Predicting the Veracity of Rumors in Social Networks: Computational Explorations

by Soroush Vosoughi

Ph.D. Thesis Proposal, Media Arts and Sciences
Massachusetts Institute of Technology
November, 2014

_____ Date: _____

Prof. Deb Roy
Associate Professor of Media Arts and Sciences
Massachusetts Institute of Technology

_____ Date: _____

Dr. Allen Gorin
Research Associate
JHU Center of Excellence for Human Language Technology
Visiting Scholar
MIT Laboratory for Social Machines

_____ Date: _____

Prof. Aral Sinan
Associate Professor of Information Technology and Marketing
Massachusetts Institute of Technology

Contents

Predicting the Veracity of Rumors in Social Networks: Computational Explorations

Soroush Vosoughi

November, 2014

Abstract

The spread of malicious or accidental misinformation in social media, especially in time-sensitive situations such as real-world emergencies can have harmful effects on individuals and society. Using computational methods, this thesis investigates the nature of rumors surrounding real-world events on Twitter and Reddit, using the April 2013 Boston Marathon bombings as a case study. With the perspective that in social media both the linguistic and the network dynamics of messages need to be taken into consideration, we propose a set of linguistic and graph-theoretic features that make up the anatomy of rumors. The key idea is that there are measurable differences in the make up of false and true rumors. We extract these features using novel natural language processing and network analytic algorithms that we have developed. In this thesis, we propose a dynamic computational model of rumors composed of these features. The model will be evaluated on the rumors surrounding the August 2014 Ferguson unrest. Once fully evaluated, the model will be used to build a real-time rumor verification system for Twitter and Reddit that can be used during real-world emergencies. This system will have immediate real-world applications for consumers of news, journalists and emergency services and can help minimize and dampen the impact of misinformation.

1 Introduction

In the last decade the Internet has become a major player as a source for news. In fact a study by the Pew Research Center has identified the Internet as the most important resource for the news for people under the age of 30 in the US and the second most important overall after television [5]. More recently, the emergence and rise in popularity of social media and networking services such as Twitter, Facebook and Reddit have greatly affected the news reporting and journalism landscapes. While social media is mostly used for everyday chatter, it is also used to share news and other important information [11, 18]. Now more than ever people turn to social media as their source of news [15, 24, 14], this is especially true for breaking-news, where people crave rapid updates on developing events in real time. As Kwak et al. (2010) have shown, over 85% of all *trending topics*[1] on Twitter are headline or persistent news [14]. Moreover, the ubiquity,

[1]Trending topics are those topics being discussed more than others on Twitter.

accessibility, speed and ease-of-use of social media have made them invaluable sources of first-hand information. Twitter for example has proven to be very useful in emergency situations, particularly for response and recovery [26]. However, the same factors that make social media a great resource for dissemination of breaking-news, combined with the relative lack of oversight of such services, make social media fertile ground for the creation and spread of unsubstantiated and unverified information about events happening in the world.

This unprecedented shift from traditional news media, where there is a clear distinction between journalists and news consumers, to social media, where news is crowd-sourced and anyone can be a reporter, has presented many challenges for various sectors of society, such as journalists, emergency services and news consumers. Journalists now have to compete with millions of people online for breaking-news. Often time this leads journalists to fail to strike a balance between the need to be first and the need to be correct, resulting in an increasing number of traditional news sources reporting unsubstantiated information in the rush to be first [6, 7]. Emergency services have to deal with the consequences and the fallout of rumors and witch-hunts on social media, and finally, news consumers have the incredibly hard task of sifting through posts in order to separate substantiated and trust-worthy posts from rumors and unjustified assumptions. A case in point of this phenomenon is the social media's response to the Boston Marathon bombings. As the events of the bombings unfolded, people turned to social media services like Twitter and Reddit to learn about the situation on the ground as it was happening. Many people tuned into police scanners and posted transcripts of police conversations on these sites. As much as this was a great resource for the people living in the greater Boston, enabling them to stay up-to-date on the situation as it was unfolding, it led to several unfortunate instances of false rumors being spread, and innocent people being implicated in witch-hunts [13, 16, 25]. Another example of such phenomenon is the 2010 earthquake in Chile where rumors propagated in social media created chaos and confusion amongst the news consumers [17].

In this thesis, we plan to develop and combine a set of natural language processing and complex network analysis tools and algorithms that enable the study and analysis of the underlying processes that develop on social media in emergency situations. More generally, we are interested in using social media as an experimental ground for studying and quantifying the nature of communicative discourse in highly connected, complex and massive communication networks (such as social media), in order to better understand and model the dynamic processes that evolve on these networks and the underlying signals driving them. Through modeling these signals and processes we attempt to explain, predict and modify how these systems behave under different conditions. As mentioned, one such behavior that we are interested in modeling is how such systems behave during real-world emergencies (e.g., natural disasters, terrorist attacks, plane crashes, etc). Specifically, we want to model the emergence, evolution, propagation and impact of unverified assertions (or rumors) on social media during emergency situations. We then plan to use these models to predict the veracity of assertions made about such events on social media, with the goal of creating a rumor verification tool for use in emergencies. Finally, we plan to study and experiment with possible approaches for intervening and minimizing the impact and spread of false information in these networks.

2 Background

Although there has been extensive work done on measuring and quantifying information credibility and modeling the spread of information in networks, most have approached this problem either through a text and language processing or network science and complex system analytics framework. The research done in the network science domain have mainly focused on modeling various diffusion and cascade structures [8, 10], the spread of "epidemics" [20, 19, 9], knowledge [8] and information and propaganda [21]. Work has also been done on identifying influential players in spreading information through a network [28, 1] and identifying sources of information [22]. In a work more directly related to our research direction, Mendoza et al, have looked at the difference in propagation behavior of false rumors and true news on Twitter [17]. In all of these cases the properties of the actual entity that is being spread–be it a message, knowledge, or a virus– is never analyzed or taken into consideration in the models. In contrast, our work will be looking at the content of the messages being spread in addition to the propagation behavior of these messages and any information that might be available about the agents involved in the propagation.

Relevant research done in text and language processing domain primarily falls either under information retrieval and comparison or semantic and sentiment analysis. The former involves using various NLP techniques to retrieve relevant information from text (or speech) and then comparing the information against a database of known-facts. The Washington Post's *TruthTeller*[2] which attempts to fact check political speech in real time is a great example of such work. The latter research attempt to detect non-literal text (text that is not supposed to be taken at face value) such as sarcasm [12], satire [3] and hostility (flames) [23] through a combination of semantic and sentiment analytic techniques.

As far as we can tell, there have been very few studies that take all these factors into consideration. Most relevant is the work of Castillo et al [4], where the authors have looked at a combination of linguistics and propagation factors that can be used to approximate users' subjective perceptions of credibility on Twitter (i.e. whether users believe the tweets they are reading), they do not however focus on objective credibility of messages.

3 Thesis Summary and Methodology

This work uses Twitter's and Reddit's response to the April 2013 Boston Marathon bombings as a case study to analyze and model the genesis, evolution and propagation of rumors. The work starts by annotating more than 20 rumors that spread about the events surrounding the bombings, followed by processing and parsing raw tweets and posts using various NLP and network analytic tools which we have developed. Leading to a computational analysis of rumors and predictive models for estimating the veracity of assertions in these mediums and finally evaluating these models on tweets and reddit posts about other real-world events and emergencies, such as the August 2014 unrest in Ferguson.

[2] http://truthteller.washingtonpost.com/

This section will explaining the following:

- The definition of rumors.
- The process through which messages on Twitter and Reddit are operationalized as a collection of computationally measurable and quantifiable features.
- The tools that have been built and need to be built to extract these features.
- The creation of a computational model of rumors using these features.

3.1 Anatomy of an Assertion in Social Media

At any given time an assertion on social media can potentially be broken down into the following parts:

- **The Form/Style**: How is the message presented? Is it well polished? Grammatical? Does it use slang?
- **The Function/Content**: What is the message about? What is it intended to achieve?
- **The Agents/Users**: Who is presenting the message? Which platform is being used? Which social group does the author belong to? What is the history of the author?
- **The Propagation/Cascade Dynamics**: What was the speed at which the message spread? What did the propagation tree look like? How many "influential nodes" did it pass through? How fast did its spread decay?

By breaking down assertions along these dimensions over time, we can create a dynamic fingerprint for each assertion. We can then group false and true assertions together and look for common structural properties between assertions in each group. In addition we can look for possible signals that can differentiate between the false and true assertions. Even though our work focuses on rumors, similar characterization techniques can be used to analyze different phenomenons in social networks.

3.2 Quantifying and Operationalizing Assertions

In the section above we briefly talked about how assertions in social media can be characterized by a combination of their form, function, agents and propagation dynamics. In this section we will explain in greater detail the nature of these four dimensions and describe how they are quantified.

3.2.1 The Form/Style

The form of a message captures how it is presented and is assumed to be independent of its information content. There are many ways to encode the form of a message, however we have found two aspects of the form to carry the most information (and thus be more predictive) about the nature of a signal in social media. These two aspects are the sophistication and the formality of a message. The sophistication of a message is captured through the following features:

- Type/token ratio. (E.g., number of adjectives, etc.)
- Complexity of the sentences. (E.g. embedded clauses, etc.)
- Complexity of words. (E.g., number of syllables, rarity.)

The formality of a message is captured through these features:

- Grammatical correctness of the message.
- Usage of emoticons.
- Usage of Internet slang. (E.g., lol, etc.)

The form of any message can be quantified using the metrics above. Our own preliminary studies have shown all of these metrics to be predictive of how the style of a message is perceived by users (i.e. whether the test users classify the message as well-formed or not). Some of these metrics can be calculated very simply (for example, the number of syllables for a word can be looked up in a dictionary), other metrics, such as the grammatical correctness of a message require more complicated algorithms that we have either developed or are planning to develop.

3.2.2 The Function/Content

The function of a message captures both the information content of the message and the intention behind it (also known as the semantics and pragmatics of a message). We encode the function of a message by looking at its topics, references (such as links to news articles, books, etc), sentiment and speech acts. Below we will explain how these are measured:

- Topics: Instead of looking at each individual word in a message we use hierarchical topic modeling techniques (e.g., HLDA) to categorize the message as a whole.
- References: External material that the message might be referencing. We extract the source and the content of the material.
- Sentiment: What is the attitude conveyed in the message? (E.g., Is it a hateful message or a message of encouragement?)
- Speech acts: Captures what the message is intended to achieve. (E.g., Is it meant to inform or is it merely asking a question?)

As with the form, our preliminary studies have shown all of these metrics to be useful in characterizing messages. All of these metrics require sophisticated algorithms to be properly measured. We have already developed most of these algorithms (such as the speech act and sentiment classifiers) and plan to use off-the-shelve algorithms for the remaining metric (such as topic modeling).

3.2.3 The Agents/Users

Messages in social media do not appear out of thin air and they do not spread by themselves, they require users to create and spread them and they require a medium to be spread through. We try to capture this aspect of a message by looking at the history and profile of the users that create, spread or reply to the message. Specifically, we look at the community, influence, role and reputation of these agents. Below we will explain these factors in more detail:

- Community: The political/social communities that the user identifies with. (E.g., right/left, conspiracy theorists, apolitical, fiscally conservative, socially liberal, etc.)
- Influence: Number of followers/friends. Have they been recognized as an influential member of the community?
- Reputation: Are they known for spreading false information? Are they trusted in the community (e.g., Karma points in reddit)? Are they a controversial/polarizing figure?
- Role: The agents' role in their social network. Are they usually a source of novel messages or do they just repeat others? Do they comment on or modify the messages they forward or just forward them without modification? We also take into account the agents' role in the real-world. (E.g., journalist, news consumer, law enforcement, etc.)

In contrast with the form and the function of a message, not all of the metrics mentioned here can be measured algorithmically. A few of these metrics such as the reputation and role require some level of manual annotation and expert knowledge.

3.2.4 The Propagation/Cascade Dynamics

Unless a message is being observed right at the source, any message will have traversed a path through a network of agents before being observed by an observer. The path that a message takes can tell a lot about the nature of the message. Our preliminary studies have shown that the shape and the speed of a message's spread is predictive of not only the future spread, influence and impact of the signal but also the information content of the message. Below we will explain these factors in more detail:

- Adoption Shape: This captures the tree shape through which a message is adopted by different agents in a network. (For example, is it a wide and shallow tree, or is it a narrow and deep tree?)

- Speed: Propagation rate. What was the rate of adoption of the message over time? When was its peak? How fast did it decay?

Through preliminary studies, we have already shown that the shape and speed of most signals can be grouped into a handful of classes. We are currently working on automatic classification algorithms for this task which will make quantifying the propagation dynamics of a message much easier.

3.3 A Computation Model of Rumors

Once we are able to operationalize assertions in social media along these four dimensions we can create dynamic fingerprints for assertions. We can then employ computational methods to analyze and study these fingerprints to look for signals differentiating false and true assertions. In order to do that we need to first manually annotate several rumors in our data-set to use as ground truth. We need to identify what the rumors are about, when and how they were started and finally when they were "resolved" as true or false. Below we will explain how we do this.

3.3.1 What is a Rumor?

In order to study rumors, we first need to clearly define them so that we can mark their birth and spread in our complex system. A rumor starts from one or more sources and spreads over time from node to node in our system. A rumor can end in three ways: it can be resolved as either true (factual), false (non-factual) or remain unresolved. There are usually several rumors about the same topic, any number of which can be true or false. The resolution of one or more rumors automatically resolves all other rumors about the same topic. For example, take the number of perpetrators in the bombings; there could be several rumors about this topic:

1. *Only one person was responsible for this act.*
2. *This was the work of at least 2 or more people.*
3. *There are only 2 perpetrators.*
4. *It was at least a team of 5 that did this.*

Once rumor number 3 was confirmed as true, it automatically resolved the other rumors as well. (In this case, rumors 1 and 4 resolved to be false and rumor 2 resolved to be true.) Detecting the birth of rumors in our system is relatively straight forward because we can look for the first mention of a rumor and use that as the approximate birth of that rumor. (Approximate, because we cannot be sure that it was truly the first mention.) In our dataset, rumors almost always emerge from Reddit and Twitter.

Next we need to have a criterion for the resolution of rumors. When are rumors about a particular topic considered resolved? For now, we consider rumors to be officially resolved once Wikipedia has closed the case on the topic. This is a very conservative estimate as Wikipedia editors usually require multiple confirmations from government and news agencies before they consider a case closed. Using this procedure, we manually annotated 24 rumors about the Boston Marathon bombings.

3.3.2 Predictive Model and Real Time Rumor Verification

Now that we have an official resolver of rumors, we would like to devise an algorithm that can predict the veracity of rumors before they are officially resolved by Wikipedia. We can operationalize our rumors using the features discussed earlier. By doing so, we can create a computational fingerprint for rumors. Our hypothesis is that over time, the fingerprints of true and false rumors about a particular topic will exhibit subtle but significantly different dynamic structures that can be detected algorithmically. Importantly, we believe that on average this signal will be detectable before the rumors are officially resolved and thus allowing us to verify or reject a rumor before official resolution.

The end goal is to create a rumor verification system that can analyze and predict the veracity of rumors about emergency events in real-time as they unfold. We can use our analysis of rumors surrounding the Boston Marathon bombings to create reference fingerprints for false and true rumors. In a new emergency, the system would track rumors as they spread in real-time and dynamically update their fingerprints. At every step the new fingerprints are compared with our reference false and true fingerprints and given a likelihood of match along with a confidence score. This system will be semi-supervised as it requires a human to tune it to real-world events.

The real-time rumor verification tool will be a website that allows users to tune into an event through searching for terms describing the event. (E.g., the terms "Boston", "marathon" and "bomb" for the Boston Marathon bombings). The tool then uses standard query expansion techniques [27] to capture more terms relevant to the event in order to increase the recall of the search. Once the tool is fully tuned into an event, it uses unsupervised topic modeling and text clustering algorithms such as Latent Dirichlet allocation (LDA) [2] to group the conversation around the event into different topics and sub-topics. (E.g., one topic could be "the number of bombs that exploded", or "the Saudi student was responsible for the bombings", etc). The tool then uses our predictive model to estimate the veracity of each of these claims which is then displayed to the user on a "Truthiness Meter". The website will also provide the users with an option to disagree or agree with the algorithm. This feedback is used in future training of the system. There will also be a section for users to discuss the rumors amongst themselves. A preliminary sketch of the website is shown in the appendix.

4 Evaluation

In its most general form, this thesis presents a computational model of assertions (including rumors) about real-world emergencies in social media, specifically Twitter and Reddit. The Boston Marathon bombings is the real-world event on which this analysis and modeling is performed. One of the primary ways in which this model can be evaluated is by testing it on a completely different data-set about another real-world event. We have decided to use the August 2014 Ferguson unrest[3] following the police shooting of Michael Brown as our test data-set.

The evaluation will consist of first manually annotating the rumors around the Ferguson unrest using the same technique as was used for the Boston Marathon bombings rumors. This

[3] http://en.wikipedia.org/wiki/2014_Ferguson_unrest

entails identifying the major rumors about the event and figuring out the time-line of the rumors, beginning from the origin all the way until the rumors are "officially" resolved as false or true. Next we run our system on the tweets and posts about the Ferguson event as if it was happening in real-time. Our system will be quantifying the messages using the techniques explained earlier and create a dynamic fingerprint for each rumor. These fingerprints are not static and will change as new messages about the rumors are fed into the system and processed. At each time step, our system feeds these fingerprints into our model, which was trained on the Boston Marathon bombings data-set to get a prediction (along with a confidence score) of the veracity of the rumor. Once the confidence of our model is above a certain threshold we consider the rumor resolved. We then compare the prediction of our model for each rumor to the manually annotated ground truth and measure the accuracy of the model.

All the algorithms developed for feature extraction are evaluated using standard cross-fold validation. The ground truth for these evaluations are mostly collected through *Amazon Turk's*[4] crowd-sourced annotation services.

5 Conclusion

This thesis investigates the nature of rumors surrounding real-world events on Twitter and Reddit, using these service's response to the Boston Marathon bombings as a reference. The work spans both the technical algorithms needed to deconstruct messages into their computational bits and the creation of a computational model of rumors. The focus of study is the understanding of the anatomy of rumors and what the anatomy reveals about their veracity. With the perspective that in social media, both the linguistic and the network dynamics of messages need to be taken into consideration, we propose a set of 16 features, divided into 4 categories: Function, Form, Agents, and Propagation Dynamics. We then propose a computational model of rumors composed of these 16 features. We plan to train this model on manually annotated examples of false and true rumors and then evaluate the model by running it on the rumors surrounding the August 2014 Ferguson unrest.

The expected contributions of this work fall into several categories. The operationalization of messages in social media through linguistic and graph-theoretic features. Methodology for annotating rumors in social media. The algorithms developed for analyzing and extracting interesting features from messages which contribute to the body of tools in NLP and network analytics. And finally, by creating a computational model of rumors, we can build a real-time rumor verification system for Twitter and Reddit that can be used during real-world emergencies such as earthquakes, fires, riots and terrorist attacks. This system will have immediate real-world applications for consumers of news, journalists and emergency services.

[4] https://www.mturk.com

6 Research Plan

6.1 Completed work

The dataset needed for this project has already been collected and mostly annotated and a number of the natural language and network analytic tools and algorithms, such as sentiment, speech act, formality and sophistication classifiers have already been implemented and evaluated.

6.2 Timeline

Early November	Submit thesis proposal.
November 2014	Finish work on semantic analysis algorithms.
Late November	Defend thesis proposal.
December 2014	Work on network analytics algorithms.
Early January 2015	ICWSM2015 deadline; a good deadline for publishing some of the algorithms.
January 2015	Finish final computational model.
February 2015	Finish evaluation of model on Ferguson data-set. Finish thesis document outline.
March 2015	Finish thesis draft.
April 2015	Thesis defense.
Late April - May 2015	Submit final thesis document.

6.3 Required resources

The main resource required to complete this thesis is access to Twitter, Reddit and Wikipedia posts made about the Boston Marathon bombings, starting from the day of the event until two weeks after (April 15, 2013 to April 29, 2013), and the Ferguson unrest, from the day of the shooting of Michael Brown on August 9, 2014 to the end of the unrest around August 23, 2014. We already have collected and annotated the Boston Marathon bombings data-set in the past year and we are in the process of doing the same for Ferguson data-set.

7 Author Biography

Soroush Vosoughi is a Ph.D. candidate under Professor Deb Roy in the Laboratory for Social Machines at the MIT Media Lab where he is studying and modeling the behavior of highly connected, complex and massive communication systems such as social networks. His main interests lie at the intersection of natural language processing, machine learning, complex systems and computational cognitive science. In the past as a member of the Cognitive Machines group at the MIT Media Lab, he has worked on the language, vision and planning subsystems of an interactive physical robot that learns to communicate in human-like ways and has been involved in studying and modeling the underlying processes involved in child language acquisition by analyzing behavior in massive audio-video corpora. Soroush received his M.Sc working in the

Cognitive Machines group and his Sc.B. in computer science from the Massachusetts Institute of Technology.

8 Committee Biographies

Allen Gorin is a visiting scholar at the MIT Laboratory for Social Machines, and also a research associate at the JHU Center of Excellence for Human Language Technology. He retired in 2012 as Director, Human Language Technology (HLT) Research from the U.S. Department of Defense at Fort Meade, where he focused on creating HLT technologies for coping with information overload. Before that, he was at AT&T Labs, leading the research team that created AT&T's "How May I Help You?" natural language voice service, which was deployed nationally in 2001. He is a Fellow of the IEEE, was awarded the AT&T Science and Technology Medal and the AT&T Strategic Patent Award. He has published 109 papers, and has been granted 44 U.S. Patents. He received the B.S. and M.A. degrees in Mathematics from SUNY at Stony Brook, and the Ph.D. in Mathematics from the CUNY Graduate Center.

Deb Roy is the Director of the Laboratory for Social Machines, Associate Professor at MIT, and Chief Media Scientist of Twitter. He leads research at MIT at the intersection of human and machine communication, advises technology start-up companies, and serves on the World Economic Forums Global Agenda Council on Social Media. He was co-founder and CEO of Bluefin Labs, a social TV analytics company, which MIT Technology Review named as one of the 50 most innovative companies of 2012. Bluefin was acquired by Twitter in 2013, Twitters largest acquisition at the time. An author of over 100 academic papers in machine learning, cognitive modeling, and human-machine interaction, his TED talk, Birth of a Word, has been viewed over 3 million times. A native of Canada, Roy received a Bachelor of Applied Science (computer engineering) from the University of Waterloo and a PhD in Media Arts and Sciences from MIT.

Sinan Aral is the David Austin Professor of Management and an Associate Professor of Information Technology and Marketing at the MIT Sloan School of Management. His research focuses on social contagion, product virality and measuring, and managing how information diffusion in massive social networks such as Twitter and Facebook affects information worker productivity, consumer demand, and viral marketing. This research has won numerous awards including the Microsoft Faculty Fellowship (2010), the PopTech Science and Public Leaders Fellowship (2010), an NSF Early Career Development (CAREER) Award (2009), the Best Overall Paper Award at the International Conference on Information Systems (ICIS) (in both 2006 and 2008), the ICIS Best Paper in IT Economics Award (2006), the ICIS Best Paper in IT Business Value Research Award (2006), the ACM SIGMIS Best Dissertation Award (2007), and the IBM Faculty Award (2009). He has worked closely with Facebook, Yahoo, Microsoft, IBM, Cisco, Intel, the New York Times, Oracle, SAP, and many other leading Fortune 500 firms on realizing business value from social media and information technology investments. Sinan has been a Fulbright Scholar, has served as chief scientist and on the board of directors of SocialAmp, a social commerce company that enables targeting and peer referral in social media networks

(which was sold to Merkle in January, 2012). He is currently chief scientist of Humin and an organizer of the Workshop on Information in Networks (WIN): http://www.winworkshop.net. He is a frequent speaker at such thought leading events as TEDxSiliconValley, TEDxColumbia Engineering, TEDxNYU, Wireds Nextwork, and PopTech. He has been the keynote speaker at executive gatherings such as Omnicoms Global Emerge Summit. His work has been published in leading journals such as the American Journal of Sociology, Information Systems Research, Management Science, Marketing Science, Nature, the Proceedings of the National Academy of Sciences (PNAS), Science, Organization Science, the Harvard Business Review, and the Sloan Management Review. Hiw work is often featured in popular press outlets such as the Economist, the New York Times, Businessweek, Wired, Fast Company, and CIO Magazine. Sinan is a Phi Beta Kappa graduate of Northwestern University. He holds an MSc from the London School of Economics and an MPP from Harvard University, and received his PhD from the MIT Sloan School of Management.

References

[1] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74. ACM, 2011.

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

[3] Clint Burfoot and Timothy Baldwin. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 161–164. Association for Computational Linguistics, 2009.

[4] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.

[5] The Pew Research Center. Internet overtakes newspapers as news outlet, December 2008. http://pewresearch.org/pubs/1066/internet-overtakes-newspapers-as-news-source[pewresearch.org; posted 23-December-2008].

[6] The Pew Research Center. Public evaluations of the news media: 1985-2009. press accuracy rating hits two decade low, September 2009. http://www.people-press.org/2009/09/13/press-accuracy-rating-hits-two-decade-low/[people-press.org; posted 13-September-2009].

[7] The Pew Research Center. Further decline in credibility ratings for most news organizations, August 2012. http://www.people-press.org/2012/08/16/further-decline-in-credibility-ratings-for-most-news-organizations/[people-press.org; posted 16-August-2012].

[8] Robin Cowan and Nicolas Jonard. Network structure and the diffusion of knowledge. *Journal of economic Dynamics and Control*, 28(8):1557–1575, 2004.

[9] Ayalvadi Ganesh, Laurent Massoulié, and Don Towsley. The effect of network topology on the spread of epidemics. In *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, volume 2, pages 1455–1466. IEEE, 2005.

[10] Sharad Goel, Duncan J Watts, and Daniel G Goldstein. The structure of online diffusion networks. In *Proceedings of the 13th ACM conference on electronic commerce*, pages 623–638. ACM, 2012.

[11] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[12] Roger J Kreuz and Gina M Caucci. Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on computational approaches to Figurative Language*, pages 1–4. Association for Computational Linguistics, 2007.

[13] Lalit Kundani. "when the tail wags the dog: Dangers of crowdsourcing justice", July 2013. http://newamericamedia.org/2013/07/when-the-tail-wags-the-dog-dangers-of-crowdsourcing-justice.php/[newamericamedia.org; posted 27-July-2013].

[14] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[15] Sam Laird. "how social media is taking over the news industry", April 2012. http://mashable.com/2012/04/18/social-media-and-the-news/[mashable.com; posted 18-April-2012].

[16] Dave Lee. "boston bombing: How internet detectives got it very wrong", April 2013. http://www.bbc.com/news/technology-22214511/[bbc.com; posted 19-April-2013].

[17] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we rt? In *Proceedings of the first workshop on social media analytics*, pages 71–79. ACM, 2010.

[18] Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 189–192. ACM, 2010.

[19] Mark EJ Newman. Spread of epidemic disease on networks. *Physical review E*, 66(1):016128, 2002.

[20] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

[21] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Detecting and tracking the spread of astroturf memes in microblog streams. *arXiv preprint arXiv:1011.3768*, 2010.

[22] Devavrat Shah and Tauhid Zaman. Rumors in a network: Who's the culprit? *Information Theory, IEEE Transactions on*, 57(8):5163–5181, 2011.

[23] Ellen Spertus. Smokey: Automatic recognition of hostile messages. In *AAAI/IAAI*, pages 1058–1065, 1997.

[24] Wilma Stassen. Your news in 140 characters: exploring the role of social media in journalism. *Global Media Journal-African Edition*, 4(1):116–131, 2010.

[25] Manuel Valdes. "innocents accused in online manhunt", April 2013. http://www.3news.co.nz/Innocents-accused-in-online-manhunt/tabid/412/articleID/295143/Default.aspx/[3news.co.nz; posted 22-April-2013].

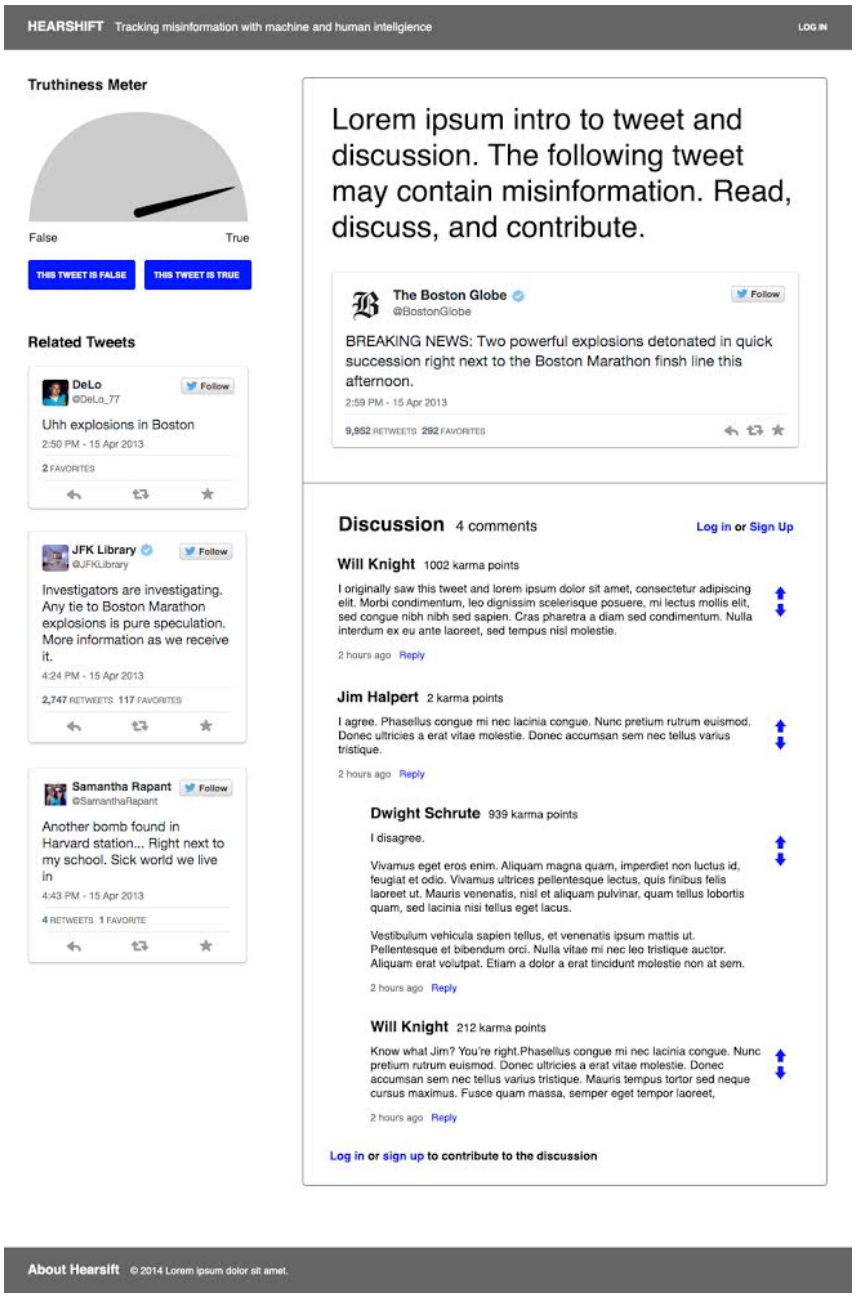
[26] Sarah Vieweg. Microblogged contributions to the emergency arena: Discovery, interpretation and implications. *Computer Supported Collaborative Work*, pages 515–516, 2010.

[27] Ellen M Voorhees. Query expansion using lexical-semantic relations. In *SIGIR94*, pages 61–69. Springer, 1994.

[28] Duncan J Watts and Peter Sheridan Dodds. Influentials, networks, and public opinion formation. *Journal of consumer research*, 34(4):441–458, 2007.

Appendices

A Sketch of the real-time rumor verification tool.



Addendum: Response to the Reviewers

Thanks to all the reviewers for their thoughtful comments. I have identified several key points from the reviews that I wish to address here:

- The few typos in the documents have been fixed and the document has been proofread once more.
- I have cited two references (6 and 7) to substantiate the claim about the role of technology on the veracity of reporting.
- One reviewer wanted me to front-load the goals of the project. I have made it clear in the introduction and the abstract that I am modeling the linguistic and diffusion characteristics of rumors. I am then using these models to predict the veracity of rumors. I mention this in the abstract: "we propose a set of linguistic and graph-theoretic features that make up the anatomy of rumors. The key idea is that there are measurable differences in the make up of false and true rumor.". I also talk about this in the last paragraph of the introduction: "Specifically, we want to model the emergence, evolution, propagation and impact of unverified assertions (or rumors) on social media during emergency situations. We then plan to use these models to predict the veracity of assertions made about such events on social media, with the goal of creating a rumor verification tool for use in emergencies."
- I have included more details about the real-time verification tool in section 3.3.2 of the proposal as was requested. I have also included a rough sketch of what the tool's interface would look like in the appendix.
- One reviewer raised the point that the "Boston Marathon bombings" and the "Ferguson unrests" data-sets are quite different from each other. I agree that this is indeed the case. I intentionally wanted to look at real-world emergencies that were not too similar, in order to show that our models can be generalized to different types of events. However, the reviewer's point about looking at a more similar event as a first stage evaluation makes great sense. Though I do not currently have such a data-set I will definitely be searching for one. (Maybe the reviewer has one in mind?)
- It was pointed out that the difference between rumors associated with a given context can be in kind (content) or degree (emphasis). This is certainly true and an important distinction. This thesis focuses on measuring the difference in kind. Looking at the different in degree of rumors is an interesting idea but one that falls somewhat outside the scope of this thesis. However I have given this some thought and if time permits, plan to develop an algorithm for predicting the "impact" a rumor might have so that intervention strategies can focus on dampening the effects of high impact rumors.
- As it was correctly observed in the reviews, our system treats "truth" as a binary variable. A claim is defined to be true if it was confirmed by multiple trusted sources (this is

discussed in section 3.3.1) and false if it was denied by the those sources. (If a rumors was neither confirmed nor denied, it is left as unresolved.) I agree with the reviewer's premise that human nature is far from binary and that false rumors may turn into self-fulfilling prophecies. Even though this is outside the scope of this thesis, I have given this problem some thought and have plans for future work to address these issues. If time permits, I would like to develop another measure for rumors to estimate the "justifiability" of rumors in addition to the veracity. As the name suggests, this justifiability metric would capture how justifiable a claim is at the time that it is made. So a false rumor that turns into a true fact would still have low justifiability even though its veracity would change. The differentiation between intentional and unintentional falsification is also an important one which is again beyond the scope of this work but is a natural extension of this work.

- Finally, one reviewer's comment about the overlap of rumor propagation and spatial event space propagation is a very interesting one. Specifically,the suggestion by the reviewer to map between the rumor space and the physical space to predict upcoming locations of the near future emergency spaces is very intriguing and one that I had not considered. This idea would have immediate applications for law enforcement agencies. One possible obstacle to looking at spatial event space is the relative lack of location information in our data-set. Only a very small percent of tweets are geo-tagged and there is no location information for Reddit posts. Though it is possible to extract location information from text using named-entity and event extraction algorithms, this currently falls outside the scope of this thesis. However, this idea is extremely powerful and I would be very happy to discuss this further with the reviewer offline. We might be able to find possible ways to extend this thesis in that direction once the current proposed work has been finished, defended and hopefully published!