A survey on Analyzing and Measuring Trustworthiness of User-Generated Content on Twitter during High-Impact Events

Student Name: Aditi Gupta

IIIT-D-PhD-CS-1001

April, 2013

Indraprastha Institute of Information Technology
New Delhi

Evaluation Committee
Dr. Anupam Joshi, UMBC
Dr. Srikanta Bedathur, IIIT-Delhi
Dr. Vikram Goyal, IIIT-Delhi
Dr. Ponnurangam Kumaraguru (PK) (Advisor)

Submitted as part of the requirements
for the Comprehensive Examination for Ph.D. in Computer Science

Abstract

Over the past few years, emergence and advances in the Internet and mobile services have revolutionized the landscape of communication by common people. Online social media provides people with an open platform to share information and opinions on diverse topics. Twitter is a micro-blogging service, which has gained popularity as one of the prominent news source and information dissemination agent over last few years. During real-world impactful events like earthquakes, elections and social movements, we see a sudden rise in activity over the Internet. People log on to Twitter and other social media, to check for updates about these events, and to share information about the events. As this data is generated in real-time by users, who are directly or indirectly linked to the event, mining this data can yield useful knowledge about the event.

The content on Twitter can provide rich information about the event, however, this vast resource of information is often is not credible, unstructured and full of noise. The unmonitored and anonymous nature of online social media makes it difficult to assess the accuracy, credibility or truthfulness of the information. Spread of misinformation, rumors, spams and inflammatory content on Twitter, 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. We present a literature survey of the various research work done around the domain of analyzing and measuring trustworthiness of user-generated content onTwitter to study high-impact events.

Keywords: Twitter, social network analysis, real-world events

Contents

Chapter 1

Introduction

Online Social Media (OSM) has evolved and gained popularity exponentially, over last few years. OSM provides a medium which has a large reach and visibility to people around the world. OSM is a platform, which spans across barriers of country, region, religion, race and language. Social media services has revolutionized the way people access information and news about current events. Unlike traditional news media, online social media such as Twitter are a bidirectional media, in which common people also have a direct platform to share information and their opinions about the news events. Twitter is a micro-blogging 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 (source) who posted the tweet.

People log on to social media websites to check for updates about events and also to share information about the event with others. In such situations, social media content provide a vast resource of unmonitored and unstructured, but rich information about events. Since the data is generated in real time and by users, many of whom are directly or indirectly involved or affected by the actual event; mining this content can yield quite useful knowledge about the ground situation. 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. Presence of noise, spam, advertisements, personal opinions, etc. makes the quality of content on Twitter questionable. Hence, there is dire need to quantify, measure, detect and filter trustworthy content from Twitter. Extracting correct and accurate information is one of the biggest challenges in utilizing information from Twitter. Filtering out spam, phishing content, detecting rumors, extracting credible information, are some of the research problems which address the challenges

of analyzing trustworthiness information on Twitter. In this survey report, we present the various research work done in the field of analyzing and measuring trustworthiness of user-generated content onTwitter to study high-impact events.

1.1 Popularity and Scope Twitter

Few of the popular online social media today are Facebook, Twitter, YouTube, Flickr, MySpace, etc. Each social network has its unique characteristics and properties. As per the 2012 official Twitter statistics, it has more than 500 million registered users, also there were 175 million tweets sent from Twitter every day throughout 2012. [1] As of, Oct 24th, 2012, Facebook touched 1 billion user accounts. [2] Twitter released official numbers that stated that more than 20 million tweets were posted from 27th Oct - Nov 1st, 2012, during the Hurricane Sandy. [3] Similarly, some of the new social media like Pinterest (started in March 2010), is one of the fastest growing social networks, it has acquired more than 25 million active users in last two years. [4]

1.2 Challenges in OSM based Research

In this section, we discuss some of the classical challenges in conducting OSM based research from a computer science researcher's perspective. All of these challenges are also applicable for research based on Twitter.

- *Volume of Content*: 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.
- *Non-Standard Grammar*: 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

[1] http://www.telegraph.co.uk/technology/twitter/9098557/Twitter-to-hit-500-million-registered-users.html
[2] http://online.wsj.com/article/SB10000872396390443635404578036164027386112.html
[3] https://twitter.com/twitter/statuses/264408082958934016
[4] http://news.cnet.com/8301-1023_3-57539742-93/pinterest-pierces-top-50-most-visited-sites-list/

on OSM data. Thus, there is a dire need to construct special and customized toolkits to analyze OSM data.

- *Anonymity*: 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.

- *Real-time Analysis*: 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.

- *Data Accessibility*: Due to privacy and anti scrapping policies of most OSM, data collection for research is a big challenge. Researchers can extract only public data, which ofter represents a very small percentage of the actual content, for e.g. Twitter and Facebook. Though, APIs are provided by most of the social media websites, but rate limits are imposed on data collection via them, for e.g. a maximum of 350 requests per hour on Twitter Rest API.

To overcome the above challenges and extract good quality information from online social networking websites, researchers have analyzed, studied and experimented on OSM data and formulated new techniques and methodologies for the same. The rest of the survey highlights and discusses the research work covering the above.

1.3 Focus of this Survey

The focus of this survey is to discuss the research work done in the space of extracting and analyzing trustworthy and credible information from Twitter during real world events. Over the past few years there has been increase in the usage of Twitter 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 Twitter, not all of the information is of quality with respect to the event. It may be fake, incorrect or noisy. Extracting good quality information is one of the biggest challenges in utilizing information from Twitter. Over last few years, people have highlighted how Twitter can be useful in extracting trustworthy information about real

life events. On the other hand, there have many instances which have highlighted the negative effects of content posted on online social media on the real life events. Figure 1.1 shows some such real instances during the *London riots (2011)* and *Assam riots (2012)*. In both the events, wrong and incorrect information via social media caused chaos and confusion in the real world. The problem definition of this survey is *Analyzing and Measuring Trustworthiness of User-Generated Content on Twitter to study High-Impact Events*. By trustworthiness we mean the truthfulness, credibility or accuracy of the content. By user-generated content we mean the messages and the user profile data on the social media websites. For us high-impact events are any events which affect a lot of people or have large geographical and temporal impact.

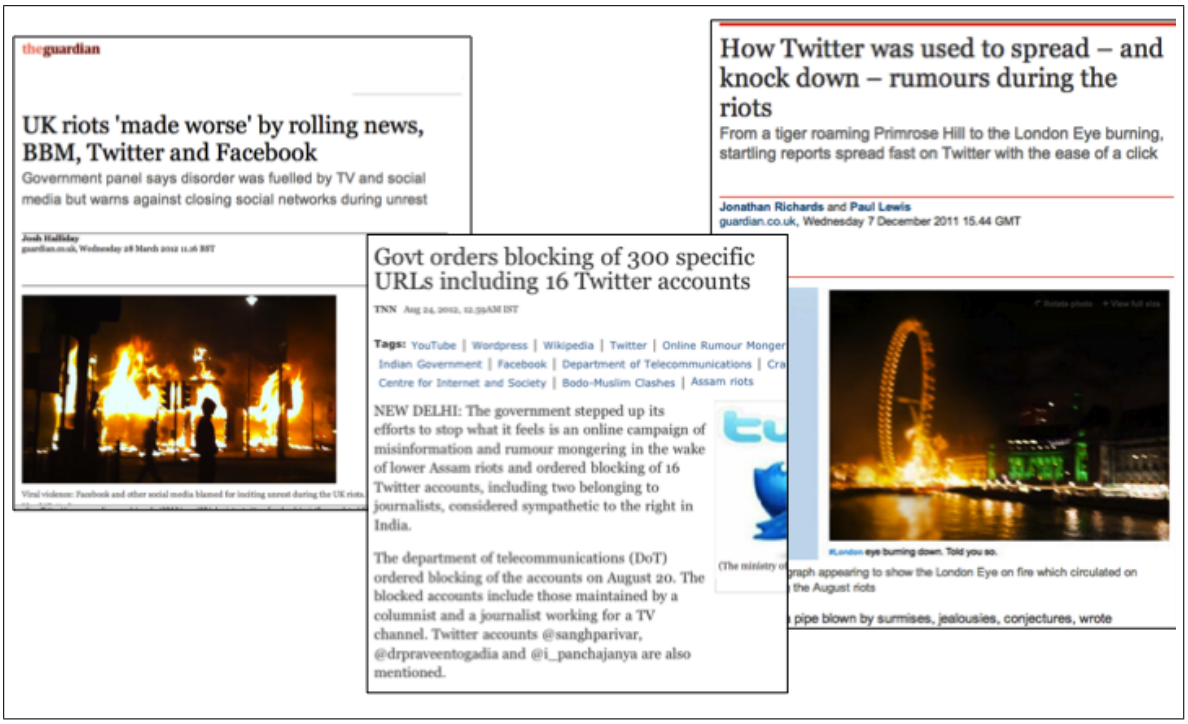


Figure 1.1: Some of the daily newspapers covered certain events which caused harmful effects in real life.

Figure 1.2 gives some sample tweets for *Hurricane Irene, USA, 2011*, that showcase variation in quality of the content related to the same event. All three tweets contain the words matching to the event and were posted while *Hurricane Irene* was the trending topic. The top-left tweet provides correct and credible information about the event. The top-right tweet, is related, but contains no information about the event, it expresses personal opinion of the user. Even though the bottom tweet 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. Thus, the major challenge in extracting credible information or just information on Twitter, is to filter out content which is related to a particular

topic yet provides no useful or actionable information about it.

Figure 1.2: Sample tweets for the event 'Hurricane Irene'. The top-left tweet provides correct and credible information about the event. The top-right tweet, contains personal opinion of the Twitter user. The bottom tweet is a spam tweet.

1.4 Map of Report

In this survey report, we would discuss and present the current state of the art in research for analyzing and measuring trustworthy information from Twitter. Next chapter, discusses the research work to assess, measure, quantify and detect useful and quality content from Twitter; In chapter 3 we would present the research done to characterize the role of Twitter during real world events. In the last chapter we discuss the implications and research gaps in analyzing trustworthy information from Twitter during real-world events.

-

Chapter 2

Analyzing and Measuring Trustworthy Information on Twitter

In this chapter, we discuss some of the research work done over the past few years, to assess, detect, quantify trust on Twitter. Different social media have difference in characteristics, type and quality of information shared through them. Events in real life correspond to a lot of content generated on Twitter, which is opinions, reactions and information from users. One major challenge in consuming content from Twitter as information, is that good quality content needs to be extracted from all the data generated. The quality of content on social media 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 [2] [16] [6]. Ranking algorithms have been applied and fine tuned by researchers for questions pertaining to trust related problems such as credibility and spam [15] [17]. Each of the above mentioned work are discussed in detail, later in the chapter.

2.1 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 [21]. 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) [40]. 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 [22]. The methods proposed act like novel event detection techniques. The study analyzed 900 news events through 2010-11.

2.2 Spam and Phishing Detection on Twitter

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 [7]. 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 [2]. 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 2.1 shows the *PhishAri* system developed by them.

Grier et al. characterized spam spread on Twitter via URLs. They found that 8% of 25 million URLs posted on Twitter point to phishing, malware, and scams listed on popular blacklists [16].

[1] https://bitly.com/

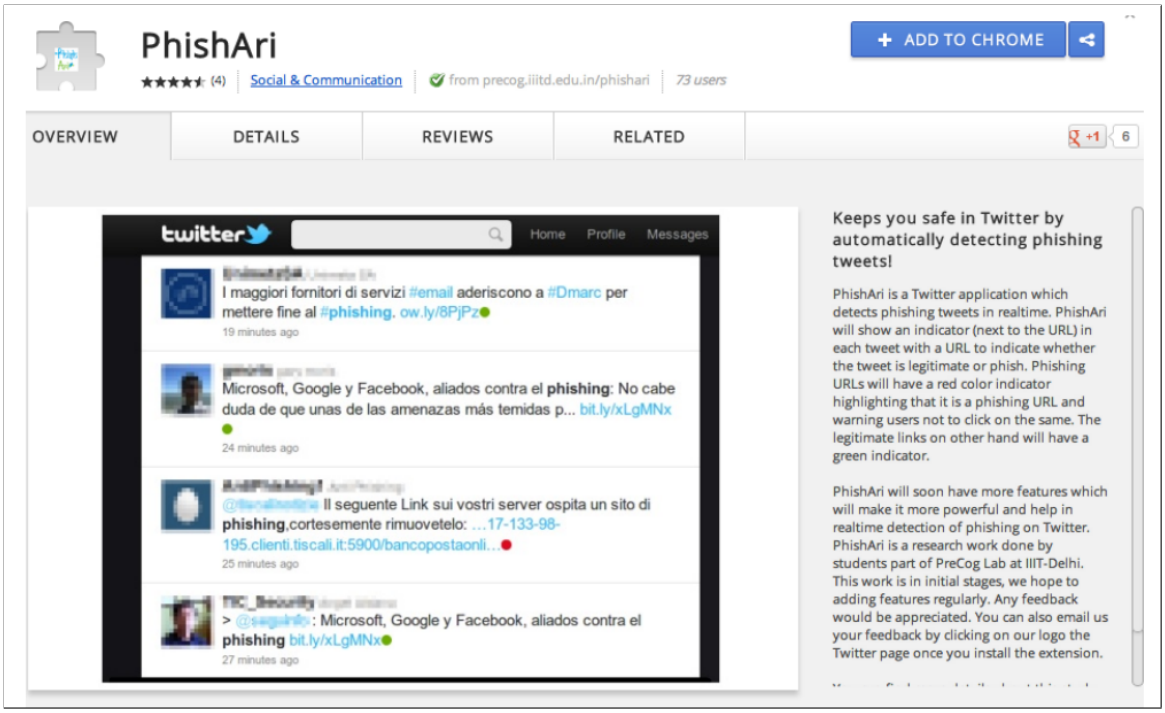


Figure 2.1: A Chrome based browser plugin to detect phishing URLs on Twitter.

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 [15]. 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 [38]. 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 [39]. 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 [3]. 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 [25]. 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.

2.3 Trust / 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 [30]. In their work, they presented certain cases of abusive behavior by Twitter users. Truthy is a live webservice built upon the above work. Figure 2.2 provides a snapshot 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 [6]. 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

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

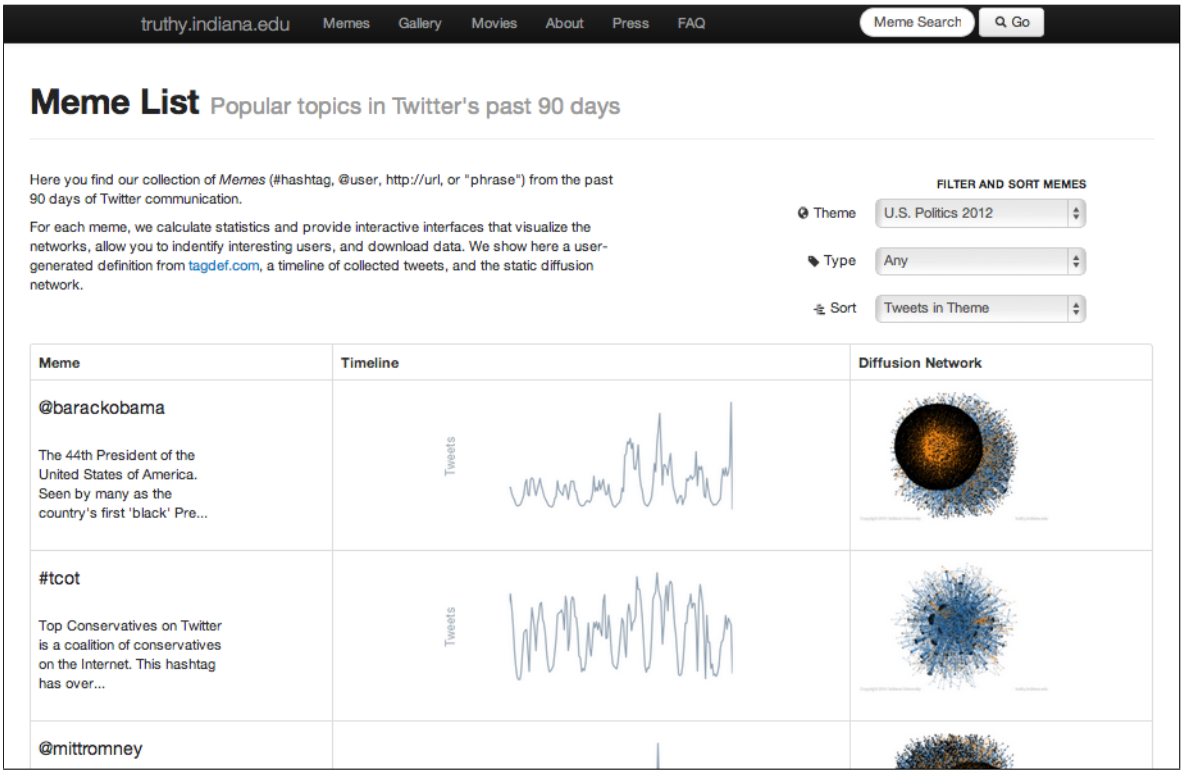


Figure 2.2: A real-time system to study information diffusion on Twitter.

news. The above work 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 [5]. 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. Gupta et al. in their work on analyzing tweets posted during the terrorist bomb blasts in Mumbai (India, 2011), showed that majority of sources of information are unknown and with low Twitter reputation (less number of followers) [18]. This highlights the difficulty in measuring credibility of information and the need to develop automated mechanisms to assess credibility of information on Twitter. The authors in a follow up study applied machine learning algorithms (SVM Rank) and information retrieval techniques (relevance feedback) to assess credibility of content on Twitter [17]. They analyzed fourteen high impact events of 2011; their results showed that on average 30% of total tweets posted about

an event contained situational information about the event while 14% was spam. Only 17% of the total tweets posted about the event contained situational awareness information that was credible. Another, very similar work to the above was done by Xia et al. on tweets generated during the England riots of 2011 [36]. They used a supervised method of Bayesian Network is used to predict the credibility of tweets in emergency situations. 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 [26]. 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 [24]. 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 [14]. Their techniques rely on the wisdom of the Twitter crowds i.e. they used the Twitter Lists feature to identify experts in various topics.

2.4 Hate and Inflammatory Content on OSM

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 [33]. 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 [37]. 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.

2.5 Chapter Summary

In this chapter, we presented the research work done in filtering and cleaning the online social media content. First we discussed, how OSM have emerged a prominent medium for people to obtain information and news. Then, we showed how research problems like, trust / credibility assessment, spam / phishing detection and hate content analysis aid in making information from social media useful and consumable. The spam and phishing research on social media done, over time highlight the dynamic and unique challenges posed. The malicious users adapt and counter the detection techniques, and hence new and dynamically adaptable algorithms need to be build to filter out malicious and unwanted content from OSM. Research based tools like PhishAri and Truthy provide a great way, to separate good quality content from bad, in real-time.

Chapter 3

Role and Analysis of Twitter for Real World Events

The posts and activity on online social media, in particular Twitter, impacts and plays a vital role in various real world events. Role of social media 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 utilized various classical research techniques for the above analysis, like machine learning, information retrieval, social network analysis and image and video analysis. In this section, we discuss some of the prominent research works done to analysis various events on Twitter.

3.1 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 [28]. They viewed people

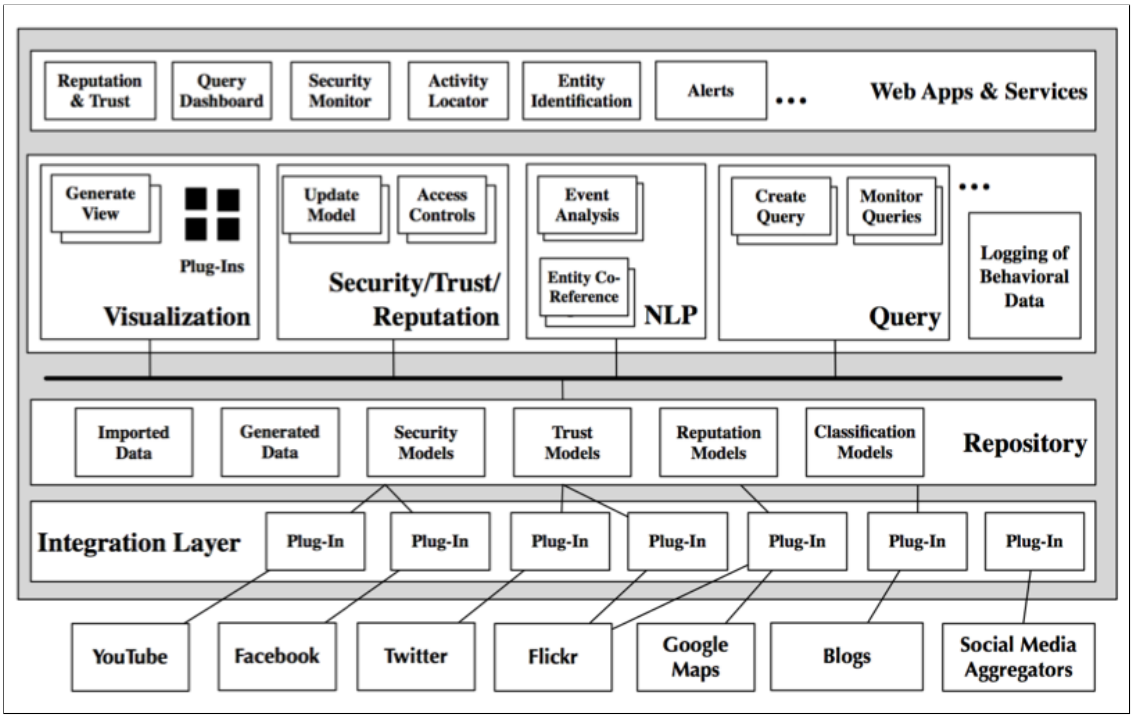


Figure 3.1: The Information Integration Landscape as suggested by Palen et al. [28]

collectively as an important resource that can play a critical role in crisis. In Figure 3.1 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 [29]. 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 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 emergencies management, response and support.

During the earthquakes in Japan in 2008-09, 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

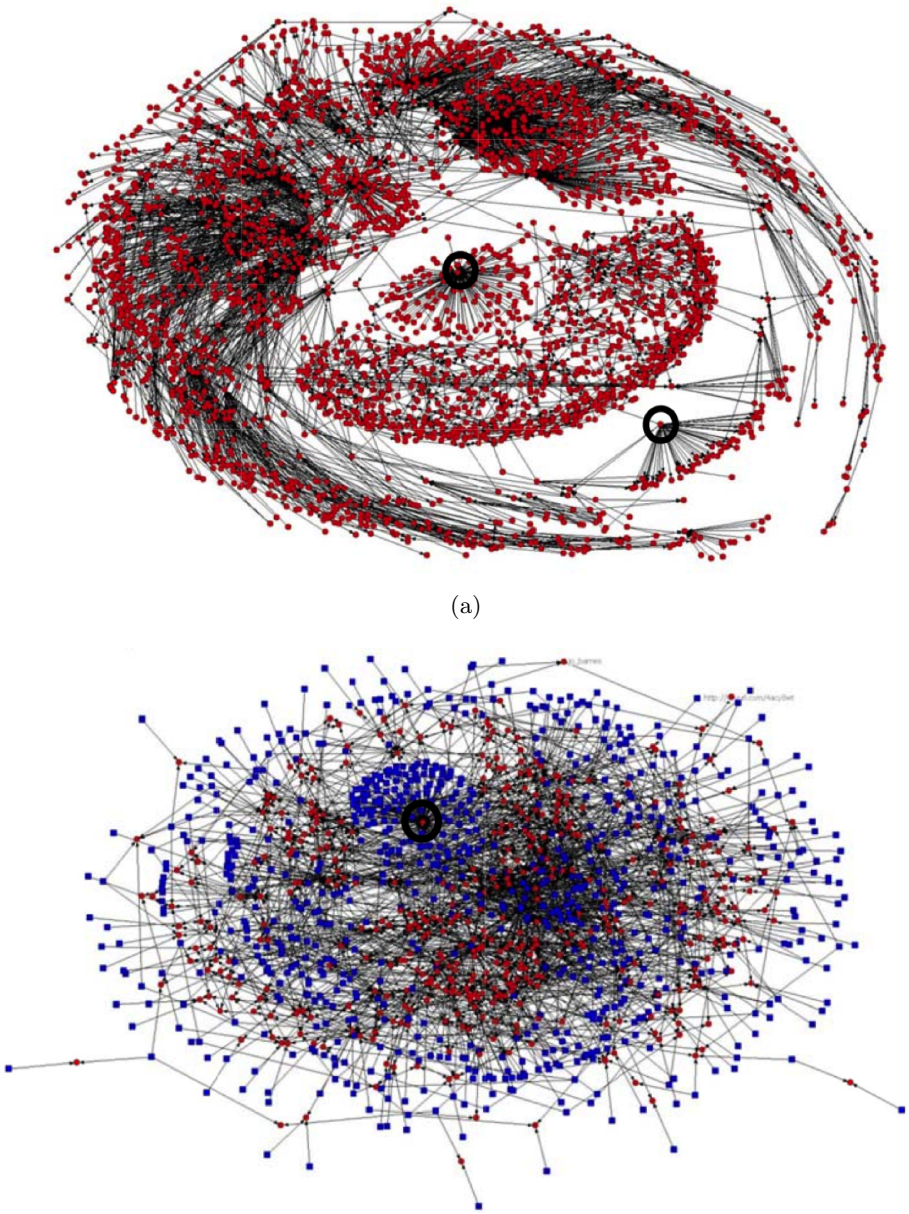
the center and trajectory of an event using Kalman and particle filtering techniques. Based upon the above models, they created an earthquake reporting application for Japan, which detected the earthquake occurrences based on tweets and sent users alert emails [31]. Sakaki et al. in a different research work, analyzed tweet trend to extract the events that happen during a crisis from the Twitter log of user activity analyzed Japanese tweets on all earthquakes during 2010-2011 [32]. 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 [13]. 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. In addition to the users-users graph formed above, they also performed social network analysis on users-resources graphs. 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. Figure 3.2 presents the two kinds of graphs.

3.2 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. [1] 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 [35]. Verma et al. used natural language techniques to build an automated classifier to detect messages on Twitter that may contribute to situational awareness [34]. Another closely related work was done by Oh et al., where they analyzed Twitter stream during the 2008 Mumbai terrorist attacks [27]. Their analysis showed how information

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



(a)

(b)

Figure 3.2: Social network analysis performed by Cheong et al. on Twitter dataset during 2011 Australian floods [13].

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 [8]. 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. Mendoza et al. used the data from 2010 earthquake in Chile to explore the behavior of Twitter users for emergency response activity [23]. 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. 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 [9]. 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. [2] Hughes et al. in their work compared the properties of tweets and users during an emergency to normal situations [1]. 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 dealing with war [4]. 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.

[2] http://leifhanlen.wordpress.com/2011/07/22/crisis-management-using-twitter-and-facebook-for-the-greater-good/

3.3 News Coverage on Role of OSM during Two Events

The role and influence of OSM (including Twitter) over last few years has been curiously and closely tracked by many news reports and articles. In this section we discuss two major events of 2011, which highlighted the strong presence of OSM during events.

3.3.1 England Riots, 2011

We discuss some of the stories and articles corresponding to the positive and negative effects of Twitter and other online social media that got featured in news. An article by Richards et al. instantiated how rumors, such as, *"a tiger roaming Primrose Hill"* and *"the London Eye burning"*, were spread on Twitter during the London riots. [3] They highlighted that the features such as re-tweet on Twitter makes it more susceptible to the spread of rumors. Various articles analyzed the spread of information via Twitter. These articles attempted to analyze whether Twitter had a role in supporting / instigating the riots and helping in the clean-up process. [4] [5] Various news media agencies created spatio-temporal impact and timeline charts to represent the incidents during the UK riots. [6] [7]

3.3.2 Mumbai Blasts, 2011

Even in developing countries like India, with lower levels on Internet penetration, there has been considerable influence of OSM in shaping public reactions, sentiments and efforts of people. During the Mumbai blasts on 13th July, 2011 there was a surge in activity on Twitter and Facebook as quoted by many national level newspaper dailys. [8] A Twitter user, Nitin Sagar [9] created a spreadsheet on Google to coordinate relief operation among people. Within hours hundreds of people registered on the sheet via Twitter. People asked for or offered help on the google spreadsheet 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. Figure 3.3 shows a snapshot of this map created for the Mumbai blasts, 2011.

[3] http://www.guardian.co.uk/uk/2011/dec/07/how-twitter-spread-rumours-riots
[4] http://www.guardian.co.uk/uk/2011/dec/07/twitter-riots-how-news-spread
[5] http://www.bbc.co.uk/news/technology-14442203
[6] http://www.bbc.co.uk/news/uk-14436499
[7] http://www.guardian.co.uk/news/datablog/interactive/2011/aug/09/uk-riots-incident-map
[8] http://articles.timesofindia.indiatimes.com/2011-07-13/mumbai/29768846_1_blast-victims-mumbai-blasts-twitter-and-facebook
[9] @nitinsgr on Twitter

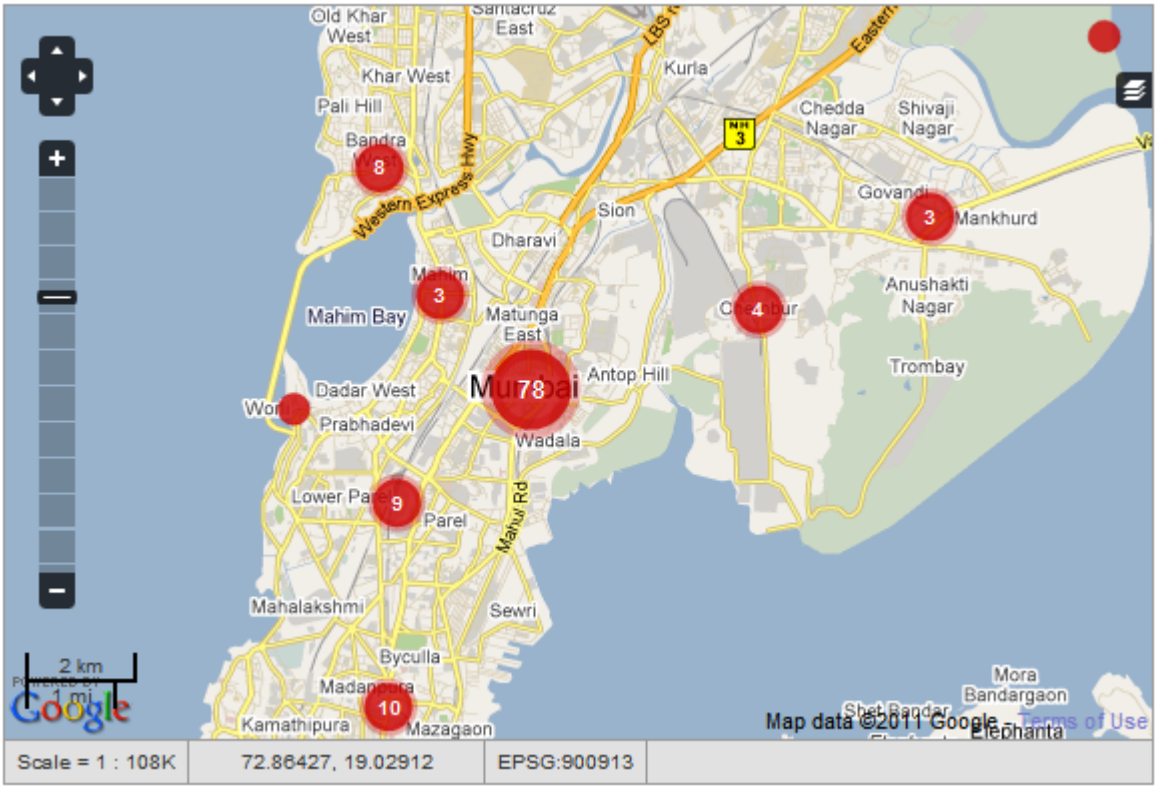


Figure 3.3: A crowd-sourcing based map for crisis management created for the Mumbai blasts. The map showed the blast locations and other related information that people posted social media like Twitter.

3.4 Chapter Summary

Figure 3.4 places some of the events analyzed over last few years on a world map. Machine learning, information retrieval and graph theory are the most frequently applied techniques computational techniques for Twitter based analysis. There have been many research papers which have analyzed the individual real world events, yet there is a need for larger and more generalizable studies. The researchers have highlighted and shown considerable difference in activity and trends of Twitter during particular events like crisis events, hence the existing models and algorithms may not be appropriate for them.

Figure 3.4: Real world events for which online social media were analyzed.

Chapter 4

Discussion and Research gaps

The research work presented in this report 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. Some of our observations from the above study were that, 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. The solutions proposed are naive, and often challenged and counter attacked by offenders. There is dire need, to study closely the content generated on Twitter during real world events, develop metrics to assess its quality, formulate algorithms to filter out the good quality information from the noise or spam generated. Some of the specific open research problems in this domain are:

- Developing self learning and adaptive techniques to detect spam and phishing content on Twitter.
- Using crowd sourced models in extracting or assessing quality of information on Twitter during high impact events.

- Separating bot and automated accounts from legitimate accounts on Twitter, that post content during high impact events.
- Automated event detection, from Twitter, in real-time for corresponding events happening in the real world.

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 [10] [11] [12]. 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 on social media. Similar work, was done for email phishing scams by Kumaraguru et al. [19] [20].

The research work discussed is this survey, 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 adaptable solutions for the same. Through this survey report, we aimed to highlight the current state of the art in analyzing Twitter activity during real world events and highlight research gaps and open problems for the community to solve.

Bibliography

[1] Amanda L. Hughes, L. P. Twitter Adoption and Use in Mass Convergence and Emergency Events. *ISCRAM Conference* (2009).

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

[3] 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.

[4] 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.

[5] 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).

[6] 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.

[7] 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.

[8] 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).

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] Gupta, A., and Kumaraguru, P. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media* (New York, NY, USA, 2012), PSOSM '12, ACM, pp. 2:2–2:8.

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

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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).

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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).

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

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

[40] 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.