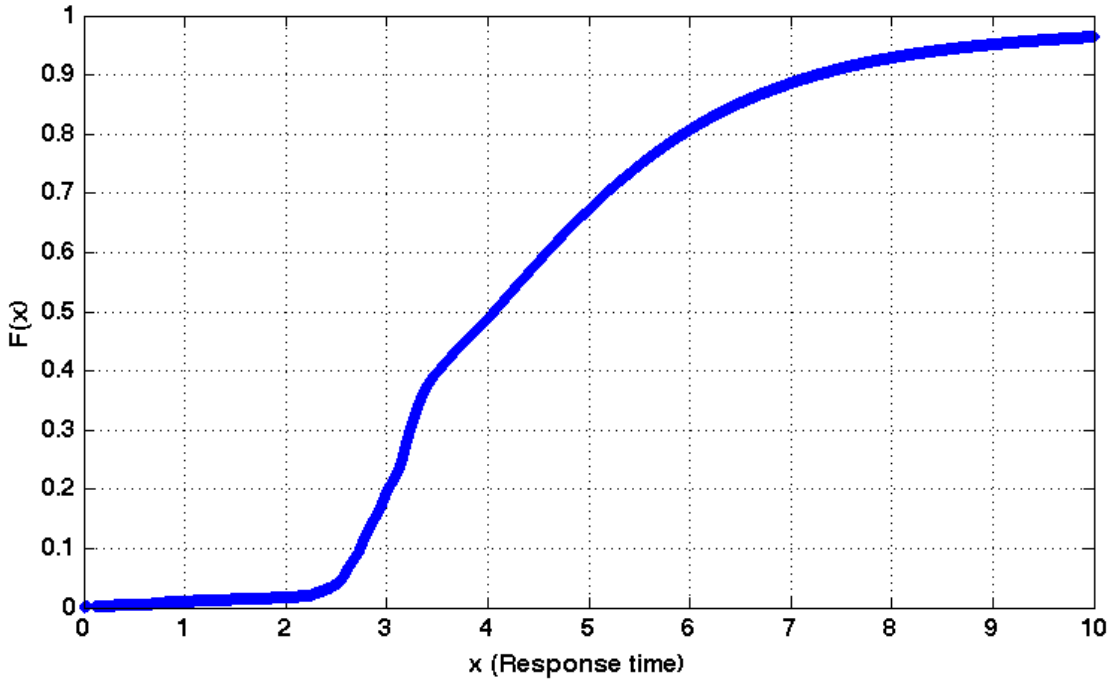



Figure 9.7: CDF of response time of TweetCred. For 80% of the users, response time was less than 6 seconds and for 98% of the users, the response time was under 10 seconds.

## 9.5 User Testing

We uploaded TweetCred to the Chrome Web Store, [12] and advertised its presence via social media and blogs. We analyzed the deployment and usage activity of TweetCred from April 27th, 2014 to March 28th, 2015. A total of 1,808 unique Twitter users used TweetCred. They constitute a diverse sample of Twitter users, from users having very few followers to two users having more

[12] `https://chrome.google.com/webstore/detail/tweetcred/fbokljinlogeihdnkikeeneiankdgikg`

Table 9.7: Summary statistics for the usage of TweetCred.

| Date of launch of TweetCred | 27 Apr, 2014 |
|---|---|
| Credibility score requests recieved | 14,234,131 |
| Unique Twitter users | 1,808 |
| Feedback was given for tweets | 1,654 |
| Unique users who gave feedback | 364 |
| Unique tweets which received feedback | 1,642 |

than 2 million followers. Figure 9.8 shows the distribution of number of followers and friends of TweetCred users.

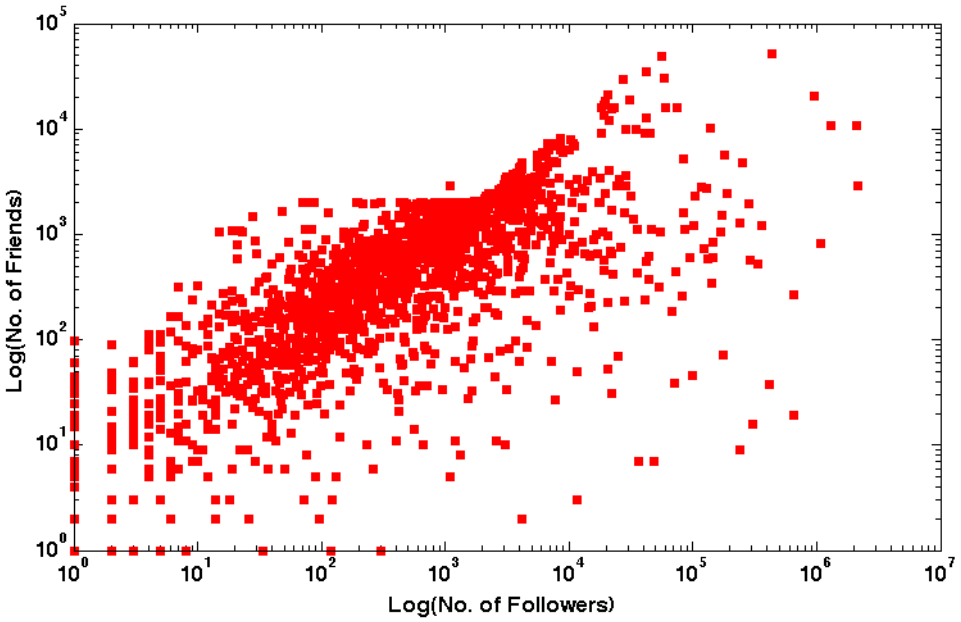



Figure 9.8: Distribution of number of followers and friends of TweetCred users.

Table 9.7 presents a summary of usage statistics for TweetCred. In total 14,234,131 API requests for the credibility score of a tweet were received as on March 28th, 2015.

We received feedback from users of our system in two ways. First, the users could give their feedback on each tweet for which a credibility score was computed. Secondly, we asked users to fill a usability survey on our website.

### 9.5.1 User Feedback

Out of the 14.2 million credibility score requests served by TweetCred, we received feedback for 1,654 of them. When providing feedback, users had the option of either agreeing or disagreeing with

Table 9.8: Feedback given by users of TweetCred on specific tweets.

| | Observed | 95% Conf. interval |
|---|---|---|
| Agreed with score | 40.14 | (36.73, 43.77) |
| Disagreed with score | 59.85 | (55.68, 64.26) |
| Disagreed: score should be higher | 48.62 | (44.86, 52.61) |
| Disagreed: score should be lower | 11.23 | (9.82, 13.65) |
| Disagreed by 1 point | 8.71 | (7.17, 10.50) |
| Disagreed by 2 points | 14.29 | (12.29, 16.53) |
| Disagreed by 3 points | 12.80 | (10.91, 14.92) |
| Disagreed by 4 points | 10.91 | (9.17, 12.89) |
| Disagreed by 5 points | 6.52 | (5.19, 8.08) |
| Disagreed by 6 points | 6.59 | (5.26, 8.16) |

our score. In case they disagreed, they were asked to mark the correct score according to them. Table 9.8 shows the break-down of the received feedback. We observed that for 40% of tweets for which user's provided feedback agreed with the credibility score given by TweetCred, while 60% disagreed—this can be partially explained by self-selection bias due to cognitive dissonance: users are moved to react when they see something that does not match their expectations.

### 9.5.2 Credibility Rating Bias

For the approximately 60% tweets the users disagreed with our score, and for 49% of the tweets the users felt that credibility score should have been higher than the one given by TweetCred, while for approximately 11% thought it should have been lower. This means TweetCred tends to produce credibility scores that are lower than what users expect. This may be in part due to the mapping from training data labels to numeric values, in which tweets that were labeled as "not relevant" or "not related" to a crisis situation were assigned lower scores. To test this hypothesis, we use keyword matches to sub-sample, from the tweets for which a credibility score was requested by users, three datasets corresponding to crisis events that occurred during the deployment of TweetCred: the crisis in Ukraine (3,637 tweets), the Oklahoma/Arkansas tornadoes (1,362 tweets), and an earthquake in Mexico (1,476 tweets).

Figure 9.9 compares the distribution of scores computed in real-time by TweetCred for the tweets on these three crisis events against a random sample of all tweets for which credibility scores were computed during the same time period. We observe that in all crisis events the credibility scores are higher than in the background distribution. This confirms the hypothesis that TweetCred gives higher credibility scores to tweets that are related to a crisis over general tweets.

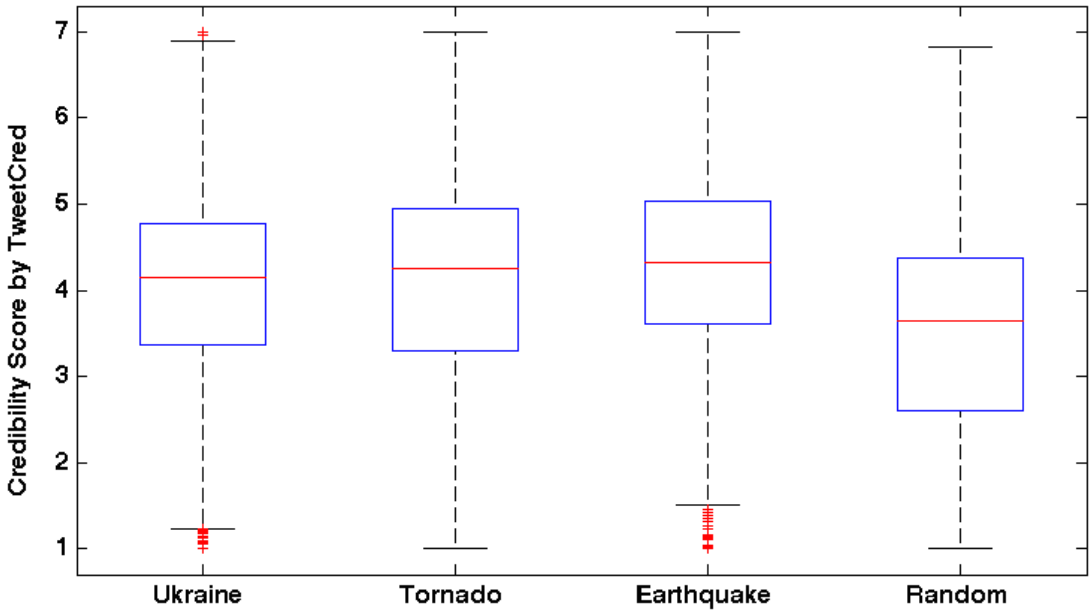



Figure 9.9: Distribution of credibility scores. We observe that during crisis events larger percentage of tweets have higher credibility than during non-crisis.

### 9.5.3 Usability Survey

To assess the overall utility and usability of the TweetCred browser extension, we conducted an online survey among its users. An unobtrusive link to the survey appeared on the right corner of Chrome's address bar when users visited Twitter. [13] The survey link was accessible only to those users who had installed the extension, this was done to ensure that only actual users of the system gave their feedback. A total of 73 users participated. The survey contained the standard 10 questions of the *System Usability Scale* (SUS) [10]. In addition to SUS questions, we also added questions about users' demographics such as gender, age, etc. We obtained an overall SUS score of 70 for TweetCred, which is considered above average from a system's usability perspective [86]. In the survey, 74% of the users found TweetCred easy to use (agree/strongly agree); 26% of the users thought there were inconsistencies in the system (agree/strongly agree); and 78% said that they may like to use TweetCred in their daily life. The complete survey responses are summarized in Appendix section of this thesis.

### 9.5.4 User Comments

TweetCred system was appreciated by users for its novelty and ease of use. Users also expressed their desire to know more about the system and its backend functionality. One recurring concern of

[13] `http://twitdigest.iiitd.edu.in/TweetCred/feedback.html`

users was related to the negative bias of the credibility scores. Users expressed that the credibility score given by TweetCred were low, even for tweets from close contacts in which they fully trust. For instance, one of the user of TweetCred said: *"People who I follow, who I know are credible, get a low rating on their tweets"*. Such local friendships and trust relationships are not captured by a generalized model built on the entire Twitter space. Other comments we received about TweetCred in the survey and from tweets about TweetCred were:

- "I plan on using this to monitor public safety situations on behalf of the City of [withheld]'s Office of Emergency Management."
- "Very clever idea but Twitter's strength is simplicity - I found this a distraction for daily use."
- "It's been good using #TweetCred & will stick around with it, thanks!"
- "It's unclear what the 3, 4 or 5 point rating mean on opinions / jokes, versus factual statements."
- "I'm using TweetCred and am very happy about usability and presentation of information."

## 9.6 Discussion

We have described TweetCred, a real-time web-based system to automatically evaluate the credibility of content on Twitter. The system provides a credibility rating from 1 (low credibility) to 7 (high credibility) for each tweet on a user's Twitter timeline. The score is computed using a supervised automated ranking algorithm, trained on human labels obtained using crowdsourcing, that determines credibility of a tweet based on more than 45 features. All features can be computed for a single tweet, and they include the tweets content, characteristics of its author, and information about external URLs. Our evaluation shows that both in terms of performance, accuracy, and usability, it is possible to bring automatic credibility ratings to users on a large scale. For the current deployment of TweetCred, analysis of false positive scenarios was not performed. In future, we would aim to analyze in detail the credibility score computed during the real-world experiment with particular emphasis on minimizing false positive cases.

TweetCred's deployment stirred a wide debate on Twitter regarding the problem and solutions for the credibility assessment problem on Twitter. We can say that social media users expect technologies that help them evaluate the credibility of the content they read. TweetCred is a first step towards fulfilling this expectation. Figures 9.10 - 9.14 some snapshots of media and online publicity about TweetCred.

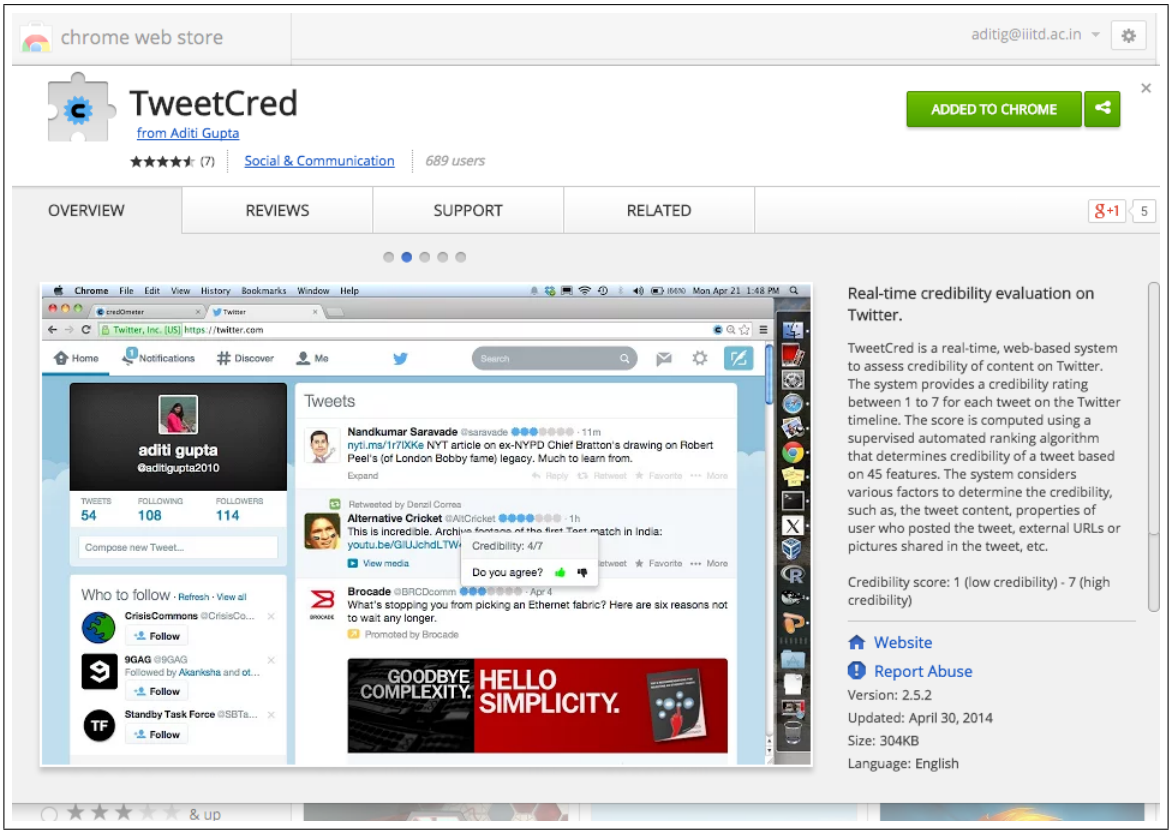



Figure 9.10: Chrome Web Store page of TweetCred

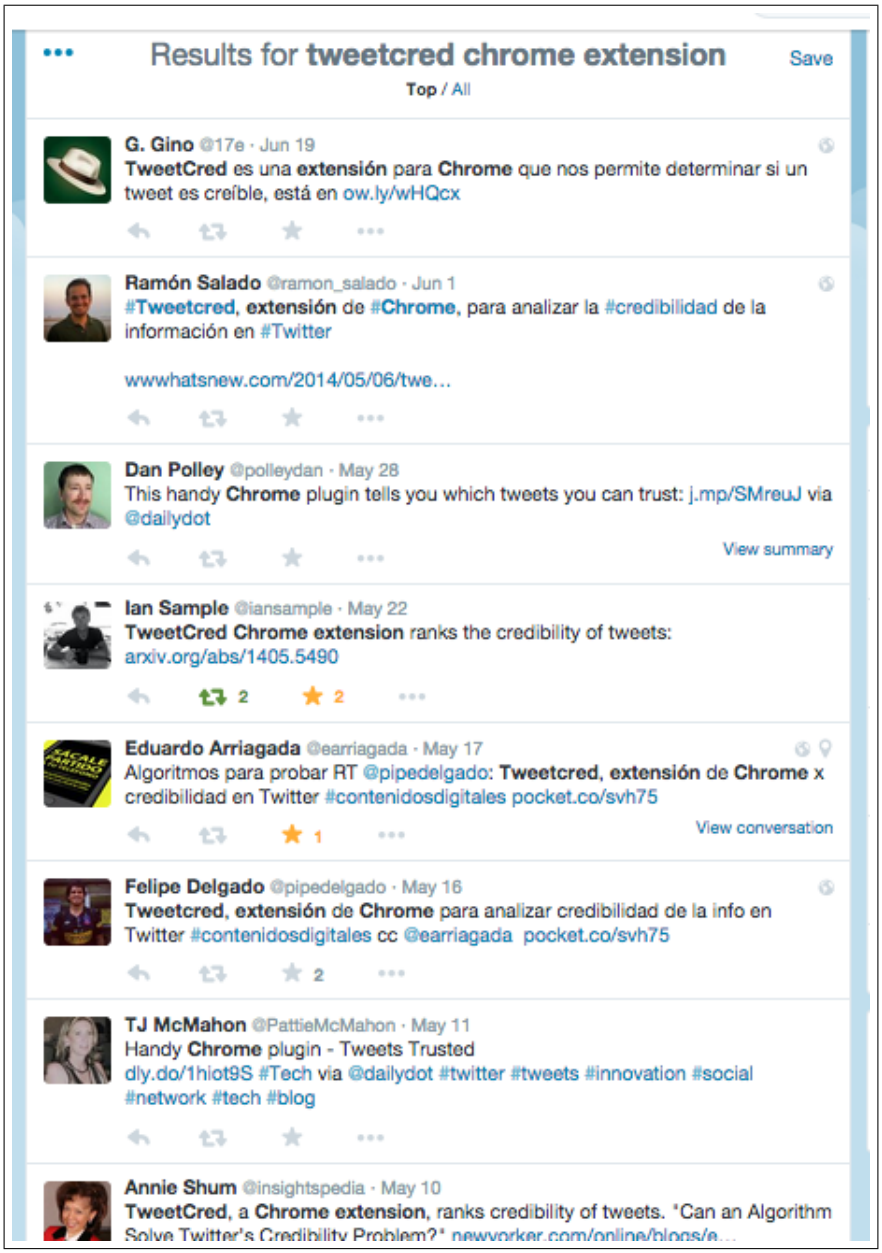
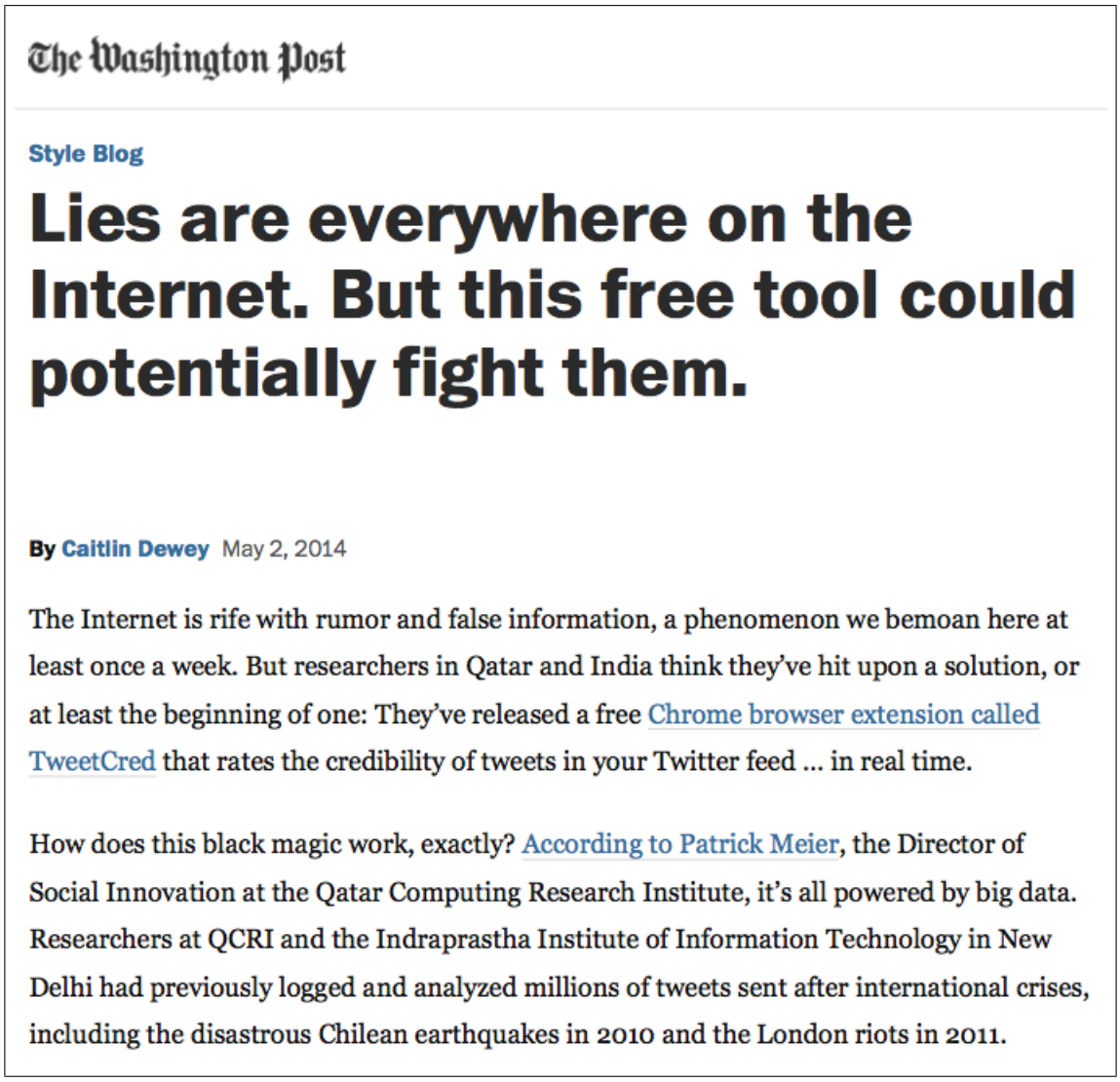



Figure 9.11: Posts on Twitter about the TweetCred Chrome Extension



The Washington Post

Style Blog

# Lies are everywhere on the Internet. But this free tool could potentially fight them.

By Caitlin Dewey May 2, 2014

The Internet is rife with rumor and false information, a phenomenon we bemoan here at least once a week. But researchers in Qatar and India think they've hit upon a solution, or at least the beginning of one: They've released a free Chrome browser extension called TweetCred that rates the credibility of tweets in your Twitter feed ... in real time.

How does this black magic work, exactly? According to Patrick Meier, the Director of Social Innovation at the Qatar Computing Research Institute, it's all powered by big data. Researchers at QCRI and the Indraprastha Institute of Information Technology in New Delhi had previously logged and analyzed millions of tweets sent after international crises, including the disastrous Chilean earthquakes in 2010 and the London riots in 2011.

Figure 9.12: Article in Washington Post about TweetCred System

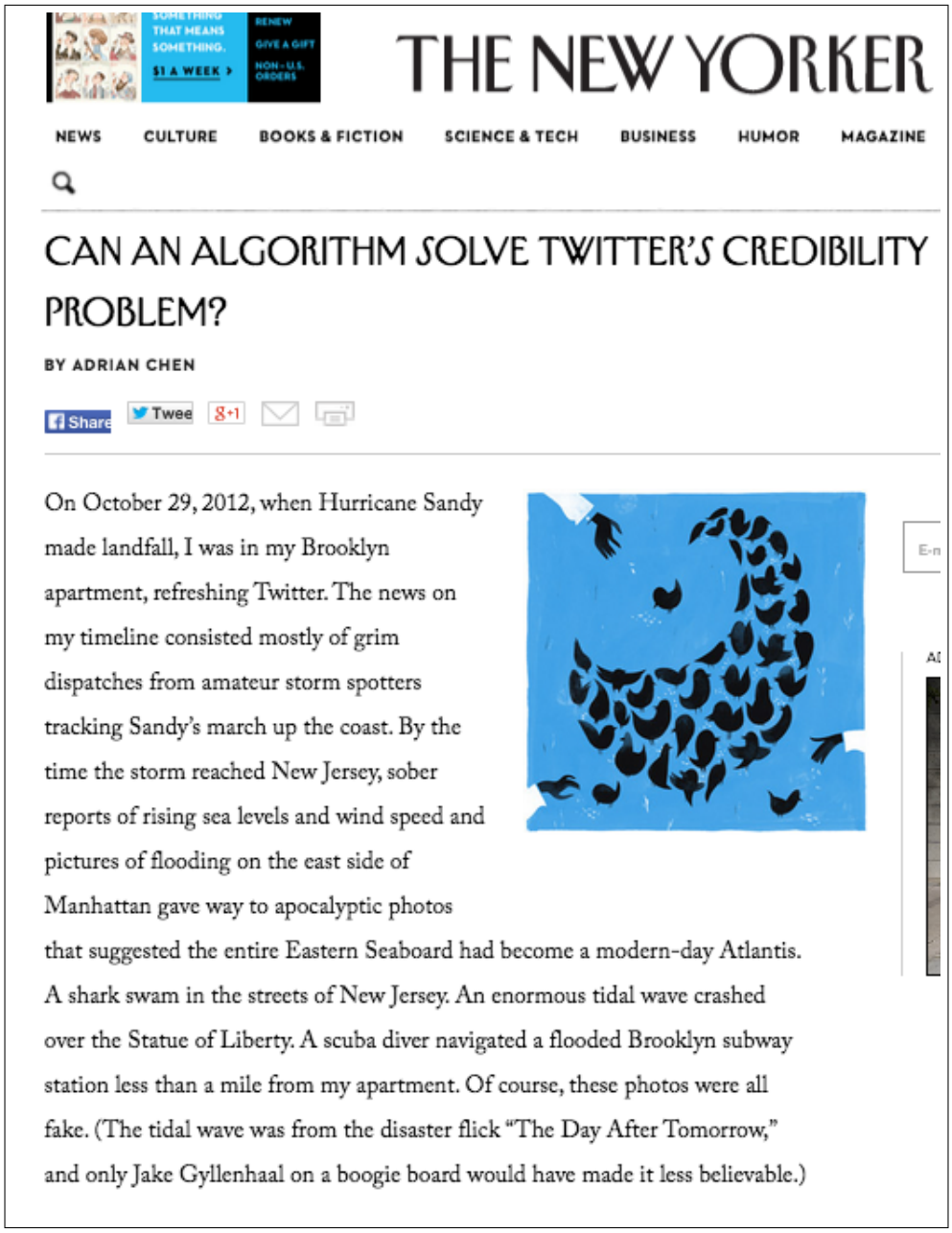



Figure 9.13: Article in NewYorker Magazine about TweetCred System

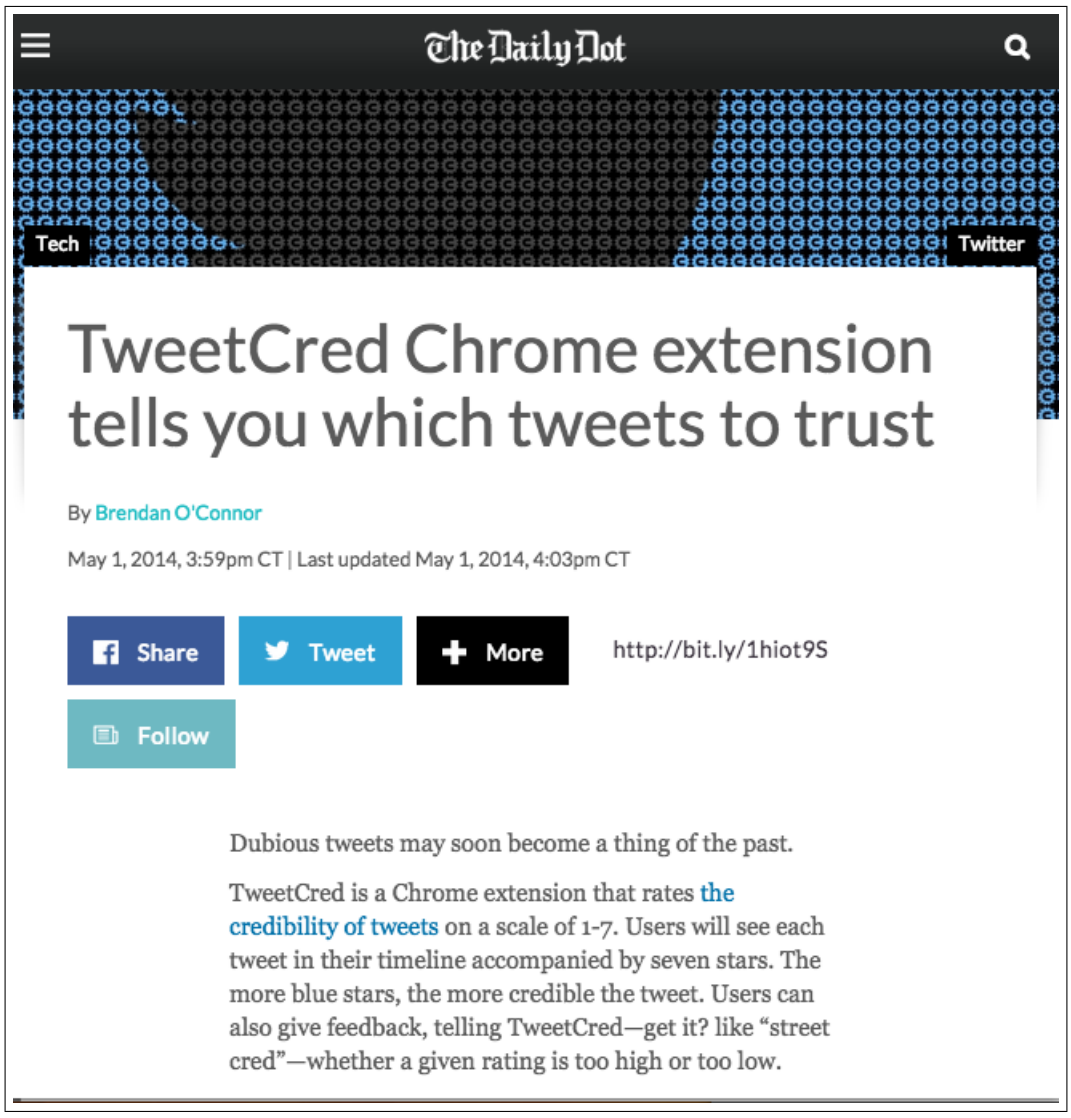



Figure 9.14: Article in DailyDot about TweetCred System.

# Chapter 10

# Conclusion and Future Work

In this chapter, we summarize the various aspects of analyzing and designing techniques to mitigate the spread of misinformation on microblogging websites. We considered Twitter, a popular microblogging platform, as our medium of study. As stated in the introduction, the goal of our research is to analyze the domain of misinformation spread on Twitter, along with proposing and evaluating a real-time solution to assess trustworthiness of content on Twitter. The first section presents a summary of our research contributions and Section 10.2 outlines directions for future research.

## 10.1 Summary

We provide a detailed analysis of misinformation spread on Twitter and propose automated techniques to assess the trustworthiness in real-time. As part of this thesis work, we built *TweetCred* that is used effectively by emergency responders, firefighters, journalists and general users to identify credibility of Twitter. We demonstrated that measuring credibility of Twitter content is possible using automated techniques, and the results are valuable for end users. The main contributions of this thesis are described below.

### 10.1.1 Characterizing Misinformation and Fake Content

We analyzed Twitter data of various events for the spread of fake images, rumors, and untrustworthy content. Some of the prominent events analyzed by us are: Mumbai blasts (2011), England Riots (2011), Hurricane Sandy (2012), Boston Marathon Blasts (2013), Polar Vortex (2014). We identified tens of thousands of tweets containing fake images, rumors, fake websites, and by malicious user profiles for these events. We performed an in-depth characterization study of how this fake versus

the real data is introduced and disseminated in the Twitter network. During Hurricane Sandy, we identified 10,350 unique tweets by 10,215 users containing fake images that were circulated on Twitter. Eighty-six percent of tweets spreading the fake images were retweets, hence very few were original tweets [6]. Our results showed that top thirty users out of 10,215 users (0.3%) were responsible for 90% of the retweets of fake image. We analyzed the role of Twitter's follower-followee network in propagating the fake images. We crawled the follower network of users and applied our algorithm to compute the overlap between fake image retweets and follower network. We found only 11% overlap between the retweet and follower graphs for the users who tweeted fake images of Sandy. This highlights that during such crisis events, users retweet information from other users irrespective of the fact whether they follow them or not. During the event of Boston Marathon Blasts, which occurred on April, 15th, 2013, a lot of fake content and malicious profiles originated on Twitter network [5]. We found that 29% of the most viral content on Twitter during the Boston event were rumors and fake; while 51% was generic opinions and comments, and rest was correct information. We found that a lot of users with late number of followers and verified accounts were responsible for spreading the fake content. Many malicious accounts were created on Twitter during the Boston event, which were later suspended by Twitter. We identified over 6,000 such user profiles; we observed that the creation of such profiles surged considerably right after the blasts occurred. We identified closed community structure and star formation in the interaction network of these suspended profiles amongst themselves.

### 10.1.2 Effectiveness of Automatic Methods to Identify Misinformation

We showed the effectiveness of supervised and unsupervised techniques in detecting fake and rumor content using meta-data, network, event and temporal features. We used classification models, to identify fake images from real images of Hurricane Sandy [6]. Decision Tree classifier gave the best results; we got an accuracy of 97% in predicting fake images from real. Tweet based features are very effective in distinguishing tweets with fake images from tweets with real images, while the performance of user based features was very poor. For rumor tweets identified during Boston Marathon, we applied linear regression model to predict how viral fake content would in future based on the attributes and impact of users currently propagating the content. In addition to basic user attributes like number of followers, friends, and verified accounts, we also defined an overall impact metric to measure impact of users who propagate a tweet in making a tweet viral. Impact of a user was computed based on attributes such as his social reputation, global engagement, topic engagement, likability, and credibility. We used regression prediction model, to verify that, overall impact of all users who propagate the fake content at a given time, can be used to estimate the growth of that content in future [5]. For three events, we also applied the unsupervised method of spectral clustering to cluster users based on a similarity metric, taking into account the content, link and meta-data similarities of the users [1]. We then used the degree centrality measure from

social network analysis to extract top users from each of the communities formed. We showed that the most central users represent the topics and opinions of the entire community with an average 81% accuracy.

### 10.1.3 Design and Build Systems to Assess Credibility of Content on Twitter

In order to mitigate misinformation spread on Twitter, we proposed that automated methods and solutions based on supervised ranking techniques can be effectively used. We proposed and evaluated a framework for assessing credibility of information in tweets using supervised machine learning and relevance feedback approach [3]. We analyzed the credibility of information in tweets corresponding to fourteen high impact news events of 2011 around the globe. From the data we analyzed, on average 30% of total tweets posted about an event contained situational information about the event while 14% was spam [2]. Only 17% of the total tweets posted about the event contained situational awareness information that was credible. Using regression analysis, we identified the important content and sourced based features, which can predict the credibility of information in a tweet. Prominent content based features were number of unique characters, swear words, pronouns, and emoticons in a tweet, and user based features like the number of followers and length of username. We adopted a supervised machine learning and relevance feedback approach using the above features, to rank tweets according to their credibility score. The performance of our ranking algorithm significantly enhanced when we applied re-ranking strategy. Results showed that extraction of credible information from Twitter can be automated with high confidence. Next we developed and deployed *TweetCred*, a system to provide an indication to Twitter users about trustworthiness of tweets in real-time within the Twitter ecosystem [4]. *TweetCred*, available as a browser plug-in, has been installed and used by 1,808 real Twitter users (available online now). [1] During ten months of its deployment, the credibility score for about 14.2 million tweets was computed, allowing us to evaluate *TweetCred* in terms of accuracy, response time, effectiveness and usability. The system provides a credibility rating from 1 (low credibility) to 7 (high credibility) for each tweet on a user's Twitter timeline. The score is computed using a supervised automated ranking algorithm, trained on human labels obtained using crowdsourcing, which determines the credibility of a tweet based on more than 45 features. All features can be computed for a single tweet, and they include the tweets content, characteristics of its author, and information about external URLs. We evaluated the performance of *TweetCred* in terms of response time, effectiveness and usability. We observed that 80% of the credibility scores are computed and displayed within 6 seconds, and that 63% of users either agreed with our automatically-generated scores or disagreed by 1 or 2 points (on a scale from 1 to 7). *TweetCred*'s deployment stirred a wide debate on Twitter regarding the problem and solutions for the credibility assessment problem on Twitter.

[1] http://twitdigest.iiitd.edu.in/TweetCred/

The browser extension featured in many news websites and blogs including the Washington Post, the New Yorker, and the Daily Dot among others.

## 10.2 Future Work and Limitations

Currently we analyzed the real-world deployment of TweetCred by analyzing the logs and feedback of the users, it may also be useful to analyze the working and usage of TweetCred by its users during a real-world event in detail. In future, we would like to analyze the role of false positives credibility scores given by TweetCred during the real-world deployment of the system. In particular we would like to analyze the occurrence of incorrect credibility scores given by TweetCred during specific events tweets versus generic tweets.

The insights obtained from this work can be used to build solutions for the problem of trustworthiness of user generated content on different online social media. As part of this thesis, we focussed only on microblogging websites, it would be interesting to apply similar methodologies on other OSM, for example on Facebook the feature vectors would be quite different from those on Twitter. There are many similarities and differences between these different services. In this thesis we focussed on only Twitter primarily for two reasons: First it is a prominent news sharing platform and and secondly it has readily accessible data for research. The methodology and research framework used in our work provide useful insights and can be used to build similar solutions for Facebook and other social networks. Due to inherent difference in the attributes of each social network, the feature sets and learning models would vary across different social networks. For example, Twitter posts have only 140 characters, while a Facebook post can be more than 5,000 characters long, indicating more extensible text analysis can be performed for Facebook.

Apart from developing techniques to automatically assess credibility, there is also a need to understand how users perceive credibility of content posted online. Understanding user perceptions can help to develop better trust assessment tools which can help end users to judge the quality of content in a better way. In future there is a dire need to build specialized real-time solutions for emergency responders and organizations, that can monitor and display crisis related tweets in an innovative way. In addition to credibility of content, the solutions should also have the functionality that can automatically identify tweets that are timely, well-written, novel, posted by reputable users, etc. Such a system would increase utilization of OSM data for crisis response and management.

We analyzed, designed solutions and evaluated effectiveness of the use of computational automated techniques in solving the problem of trust and credibility on microblogging websites. The output from our work resulted in a real-world deployment, which is till date being used by real users of Twitter.

# Bibliography


[1] Abel, F., Gao, Q., Houben, G.-J., and Tao, K. Analyzing user modeling on twitter for personalized news recommendations. In *Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization* (Berlin, Heidelberg, 2011), UMAP'11, Springer-Verlag, pp. 1–12.

[2] Adetunji, J. 'twitter terrorists' face 30 years after being charged in mexico. *http://www.guardian.co.uk/world/2011/sep/04/twitter-terrorists-face-30-years* (2011).

[3] Amanda L. Hughes, L. P. Twitter Adoption and Use in Mass Convergence and Emergency Events. *6th International Conference on Information Systems for Crisis Response and Management (ISCRAM)* (2009).

[4] Anupama Aggarwal, Ashwin Rajadesingan, P. K. Phishari: Automatic realtime phishing detection on twitter. *7th IEEE APWG eCrime Researchers Summit (eCRS)* (2012).

[5] Anwar, J. Mumbai blasts: People reach out to victims through Twitter, Facebook, 2011. Retrieved July 13, 2011, http://timesofindia.indiatimes.com/city/mumbai/Mumbai-blasts-People-reach-out-to-victims-through-Twitter-Facebook/articleshow/9215704.cms.

[6] Anwar, J. Govt blocks Twitter accounts of some journalists, 2012. Retrieved August 23, 2012, http://timesofindia.indiatimes.com/tech/tech-news/Govt-blocks-Twitter-accounts-of-some-journalists/articleshow/15612767.cms.

[7] APWG. Phishing Activity Trends Report, 2013. Retrieved April 24, 2013, by Anti Phishing Working Group (APWG), http://docs.apwg.org/reports/apwg_trends_report_Q4_2012.pdf.

[8] Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., and Gonçalves, M. Detecting spammers and content promoters in online video social networks. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 2009), SIGIR '09, ACM, pp. 620–627.

[9] Boorstin, J. FIFA?s World Cup final breaks social media records, 2014. Retrieved , http://www.cnbc.com/id/101835109.

[10] Brooke, J. SUS: A quick and dirty usability scale. In *Usability evaluation in industry*, P. W. Jordan, B. Weerdmeester, A. Thomas, and I. L. Mclelland, Eds. Taylor and Francis, London, 1996.

[11] Bryan Semaan, G. M. 'facebooking' towards crisis recovery and beyond: disruption as an opportunity. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (2012), CSCW '12.

[12] Buckley, C., Salton, G., and Allan, J. Automatic retrieval with locality information using SMART. *NIST special publication*, 500207 (1993), 59–72.

[13] Canini, K. R., Suh, B., and Pirolli, P. L. Finding credible information sources in social networks based on content and social structure. In *SocialCom* (2011).

[14] Carley, K., Dombroski, M., Tsvetovat, M., Reminga, J., and Kamneva, N. Destabilizing dynamic covert networks. In *Proceedings of the 8th International Command and Control Research* (Vienna, VA, 2003), Evidence Based Research.

[15] Castillo, C., El-Haddad, M., Pfeffer, J., and Stempeck, M. Characterizing the life cycle of online news stories using social media reactions. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (New York, NY, USA, 2014), CSCW '14, ACM, pp. 211–223.

[16] Castillo, C., Mendoza, M., and Poblete, B. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web* (New York, NY, USA, 2011), WWW '11, ACM, pp. 675–684.

[17] CBS. Brits get 4 years prison for facebook riot posts. *http://www.cbsnews.com/2100-202_162-20093364.html* (2011).

[18] Center, P. R. Social Networking Fact Sheet, 2015. Retrieved March 26, 2015, http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/.

[19] Chen, J., Nairn, R., Nelson, L., Bernstein, M., and Chi, E. Short and tweet: experiments on recommending content from information streams. CHI '10, pp. 1185–1194.

[20] Chhabra, S., Aggarwal, A., Benevenuto, F., and Kumaraguru, P. Phi.sh/$ocial: the phishing landscape through short urls. In *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference* (New York, NY, USA, 2011), CEAS '11, ACM, pp. 92–101.

[21] CNBC. False rumor of explosion at white house causes stocks to briefly plunge; ap confirms its twitter feed was hacked. *http://www.cnbc.com/id/100646197* (2013).

[22] Colbert, A. 7 Fake Hurricane Sandy Photos You're Sharing on Social Media, 2012. Retrieved October 30, 2012, http://mashable.com/2012/10/29/fake-hurricane-sandy-photos/.

[23] Cooper, M. Hurricane Cost Seen as Ranking Among Top Ten, 2011. Retrieved Aug 30, 2011, http://www.nytimes.com/2011/08/31/us/31floods.html.

[24] Corporation, S. Istr: internet security threat report 2013.

[25] Corvey, W. J., Verma, S., Vieweg, S., Palmer, M., and Martin, J. H. Foundations of a multilayer annotation framework for twitter communications during crisis events. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (Istanbul, Turkey, may 2012), N. C. C. Chair), K. Choukri, T. Declerck, M. U. Do?an, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds., European Language Resources Association (ELRA).

[26] De Longueville, B., Smith, R. S., and Luraschi, G. "omg, from here, i can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 International Workshop on Location Based Social Networks* (New York, NY, USA, 2009), LBSN '09, ACM, pp. 73–80.

[27] Dong, A., Zhang, R., Kolari, P., Bai, J., Diaz, F., Chang, Y., Zheng, Z., and Zha, H. Time is of the essence: improving recency ranking using twitter data. WWW '10.

[28] Duan, Y., Jiang, L., Qin, T., Zhou, M., and Shum, H.-Y. An empirical study on learning to rank of tweets. In *COLING '10*.

[29] Economist. A web of support, 2011. Retrieved , http://www.economist.com/blogs/babbage/2011/07/online-crisis-management.

[30] Ehealthcoordinator. Crisis management: using Twitter and Facebook for the greater good, 2011. Retrieved July 21, 2011, https://leifhanlen.wordpress.com/2011/07/22/crisis-management-using-twitter-and-facebook-for-the-greater-good/.

[31] (FEMA), F. E. M. A. National Preparedness Report, 2013. Retrieved March 30, 2013, http://www.fema.gov/media-library-data/20130726-1916-25045-0015/npr2013_final.pdf.

[32] Fogg, B., Marshall, J., Kameda, T., Solomon, J., Rangnekar, A., Boyd, J., and Brown, B. Web credibility research: a method for online experiments and early study results. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems* (2001), CHI EA '01.

[33] Fogg, B. J. Prominence-interpretation theory: explaining how people assess credibility online. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems* (2003), CHI EA '03.

[34] Fogg, B. J., and Tseng, H. The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (1999), CHI '99.

[35] France Cheong, C. C. Social media data mining: A social network analysis of tweets during the 2010-2011 australian floods. In *PACIS* (2011).

[36] Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res. 4* (2003), 933–969.

[37] Friggeri, A., Adamic, L., Eckles, D., and Cheng, J. Rumor cascades, 2014.

[38] Gabbatt, A. How companies used social media during Hurricane Sandy, 2013. Retrieved February 20, 2013, http://www.guardian.co.uk/world/us-news-blog/2013/feb/20/mta-conedison-hurricane-sandy-social-media-week.

[39] Ghosh, S., Sharma, N., Benevenuto, F., Ganguly, N., and Gummadi, K. Cognos: crowdsourcing search for topic experts in microblogs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (2012), SIGIR '12.

[40] Ghosh, S., Viswanath, B., Kooti, F., Sharma, N. K., Korlam, G., Benevenuto, F., Ganguly, N., and PhaniGummadi, K. Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st international conference on World Wide Web* (2012), WWW '12.

[41] Grier, C., Thomas, K., Paxson, V., and Zhang, M. @spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security* (New York, NY, USA, 2010), CCS '10, ACM, pp. 27–37.

[42] Gupta, A., Chhabra, A., and Kumaraguru, P. Twit-digest: Real-time twitter search portal for extracting, tracking and visualizing information. *IBM Collaborative Academia Research Exchange (IBM-ICARE)* (2012).

[43] Gupta, A., Gupta, M., and Kumaraguru, P. Twit-digest version 2: An online solution for analyzing and visualizing twitter in real-time. *2nd Security and Privacy Symposium IIT, Kanpur* (2014).

[44] Gupta, A., Joshi, A., and Kumaraguru, P. Identifying and characterizing user communities on twitter during crisis events. In *Proceedings of the 2012 workshop on Data-driven user behavioral modelling and mining from social media, in CIKM companion* (2012), ACM.

[45] Gupta, A., and Kumaraguru, P. Credibility ranking of tweets during high impact events. *Workshop on Privacy and Security in Online Social Media, Proceedings of the 21st international conference on World Wide Web companion (WWW)* (2012).

[46] Gupta, A., and Kumaraguru, P. Misinformation on twitter during crisis events. In *Encyclopedia of Social Network Analysis and Mining (ESNAM)*. Springer publications, 2012.

[47] Gupta, A., and Kumaraguru, P. Twitter explodes with activity in mumbai blasts! a lifeline or an unmonitored daemon in the lurking?

[48] Gupta, A., Kumaraguru, P., Castillo, C., and Meier, P. Tweetcred: Real-time credibility assessment of content on twitter. In *Proceedings of the 6th International Conference on Social Informatics* (2014), SocInfo 14.

[49] Gupta, A., Lamba, H., and Kumaraguru, P. $1.00 per rt #bostonmarathon #prayforboston: Analyzing fake content on Twitter. In *Proc. Eighth IEEE APWG eCrime Research Summit (eCRS)* (2013), IEEE, p. 12.

[50] Gupta, A., Lamba, H., Kumaraguru, P., and Joshi, A. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Second Workshop on Privacy and Security in Online Social Media, Proceedings of the 22nd international conference on World Wide Web companion (WWW)* (2013).

[51] Hughes, A. l., and Palen, L. Twitter adoption and use in mass convergence and emergency events. In *Proceedings of the 2009 ISCRAM Conference* (2009).

[52] Järvelin, K., and Kekäläinen, J. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems 20* (2002), 2002.

[53] Java et al. Detecting Commmunities via Simultaneous Clustering of Graphs and Folksonomies. In *Proceedings of the Tenth Workshop on Web Mining and Web Usage Analysis (WebKDD)* (August 2008), ACM.

[54] Jean Burgess, F. V., and Bruns, A. How many fake Sandy pictures were really shared on social media?, 2012. Retrieved November 6, 2012, http://www.guardian.co.uk/news/datablog/2012/nov/06/fake-sandy-pictures-social-media.

[55] Joachims, T. Optimizing search engines using clickthrough data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)* (2002), pp. 133–142.

[56] Karatzoglou, A. specc {kernlab}. http://www.inside-r.org/packages/cran/kernlab/docs/specc.

[57] Kumaraguru, P., Cranshaw, J., Acquisti, A., Cranor, L., Hong, J., Blair, M. A., and Pham, T. School of phish: a real-world evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security* (2009), SOUPS '09.

[58] Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L. F., Hong, J., and Nunge, E. Protecting people from phishing: the design and evaluation of an embedded training email system. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems* (2007), pp. 905–914.

[59] Kwak, H., Lee, C., Park, H., and Moon, S. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web* (New York, NY, USA, 2010), WWW '10, ACM, pp. 591–600.

[60] Landis, J. R., and Koch, G. G. The Measurement of Observer Agreement for Categorical Data. *Biometrics 33*, 1 (Mar. 1977), 159–174.

[61] Lu, R., Xu, Z., Zhang, Y., and Yang, Q. Life activity modeling of news event on twitter using energy function.

[62] Maron, D. F. How Social Media Is Changing Disaster Response, 2013. Retrieved June 7, 2013, http://www.scientificamerican.com/article/how-social-media-is-changing-disaster-response/.

[63] Mendoza, M., Poblete, B., and Castillo, C. Twitter under crisis: can we trust what we rt? In *Proceedings of the First Workshop on Social Media Analytics* (New York, NY, USA, 2010), SOMA '10, ACM, pp. 71–79.

[64] Metzler, D., and Croft, W. B. Linear feature-based models for information retrieval. *Information Retrieval 10*, 3 (2007), 257–274.

[65] Moffit, J. Tweeting in the Rain, Part 4: Tweets during the 2013 Colorado flood, 2014. Retrieved September 12, 2014, https://blog.gnip.com/tweeting-rain-part-4-tweets-2013-colorado-flood/.

[66] Morris, M. R., Counts, S., Roseway, A., Hoff, A., and Schwarz, J. Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (New York, NY, USA, 2012), CSCW '12, ACM, pp. 441–450.

[67] Nazir, A., Raza, S., Chuah, C.-N., and Schipper, B. Ghostbusting facebook: detecting and characterizing phantom profiles in online social gaming applications. In *Proceedings of the 3rd conference on Online social networks* (2010), WOSN'10.

[68] Newman et al. Analysis of weighted networks. *Phys Rev E Stat Nonlin Soft Matter Phys 70*, 5 Pt 2 (Nov. 2004).

[69] O'Donovan, J., Kang, B., Meyer, G., Hllerer, T., and Adali, S. Credibility in context: An analysis of feature distributions in twitter. *ASE/IEEE International Conference on Social Computing, SocialCom* (2012).

[70] Oh, O., Agrawal, M., and Rao, H. R. Information control and terrorism: Tracking the mumbai terrorist attack through twitter. *Information Systems Frontiers 13*, 1 (Mar. 2011), 33–43.

[71] Page, L., Brin, S., Motwani, R., and Winograd, T. The PageRank citation ranking: Bringing order to the web. Tech. rep., Stanford Digital Library Technologies Project, 1998.

[72] Palen, L., Anderson, K. M., Mark, G., Martin, J., Sicker, D., Palmer, M., and Grunwald, D. A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters. In *Proceedings of the 2010 ACM-BCS Visions of Computer Science Conference* (2010), ACM-BCS '10.

[73] Palen, L., and Vieweg, S. The emergence of online widescale interaction in unexpected events: assistance, alliance & retreat. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work* (New York, NY, USA, 2008), CSCW '08, ACM, pp. 117–126.

[74] Qazvinian, V., Rosengren, E., Radev, D. R., and Mei, Q. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, PA, USA, 2011), EMNLP '11, Association for Computational Linguistics, pp. 1589–1599.

[75] RADFORD, B. Social Media Ebola Hoax Causes Deaths, 2014. Retrieved October 1, 2014, http://news.discovery.com/human/psychology/social-media-ebola-hoax-causes-deaths-14100.htm.

[76] Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., and Menczer, F. Truthy: mapping the spread of astroturf in microblog streams. WWW '11.

[77] Reinwald, C. What Twitter Got Wrong During the Week Following Last Year?s Boston Marathon, 2014. Retrieved April 18, 2014, http://www.boston.com/news/local/massachusetts/2014/04/18/what-twitter-got-wrong-during-the-boston-marathon-bombing-week/ZOYLJpEydYgJ8UYNUT674H/story.html.

[78] Reuter, C., Ludwig, T., Friberg, T., Moi, M., Akerkar, R., Pratzler W, S., Gizikis, A., and Brien, T. O. Emergent: Emergency management in social media generation.

[79] Richards, J., and Lewis, P. How twitter was used to spread and knock down rumours during the riots. *http://www.guardian.co.uk/uk/2011/dec/07/how-twitter-spread-rumours-riots* (2011).

[80] Ricketts, S. How the Boston Marathon explosions reveal the two sides of Twitter, 2013. Retrieved April 15, 2013, http://www.guardian.co.uk/commentisfree/2013/apr/16/boston-marathon-explosions-reveal-twitter.

[81] Robertson, S. E., Walker, S., and Beaulieu, M. Okapi at trec-7: automatic ad hoc, filtering, vlc and interactive track. *IN* (1999).

[82] Rogers, S. #BRAvsCRO: 7 #WorldCup @TwitterData points you need to know, 2014. Retrieved June 13, 2014, https://blog.twitter.com/2014/bravscro-7-worldcup-twitterdata-points-you-need-to-know.

[83] Rogers, S. Insights into the #WorldCup conversation on Twitter, 2014. Retrieved July 14, 2014, https://blog.twitter.com/2014/insights-into-the-worldcup-conversation-on-twitter.

[84] Sakaki, T., Okazaki, M., and Matsuo, Y. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (New York, NY, USA, 2010), WWW '10, ACM, pp. 851–860.

[85] Sakaki, T., Toriumi, F., and Matsuo, Y. Tweet trend analysis in an emergency situation. In *Proceedings of the Special Workshop on Internet and Disasters* (New York, NY, USA, 2011), SWID '11, ACM, pp. 3:1–3:8.

[86] Sauro, J. Measuring Usability With The System Usability Scale (SUS), 2011. Retrieved February 2, 2011, http://www.measuringusability.com/sus.php.

[87] SEEN. Boston Marathon 2013, 2013. Retrieved April 15, 2013, http://beta.seen.co/event/boston-marathon-2013-boston-ma-2013-7033.

[88] Shi et al. Normalized cuts and image segmentation. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*.

[89] Stancati, M. Mumbai Blasts: Did Twitter Really Help?, 2011. Retrieved July 15, 2011, http://blogs.wsj.com/indiarealtime/2011/07/15/mumbai-blasts-did-twitter-really-help/.

[90] Stanford. Okapi BM25: a non-binary model, 2009. Retrieved 4 July, 2009, http://nlp.stanford.edu/IR-book/html/htmledition/okapi-bm25-a-non-binary-model-1.html.

[91] Stanford. Tf-idf weighting, 2009. Retrieved July 4, 2009, http://nlp.stanford.edu/IR-book/html/htmledition/tf-idf-weighting-1.html.

[92] Sureka, A., Kumaraguru, P., Goyal, A., and Chhabra, S. Mining YouTube to Discover Hate Videos, Users and Hidden Communities. *Accepted at Sixth Asia Information Retrieval Societies Conference* (2010).

[93] Telegraph. Boston Marathon explosions: three dead, dozens injured as 'bombs' hit race finish line, 2013. Retrieved April 16, 2013, http://www.telegraph.co.uk/news/worldnews/northamerica

[94] theTelegraph. Boston marathon blast injuries toll at 264. *http://www.dailytelegraph.com.au/boston-marathon-blast-injuries-toll-at-264/story-e6freuz9-1226628261632* (2013).

[95] Twitter. Streaming api. *https://dev.twitter.com/docs/streaming-apis* (2013).

[96] Vagias, W. M. Likert-type scale response anchors, 2006. Retrieved 2006, http://www.clemson.edu/centers-institutes/tourism/documents/sample-scales.pdf.

[97] Verma, S., Vieweg, S., Corvey, W., Palen, L., Martin, J. H., Palmer, M., Schram, A., and Anderson, K. M. Natural language processing to the rescue? extracting "situational awareness" tweets during mass emergency. In *ICWSM* (2011), L. A. Adamic, R. A. Baeza-Yates, and S. Counts, Eds., The AAAI Press.

[98] Vieweg, S., Hughes, A. L., Starbird, K., and Palen, L. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the 28th international conference on Human factors in computing systems* (New York, NY, USA, 2010), CHI '10, ACM, pp. 1079–1088.

[99] Viswanath, B., Bashir, M. A., Crovella, M., Guha, S., Gummadi, K. P., Krishnamurthy, B., and Mislove, A. Towards detecting anomalous user behavior in online social networks. In *Proceedings of the 23rd USENIX Security Symposium (USENIX Security)* / (2014).

[100] Wang et al. Detecting community kernels in large social networks. In *Proceedings of International Conference on Data Mining*, ICDM '11.

[101] Week, T. 10 fake photos of Hurricane Sandy, 2012. Retrieved October 30, 2012, http://theweek.com/article/index/235578/10-fake-photos-of-hurricane-sandy.

[102] Welch et al. Topical semantics of twitter links. In *Proceedings of the fourth international conference on Web search and data mining*, WSDM '11.

[103] Xia, X., Yang, X., Wu, C., Li, S., and Bao, L. Information credibility on twitter in emergency situation. In *Proceedings of the 2012 Pacific Asia conference on Intelligence and Security Informatics* (2012), PAISI'12.

[104] Xiang, G., Fan, B., Wang, L., Hong, J., and Rose, C. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (New York, NY, USA, 2012), CIKM '12, ACM, pp. 1980–1984.

[105] Xu, J., and Li, H. Adarank: A boosting algorithm for information retrieval. In *Proc. SIGIR* (New York, NY, USA, 2007), ACM, pp. 391–398.

[106] Xu, Z., Zhang, Y., Wu, Y., and Yang, Q. Modeling user posting behavior on social media. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2012), SIGIR '12, ACM, pp. 545–554.

[107] Yahoo. 10 fake photos of Hurricane Sandy, 2012. Retrieved October 30, 2012, http://news.yahoo.com/10-fake-photos-hurricane-sandy-075500934.html.

[108] Yang, C., Harkreader, R., Zhang, J., Shin, S., and Gu, G. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In *Proceedings of the 21st international conference on World Wide Web* (2012), WWW '12.

[109] Yardi, S., Romero, D., Schoenebeck, G., and Boyd, D. Detecting spam in a Twitter network. *First Monday 15*, 1 (Jan. 2010).

[110] Yin, H., Cui, B., Chen, L., Hu, Z., and Zhou, X. Dynamic user modeling in social media systems. *ACM Trans. Inf. Syst. 33*, 3 (Mar. 2015), 10:1–10:44.

[111] Zeldin, W. Venezuela: Twitter users arrested on charges of spreading rumors. *http://www.loc.gov/lawweb/servlet/lloc_news?disp3_l205402106_text* (2010).

[112] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European conference on Advances in information retrieval* (Berlin, Heidelberg, 2011), ECIR'11, Springer-Verlag, pp. 338–349.

# List of publications


[1] Gupta, A., Joshi, A., and Kumaraguru, P. Identifying and characterizing user communities on twitter during crisis events. In *Proceedings of the 2012 workshop on Data-driven user behavioral modelling and mining from social media, in CIKM companion* ACM, 2012.

[2] Gupta, A., and Kumaraguru, P. Credibility ranking of tweets during high impact events. *Workshop on Privacy and Security in Online Social Media, Proceedings of the 21st international conference on World Wide Web companion (WWW)*, ACM 2012.

[3] Gupta, A., and Kumaraguru, P. Misinformation on twitter during crisis events. In *Encyclopedia of Social Network Analysis and Mining (ESNAM)*. Book Chapter, Springer publications, 2012.

[4] Gupta, A., Kumaraguru, P., Castillo, C., and Meier, P. Tweetcred: Real-time credibility assessment of content on twitter. In *Proceedings of the 6th International Conference on Social Informatics*, (SocInfo'14), Springer publications, 2014.

[5] Gupta, A., Lamba, H., and Kumaraguru, P. $1.00 per rt #bostonmarathon #prayforboston: Analyzing fake content on Twitter. In *Proc. Eighth IEEE APWG eCrime Research Summit (eCRS)* , IEEE, 2013.

[6] Gupta, A., Lamba, H., Kumaraguru, P., and Joshi, A. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Second Workshop on Privacy and Security in Online Social Media, Proceedings of the 22nd international conference on World Wide Web companion (WWW)*, ACM, 2013.

[7] Gupta, A., Chhabra, A., and Kumaraguru, P. Twit-Digest: Real-time Twitter search portal for extracting, tracking and visualizing information. In *IBM Collaborative Academia Research Exchange (IBM-ICARE)*. Poster. 2012.

[8] Gupta, A., Gupta, M., and Kumaraguru, P. Twit-Digest Version 2: An Online Solution for Analyzing and Visualizing Twitter in Real-Time. In *2nd Security and Privacy Symposium IIT, Kanpur*. Poster. 2014.

[9] Gupta, A., Lamba, H., and Kumaraguru, P. Analyzing and Measuring Spread of Fake Content on Twitter during High Impact Events. In *2nd Security and Privacy Symposium IIT, Kanpur*. Poster. 2014.

# Appendix 1

A sample tweet JSON object returned by Twitter API is:

{“coordinates”: null,
“favorited”: false,
“truncated”: false,
“created_at”:“Mon Sep 24 03:35:21 +0000 2012”,
“id_str”:“250075927172759552”,
“entities”: {
“urls”: [ ],
“hashtags”: [{
“text”:“freebandnames”,
“indices”: [ 20, 34 ]} ],
“user_mentions”: [ ] },
“in_reply_to_user_id_str”: null,
“contributors”: null,
“text”:“Aggressive Ponytail #freebandnames”,
“metadata”: {
“iso_language_code”:“en”,
“result_type”:“recent”},
“retweet_count”: 0,
“in_reply_to_status_id_str”: null,
“id”: 250075927172759552,
“geo”: null,
“retweeted”: false,
“in_reply_to_user_id”: null,
“place”: null,
“user”: {
“profile_sidebar_fill_color”:“DDEEF6”,

“profile_sidebar_border_color”:“C0DEED”,
“profile_background_tile”: false,
“name”:“Sean Cummings”,
“profile_image_url”:“http://a0.twimg.com/profile_images/2359746665/
1v6zfgqo8g0d3mk7ii5s_normal.jpeg”,
“created_at”:“Mon Apr 26 06:01:55 +0000 2010”,
“location”:“LA, CA”,
“follow_request_sent”: null,
“profile_link_color”:“0084B4”,
“is_translator”: false,
“id_str”:“137238150”,
“entities”: {
“url”: {
“urls”: [ {
“expanded_url”: null,
“url”:“”,
“indices”: [ 0, 0 ] }] },
“description”: {
“urls”: []}},
“default_profile”: true,
“contributors_enabled”: false,
“favourites_count”: 0,
“url”: null,
“profile_image_url_https”:“https://si0.twimg.com/profile_images/2359746665
/1v6zfgqo8g0d3mk7ii5s_normal.jpeg”,
“utc_offset”: -28800,
“id”: 137238150,
“profile_use_background_image”: true,
“listed_count”: 2,
“profile_text_color”:“333333”,
“lang”:“en”,
“followers_count”: 70,
“protected”: false,
“notifications”: null,
“profile_background_image_url_https”:“https://si0.twimg.com/images/
themes/theme1/bg.png”,
“profile_background_color”:“C0DEED”,

“verified”: false,
“geo_enabled”: true,
“time_zone”:“Pacific Time (US & Canada)”,
“description”:“Born 330 Live 310”,
“default_profile_image”: false,
“profile_background_image_url”:“http://a0.twimg.com/images/themes/
theme1/bg.png”,
“statuses_count”: 579,
“friends_count”: 110,
“following”: null,
“show_all_inline_media”: false,
“screen_name”:“sean_cummings” },
“in_reply_to_screen_name”: null,
“source”:“Twitter for Mac”,
“in_reply_to_status_id”: null }

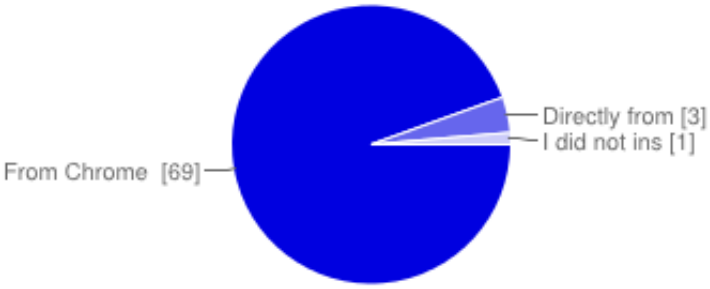
# Complete Survey Responses

# 73 responses

## TweetCred Usability Survey

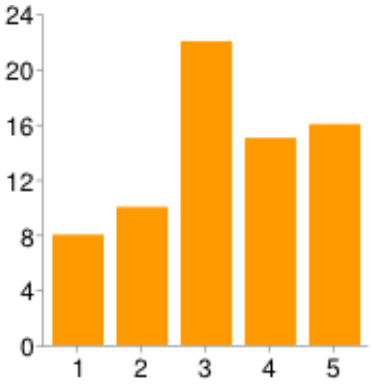
## Summary

### From where did you install the TweetCred Chrome Plugin?




| | | |
|---|---|---|
| From Chrome web store | **69** | 95% |
| Directly from a folder | **3** | 4% |
| I did not install the TweetCred Chrome Plugin | **1** | 1% |

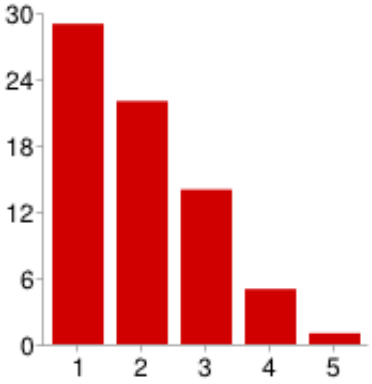
### 1. I think that I would like to use this system frequently.




| | | |
|---|---|---|
| 1 | **8** | 11% |
| 2 | **10** | 14% |
| 3 | **22** | 30% |
| 4 | **15** | 21% |
| 5 | **16** | 22% |

1: Strongly Disagree 5: Strongly Agree

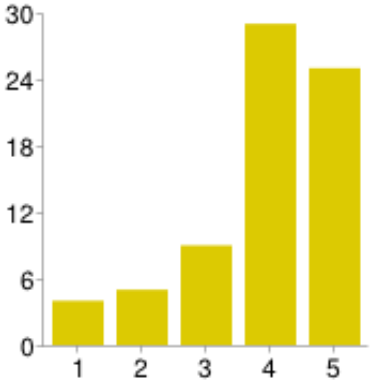
### 2. I found the system unnecessarily complex.




| | | |
|---|---|---|
| 1 | **29** | 40% |
| 2 | **22** | 30% |
| 3 | **14** | 19% |
| 4 | **5** | 7% |
| 5 | **1** | 1% |

1: Strongly Disagree 5: Strongly Agree

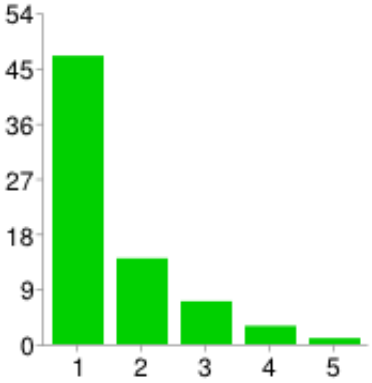
**3. I thought the system was easy to use.**




| | | |
|---|---|---|
| 1 | **4** | 5% |
| 2 | **5** | 7% |
| 3 | **9** | 12% |
| 4 | **29** | 40% |
| 5 | **25** | 34% |

1: Strongly Disagree 5: Strongly Agree

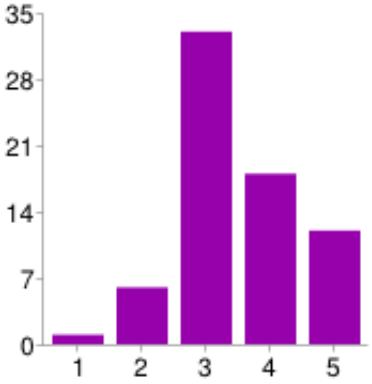
**4. I think that I would need the support of a technical person to be able to use this system.**




| | | |
|---|---|---|
| 1 | **47** | 64% |
| 2 | **14** | 19% |
| 3 | **7** | 10% |
| 4 | **3** | 4% |
| 5 | **1** | 1% |

1: Strongly Disagree 5: Strongly Agree

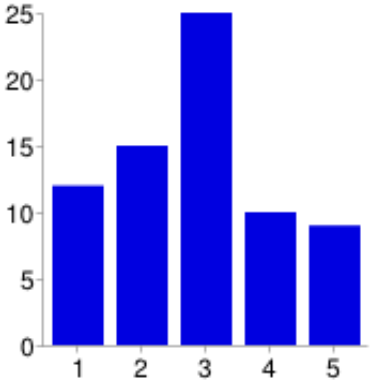
**5. I found the various functions in this system were well integrated.**




| | | |
|---|---|---|
| 1 | **1** | 1% |
| 2 | **6** | 8% |
| 3 | **33** | 45% |
| 4 | **18** | 25% |
| 5 | **12** | 16% |

1: Strongly Disagree 5: Strongly Agree

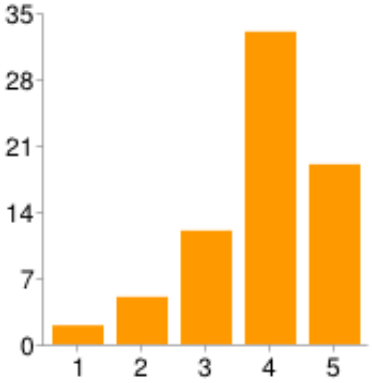
**6. I thought there was too much inconsistency in this system.**




| | | |
|---|---|---|
| 1 | **12** | 16% |
| 2 | **15** | 21% |
| 3 | **25** | 34% |
| 4 | **10** | 14% |
| 5 | **9** | 12% |

1: Strongly Disagree 5: Strongly Agree

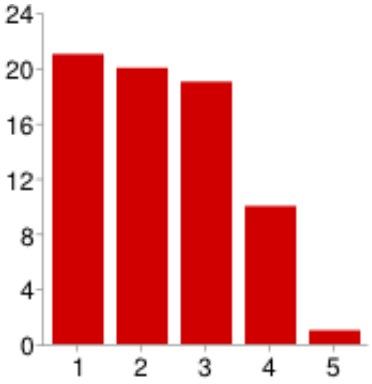
**7. I would imagine that most people would learn to use this system very quickly.**




| | | |
|---|---|---|
| 1 | **2** | 3% |
| 2 | **5** | 7% |
| 3 | **12** | 16% |
| 4 | **33** | 45% |
| 5 | **19** | 26% |

1: Strongly Disagree 5: Strongly Agree

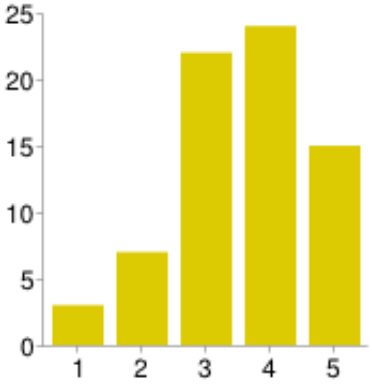
**8. I found the system very cumbersome to use.**




| | | |
|---|---|---|
| 1 | **21** | 29% |
| 2 | **20** | 27% |
| 3 | **19** | 26% |
| 4 | **10** | 14% |
| 5 | **1** | 1% |

1: Strongly Disagree 5: Strongly Agree

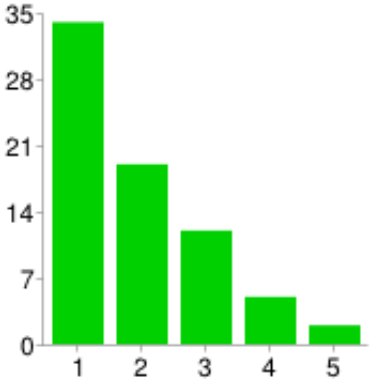
**9. I felt very confident using the system.**




| | | |
|---|---|---|
| 1 | **3** | 4% |
| 2 | **7** | 10% |
| 3 | **22** | 30% |
| 4 | **24** | 33% |
| 5 | **15** | 21% |

1: Strongly Disagree 5: Strongly Agree

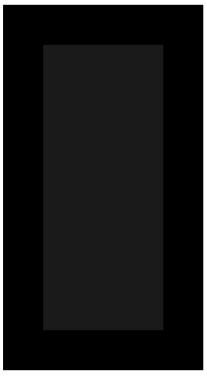
**10. I needed to learn a lot of things before I could get going with this system.**




| | | |
|---|---|---|
| 1 | **34** | 47% |
| 2 | **19** | 26% |
| 3 | **12** | 16% |
| 4 | **5** | 7% |
| 5 | **2** | 3% |

1: Strongly Disagree 5: Strongly Agree

**11. What is your Twitter username?**

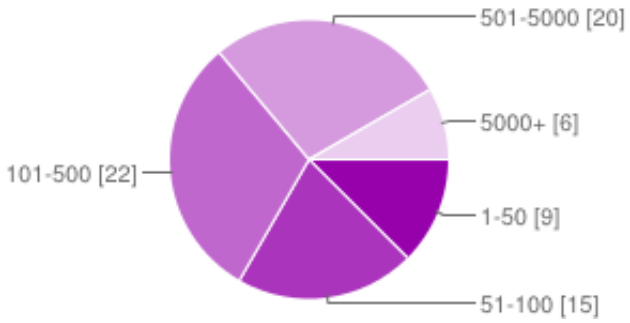


User data anonymized (60 responses)

[redacted]

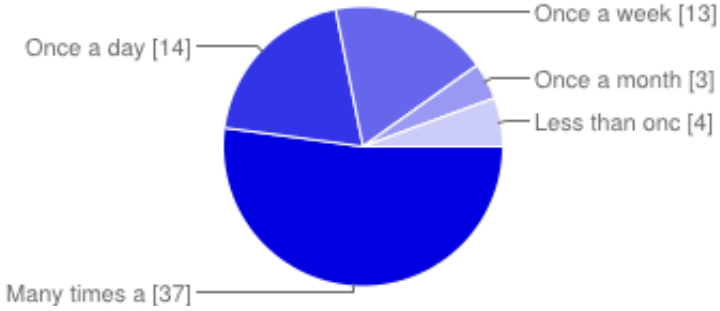
User data anonymized

## 12. How many followers do you have on Twitter?



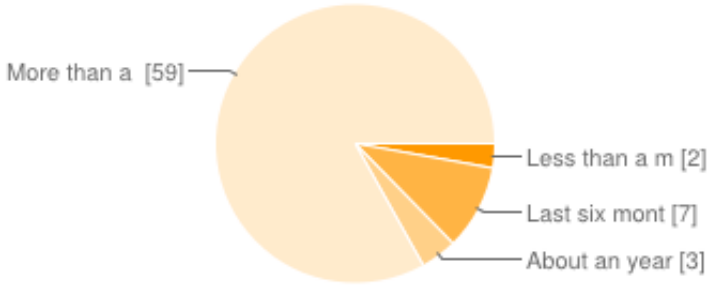

| | | |
|---|---|---|
| 1-50 | **9** | 12% |
| 51-100 | **15** | 21% |
| 101-500 | **22** | 30% |
| 501-5000 | **20** | 27% |
| 5000+ | **6** | 8% |

## 13. How often do you post on Twitter?



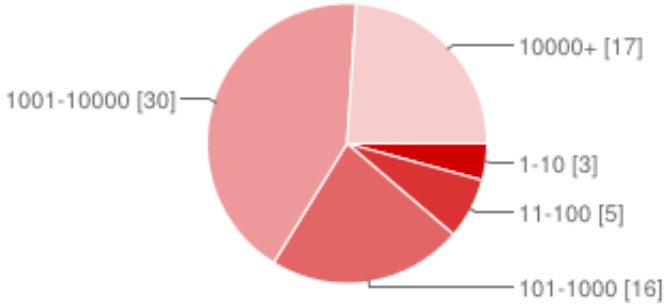

| | | |
|---|---|---|
| Many times a day | **37** | 51% |
| Once a day | **14** | 19% |
| Once a week | **13** | 18% |
| Once a month | **3** | 4% |
| Less than once a month | **4** | 5% |

## 14. When did you create an account on Twitter?



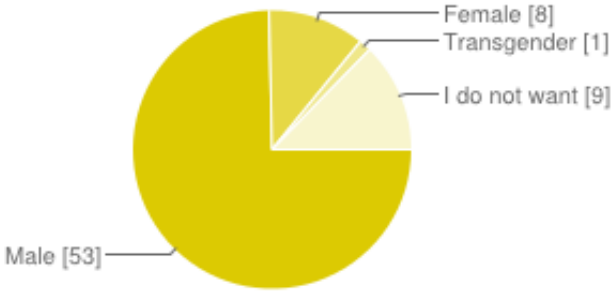

| | | |
|---|---|---|
| Less than a month ago | **2** | 3% |
| Last six months | **7** | 10% |
| About an year ago | **3** | 4% |
| More than a year back | **59** | 81% |

## 15. How many tweets have you posted in all?



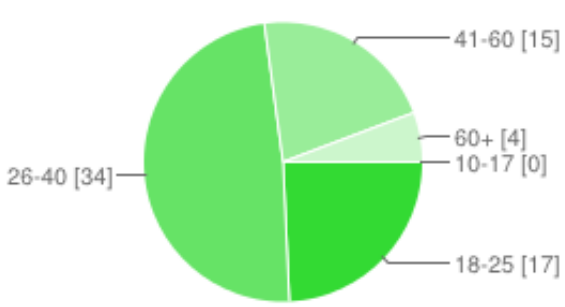

| | | |
|---|---|---|
| 1-10 | **3** | 4% |
| 11-100 | **5** | 7% |
| 101-1000 | **16** | 22% |
| 1001-10000 | **30** | 41% |
| 10000+ | **17** | 23% |

## 16. What is your gender?



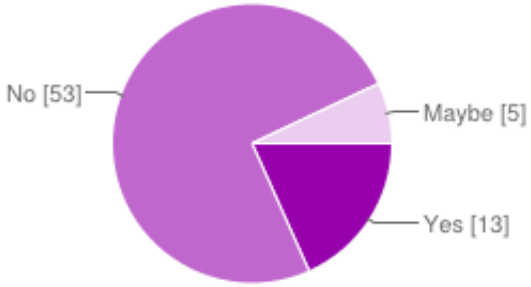
| | | |
|---|---|---|
| Male | **53** | 73% |
| Female | **8** | 11% |
| Transgender | **1** | 1% |
| I do not want to share | **9** | 12% |

### 17. What is your age group?



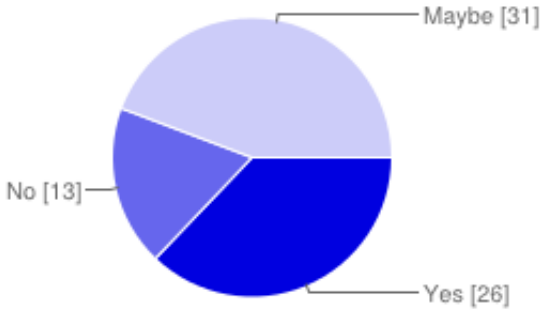

| | | |
|---|---|---|
| 10-17 | **0** | 0% |
| 18-25 | **17** | 23% |
| 26-40 | **34** | 47% |
| 41-60 | **15** | 21% |
| 60+ | **4** | 5% |

### 18. Have you seen / used any other system like (having similar features) TweetCred?




| | | |
|---|---|---|
| Yes | **13** | 18% |
| No | **53** | 73% |
| Maybe | **5** | 7% |

### 19. Will you like to use TweetCred in your daily life?




| | | |
|---|---|---|
| Yes | **26** | 36% |
| No | **13** | 18% |
| Maybe | **31** | 42% |