

University of California
Santa Barbara

Information Reliability on the Social Web
Models and Applications in Intelligent User Interfaces

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Computer Science

by

Byungkyu Kang

Committee in charge:

Professor Tobias Höllerer, Chair
Professor Matthew Turk
Professor Xifeng Yan
Dr. John O'Donovan

March 2016

The Dissertation of Byungkyu Kang is approved.

Professor Matthew Turk

Professor Xifeng Yan

Dr. John O'Donovan

Professor Tobias Höllerer, Committee Chair

March 2016

Information Reliability on the Social Web
Models and Applications in Intelligent User Interfaces

Copyright © 2016

by

Byungkyu Kang

To my parents and my dearest wife Claudia Koeun Choi!

Acknowledgements

I would like to acknowledge that, during my work on this thesis, I received financial support from grant No. W911NF-09-2-0053 of the U.S. Army Research Laboratory and grant No. 1058132 of the NSF grant IIS.

I realize that I have been extremely lucky to meet really great people and to be able to learn invaluable knowledge and wisdom. I am truly grateful for all the care and support I received from my advisor Tobias Höllerer and my research mentor John O'Donovan during this thesis. I also thank to my committee members, Matthew Turk and Xifeng Yan, for their helpful and detailed feedback on my thesis work. During my PhD years, I also received invaluable advice from George Legrady on my media art studies at UCSB, from Sibel Adalton my credibility and trust studies, and from great mentors and other researchers/students at Yahoo! Labs, Adobe research and KIST on my summer research intern projects. I also thank to all my beloved friends who have cheered me up and listened to all the anxiety and anguish of my journey including Seokwon Choi, Tom Depasquale, Teresiana Matarrese, Miguel Lastras, Anylu Perez Tapia, Eduardo Graells, Luca Chiarandini, Seojin Ko, Sangho Oh, and Hyunchul Park!

This thesis work would not have been possible without my beloved parents. I thank to my family for being with me even from before the beginning and sometimes giving everything they have and more. Finally, My beloved wife Claudia Koeun Choi! I need no calculation to say that she has given me the best part of those days thank you and love you.

Curriculum Vitæ

Byungkyu Kang

Education

- 2016 Ph.D. in Computer Science (Expected), University of California, Santa Barbara.
- 2012 M.A. in Computer Science, University of California, Santa Barbara.
- 2007 B.Eng. in Electrical Electronics Engineering, Chung-ang University

Publications

Papers Under Review

- **Byungkyu Kang**, Nava Tintarev, Tobias Höllerer and John O’Donovan, “What am I not seeing? An Interactive Approach to Social Content Discovery in Microblogs,” 10th ACM Conference on Recommender Systems (RecSys2016)
- **Byungkyu Kang**, Haleigh Wright, Tobias Höllerer, Ambuj Singh and John O’Donovan, “Through The Grapevine: A Comparison of News in Microblogs and Traditional Media,” Lecture Notes in Social Networks, Berlin, Springer-Verlag (2016)
- **Byungkyu Kang**, Tobias Höllerer and John O’Donovan, “Evaluating personal metadata and credibility perception in social media streams,” ACM Transactions on Interactive Intelligent Systems
- Ilaria Bordino, Olivier Van Laere, Yelena Mejova, **Byungkyu Kang** and Mounia Lalmas, “Beyond Entities: Promoting Exploratory Search with Bundles,” Information Retrieval, Kluwer Academic Publishers

Peer Reviewed Papers

- Nava Tintarev, **Byungkyu Kang**, Tobias Höllerer and John O’Donovan. “Inspection Mechanisms for Community-based Content Discovery in Microblogs,” Joint Workshop on Interfaces and Human Decision Making in Recommender Systems (in-tRS) held in conjunction with the 9th ACM Conference on Recommender Systems, Vienna, Austria, September 2015.
- **Byungkyu Kang**, Tobias Höllerer and John O’Donovan. “The Full Story: Automatic detection of unique news content in Microblogs,” Proceedings of the 2015 IEEE/ACM International Symposium on Foundations and Applications of Big Data Analytics (FAB) held in conjunction with the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Paris, France, August 2015.

- **Byungkyu Kang**, Tobias Höllerer and John O’Donovan. “Believe it or Not? Analyzing Information Credibility in Microblogs”, Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Paris, France, August 2015.
- Sujoy Sikdar, Sibel Adalı, Md Tanvir Amin, Tarek Abdelzaher, Kevin Chan, Jin-Hee Cho, **Byungkyu Kang**, John O’Donovan. “Finding True and Credible Information on Twitter,” 17th International Conference of Information Fusion (IEEE FUSION), Salamanca, Spain, July 2014.
- John O’Donovan, **Byungkyu Kang**. “Competence Modeling in Twitter: Mapping Theory to Practice,” International Conference on Social Computing (SocialCom 2014). Palo Alto, California., USA. May 27th-29th 2014
- **Byungkyu Kang**, George Legrady and Tobias Höllerer. “TweetProbe: A Real-Time Microblog Stream Visualization Framework,” IEEE Visualization (VIS), Arts Program (VISAP), Atlanta, Georgia, October 2013.
- Sujoy Sikdar, **Byungkyu Kang**, John O’Donovan, Tobias Höllerer, Sibel Adalı, “Understanding Information Credibility on Twitter,” IEEE/ASE SocialCom, Washington, DC, USA, 2013 (**Best paper**)
- Sujoy Sikdar, **Byungkyu Kang**, John O’Donovan, Tobias Höllerer, Sibel Adalı, “Cutting Through the Noise: Defining Ground Truth in Information Credibility on Twitter,” ASE HUMAN JOURNAL 2.1, 2013
- James Schaffer, **Byungkyu Kang**, Tobias Höllerer, Hengchang Liu and John O’Donovan, “Interactive Interfaces for Complex Network Analysis: A QoI Perspective,” IEEE International Conference on Pervasive Computing and Communications (PERCOM), 2013
- John O’Donovan, **Byungkyu Kang**, Greg Meyer, Tobias Höllerer and Sibel Adalı, “Credibility in Context: An Analysis of Feature Distributions in Twitter,” IEEE SocialCom, Amsterdam, Netherlands, 2012
- **Byungkyu Kang**, John O’Donovan and Tobias Höllerer, “Modeling Topic Specific Credibility in Twitter,” International Conference on Intelligent User Interfaces(IUI), Lisbon, Portugal, 2012
- **Byungkyu Kang**, John O’Donovan and Tobias Höllerer, “A Framework for Modeling Trust in Collaborative Ontologies,” Proceedings of the sixth Graduate Student Workshop on Computing, 2011, UC Santa Barbara, 39-40.
- **Byungkyu Kang**, Mathieu Rodrigue, Tobias Höllerer and Hwasup Lim, “Poster: Real Time Hand Pose Recognition with Depth Sensors for Mixed Reality Interfaces,” IEEE Symposium on 3D User Interfaces(3DUI), 2013.

Abstract

Information Reliability on the Social Web

Models and Applications in Intelligent User Interfaces

by

Byungkyu Kang

The Social Web is undergoing continued evolution, changing the paradigm of information production, processing and sharing. Information sources have shifted from institutions to individual users, vastly increasing the amount of information available online. To overcome the information overload problem, modern filtering algorithms have enabled people to find relevant information in efficient ways. However, noisy, false and otherwise useless information remains a problem. We believe that the concept of information reliability needs to be considered along with information relevance to adapt filtering algorithms to today's Social Web. This approach helps to improve information search and discovery and can also improve user experience by communicating aspects of information reliability.

This thesis first shows the results of a cross-disciplinary study into perceived reliability by reporting on a novel user experiment. This is followed by a discussion of modeling, validating, and communicating information reliability. A selection of important reliability attributes such as source credibility, competence, influence and timeliness are examined through different case studies. Results show that perceived reliability of information can vary across contexts. Finally, recent studies on visual analytics, including algorithm explanations and interactive interfaces are discussed with respect to their impact on the perception of information reliability in a range of application domains.

Contents

| | |
|--|-------------|
| Curriculum Vitae | vi |
| Abstract | viii |
| 1 Introduction | 1 |
| 1.1 Introduction | 2 |
| 1.2 Reliable Information | 4 |
| 1.2.1 Information Quality and Reliability | 4 |
| 1.3 Attributes of Information Reliability | 9 |
| 1.4 Opportunities and Challenges | 14 |
| 1.4.1 Intrinsic or Perceived Quality? | 14 |
| 1.5 Contributions | 16 |
| 1.6 Scope and Organization | 17 |
| 1.7 Permissions and Attributions | 17 |
| 2 Background Research | 19 |
| 2.1 Introduction | 19 |
| 2.2 The Social Web | 20 |
| 2.3 Information Reliability | 23 |
| 2.4 Perceived Reliability | 27 |
| 2.4.1 User interaction | 28 |
| 2.4.2 Visual cues | 29 |
| 2.5 Reliability Metrics | 31 |
| 2.5.1 Credibility | 32 |
| 2.5.2 Competence | 38 |
| 2.5.3 Influence | 41 |
| 2.5.4 Relevance | 43 |
| 2.5.5 Newsworthiness | 44 |
| 2.5.6 Interestingness | 45 |
| 2.6 Communicating Reliable Information | 45 |
| 2.6.1 Social Stream Filtering and Analysis | 46 |

| | | |
|----------|---|-----------|
| 2.6.2 | Real-time Visualization | 47 |
| 2.7 | Conclusion | 47 |
| 3 | Perception | 49 |
| 3.1 | Introduction | 49 |
| 3.2 | Human Perception | 50 |
| 3.2.1 | Information source | 51 |
| 3.2.2 | Visual cues and perception | 53 |
| 3.2.3 | Impact of user interface design | 55 |
| | Traditional web pages | 55 |
| | The Social Web | 55 |
| 3.3 | Experiment: Credibility Perception on Microblog Contents | 57 |
| 3.3.1 | Introduction | 57 |
| 3.3.2 | Credibility Perception Survey | 61 |
| 3.3.3 | Experimental Setup | 67 |
| | Exp1: Perception and Interaction | 67 |
| | Exp2: Artificially Controlled Content | 71 |
| | Study design | 72 |
| 3.3.4 | Results and Discussion | 77 |
| | Study Participants | 77 |
| | Influence of Metadata | 77 |
| | Cross-Feature Analysis | 78 |
| | Influence of Feature Classes | 79 |
| | Cross-Platform Analysis | 79 |
| | Impact of Different Treatments | 81 |
| | Cross-Topic Analysis | 81 |
| 3.3.5 | Summary | 83 |
| 3.4 | Conclusion | 85 |
| 4 | Modeling | 86 |
| 4.1 | Introduction | 86 |
| 4.2 | Modeling Information Reliability | 87 |
| 4.2.1 | Reliability Metrics | 88 |
| 4.3 | Modeling Credibility in Twitter | 91 |
| 4.3.1 | Experimental Setup | 92 |
| 4.3.2 | Modeling Credibility | 93 |
| | Social Model | 94 |
| | Content-based Model | 97 |
| | Hybrid Model | 98 |
| 4.3.3 | Evaluation | 100 |
| | Data Analysis | 100 |

| | | |
|----------|--|------------|
| | Predicting Credibility | 105 |
| 4.3.4 | Discussion | 108 |
| 4.3.5 | Summary | 109 |
| 4.4 | Modeling User Competence by Mapping Theory to Practice | 110 |
| 4.4.1 | Mapping Theory to Practice | 111 |
| 4.4.2 | Setup and Data Collection | 112 |
| | Data Collection | 112 |
| | Theoretical Foundation | 115 |
| | Mapping | 115 |
| | Feature Analysis | 121 |
| 4.4.3 | Evaluation | 123 |
| | Feature-based Competence Assessment | 123 |
| 4.4.4 | Summary and Discussion | 127 |
| 4.5 | Modeling News Content in Microblogs | 128 |
| 4.5.1 | Introduction | 129 |
| 4.5.2 | Content Similarity based Approach | 133 |
| | Data Collection | 133 |
| | Similarity Computation | 134 |
| | Strategy Selection | 135 |
| | Experimental Setup | 139 |
| | Evaluation | 141 |
| 4.5.3 | Network based Approach | 148 |
| | Hypotheses | 149 |
| | Data Collection and Preprocessing | 149 |
| | Labeling Tweets | 150 |
| | Network Analysis (Exp 1) | 151 |
| | Topic Association (Exp 2) | 152 |
| | Results and Discussions | 154 |
| 4.5.4 | Future Work | 158 |
| 4.5.5 | Conclusion | 163 |
| 4.6 | Modeling User Influence for Social Marketing | 165 |
| 4.6.1 | Approach | 165 |
| | Role-based User Identification | 166 |
| 4.6.2 | Model | 168 |
| 4.6.3 | Data Collection | 172 |
| 4.6.4 | Conclusion | 173 |
| 4.7 | Conclusion | 174 |
| 5 | Validation | 176 |
| 5.1 | Introduction | 177 |
| 5.2 | Features | 181 |
| 5.2.1 | Content Based Features | 181 |

| | | |
|----------|--|------------|
| 5.2.2 | User Based Features | 182 |
| 5.2.3 | Conversation Based Features | 184 |
| 5.3 | Collection and Annotation of Twitter Data | 186 |
| 5.3.1 | Annotating Twitter Data | 188 |
| 5.4 | Evaluation | 193 |
| 5.4.1 | Ground Truth Selection | 193 |
| | Predictability of Ground Truth | 196 |
| 5.4.2 | Best features in different network contexts | 201 |
| 5.5 | Guidelines for Studying Information Reliability | 202 |
| 5.6 | Summary and Discussion | 206 |
| 6 | Communicating Reliable Information | 208 |
| 6.1 | Introduction | 209 |
| 6.2 | Visual Representation of Information | 210 |
| 6.2.1 | Visualization: Data to Information and to Knowledge | 210 |
| 6.2.2 | User Interfaces for Effective Communication | 211 |
| 6.2.3 | Interactive Interfaces for Reliable Information | 212 |
| 6.3 | Real-time Systems with Visual Interfaces | 219 |
| 6.3.1 | Introduction | 219 |
| 6.3.2 | Design Considerations | 221 |
| | Real-time Message Filtering | 222 |
| | Time-window based Ranking | 222 |
| | Color-coded Visualization | 223 |
| | Sentiment with Rain Drops | 223 |
| | Logarithmic Timeline | 224 |
| 6.3.3 | System Architecture | 224 |
| | Twitter Stream Filtering | 226 |
| | Back-end Data Processing | 226 |
| | Front-end Visualization Layer | 227 |
| 6.3.4 | Visualization | 228 |
| | Sentiment Map (Raindrop Message Visualizer) | 228 |
| | Real-time Ranking Visualization | 234 |
| 6.3.5 | Deployment, Reception, and Discussion | 236 |
| | Continuum of Discontinuity | 236 |
| | A Scenario-based Observation | 237 |
| 6.3.6 | Summary | 238 |
| 6.4 | Inspectability and Personalization in Social Content Discovery | 240 |
| 6.4.1 | Introduction | 240 |
| 6.4.2 | Background | 242 |
| 6.4.3 | Formative User Study | 245 |
| 6.4.4 | HopTopics System | 246 |
| | User Interface Design | 247 |

| | | |
|----------|---|------------|
| | Interaction Design | 248 |
| | System Architecture | 250 |
| 6.4.5 | Main Experiment | 251 |
| | Experiment Design | 251 |
| | Hypotheses | 252 |
| | Participants | 255 |
| | Materials | 255 |
| | Procedure | 256 |
| | Results | 258 |
| 6.4.6 | Discussion and Future Work | 265 |
| 6.5 | Summary | 266 |
| 7 | Conclusions and Future Work | 267 |
| 7.1 | Introduction | 267 |
| 7.2 | Reliable Information on the Social Web | 268 |
| 7.3 | Objectives and Contributions | 268 |
| 7.4 | Identifying Reliable Information with Intelligent Algorithms and Robust Ground Truth | 269 |
| 7.5 | Information Reliability and the End User | 271 |
| 7.6 | Future Work | 272 |
| | Bibliography | 274 |

Chapter 1

Introduction

This dissertation is about information reliability models on the Social Web and their applications in intelligent user interfaces. In this thesis, we discuss multiple studies that investigate into how reliable information can be modeled and automatically identified on the Social Web in order to support modern information filtering algorithms for information seeking tasks. Additionally, we study how information reliability models can be applied to intelligent user interfaces and how such intelligent systems can effectively and interactively highlight reliable high-quality information from heterogeneous or noisy data on the Social Web. This chapter begins with the main motivation of the research including background information in Section 1.1. Section 1.2 sets the scope of the high-quality information we refer to throughout this thesis by discussing the framework of information quality on which we base our studies and proposing a narrow definition of information reliability. We also review several definitions that have been proposed by researchers from related fields. In Section 1.3, we present an overview of different aspects of high-quality information studied in many disciplines, such as computer science, communications and social science. To underpin the motivation and objective of our work, in Section 1.4, we discuss critical issues and difficulties in identifying such high-quality

information and expected benefits. Lastly, Section 1.6 outlines the organization of this thesis.

1.1 Introduction

Over the last decade, the Internet has evolved with many modern means of communication that allow people to connect with anyone, anytime and anywhere. Unlike personal or organizational websites, which support one-to-many communication, modern social networking platforms provide many-to-many communication as the core feature, allowing information production and exchange at an unprecedented scale. Moreover, nowadays users can share information instantly using mobile computing devices such as smartphones. This technological advancement has enabled individual users on the social networks to become active information providers and consumers at the same time, or “prosumers”. To this end, there have been significant efforts made by researchers to improve scalability, accuracy and relevance of information filtering algorithms. For example, [1, 2, 3, 4] propose methods to improve scalability of search algorithms by sorting huge amount of data in near real-time applications. Other studies [5, 6, 7] propose new algorithms for improved accuracy and relevance of the search engine results page (SERP). More effective and/or efficient models and algorithms are still being investigated and implemented by researchers.

However, the ease with which a user can access, create and share information turned our attention to other information quality problems. A large portion of the information on the Social Web has been published with no quality control. These problems have made finding reliable, high-quality information on the Web more difficult [8]. In fact, in many cases, information consumers are responsible for assessing quality of information

on the Web¹. For example, malicious or careless users publish rumors or misinformation on the Social Web, and such information can get quickly disseminated as users share it with others on the network without verification.

Moreover, users not only want to find relevant, accurate and credible information but also often search for interesting information. Searching for interesting or unique information, often referred to as “serendipitous search” (defined in [9]), is difficult since there is no established universal method to measure such subjective quality of information.

In this dissertation, we present through different case studies how we model, measure and identify such complex, high-quality information on the Social Web and how they can be effectively presented to users using interactive and intelligent user interfaces.

Why the Social Web? According to the survey conducted by PEW Research Center in 2014², 74% of Internet users (online adults) use online social networking services. Interestingly, the population is evenly distributed across different demographic factors such as gender, age, education or yearly income. Moreover, a recent study [10] found that the majority of Internet users utilize participatory Web environment (Web 2.0), particularly social networking platforms, as *information sources*, indicating that the Social Web has become the primary source of information when people search for information of their interest in various needs. Journalists also have turned their heads towards these online social platforms for producing their news contents [11, 12]. This new paradigm in the knowledge industry has been reshaping how people consume information in daily life from one-to-many to many-to-many relationships.

Please note that we frequently use the term “Social Web” to specify the Social Web applications/platforms (e.g. Twitter, Wikipedia and Reddit) examined in our studies and

¹<https://www.westernu.edu/bin/computing/online-information-quality.pdf>

²Social Networking Fact Sheet, PEW Research Center, January 2014. <http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/>

to distinguish them from other provenance of information. Including typical microblogging platforms like Twitter, such Social Web applications incorporate social dynamics and communication aspect, and thus, they are worth to be studied and analyzed to understand the contemporary practice of information production and consumption online.

1.2 Reliable Information

We briefly discussed the main motivation of our studies and the importance of Social Web applications and platforms as the information source. In this section, we discuss an established information quality framework and elaborate on how we relate the reliable information we refer to in the studies to this framework by exploring some attributes of information quality on the Web.

1.2.1 Information Quality and Reliability

Quality of information or data has long been a critical topic in information systems (IS) and other related disciplines. In the recent years, researchers in information systems have begun interpreting data or information as a value [13] or another type of product [14, 15] (e.g. output of a manufacturing process). According to Wang [15], an information product (IP) can be defined, measured, analyzed and improved like other manufacturing products. Particularly, in Wang and Strong [14], the authors claim that “poor data quality can have substantial social and economic impacts.” They also assert that “high-quality data should be intrinsically good, contextually appropriate for the task, clearly represented, and accessible to the data consumer.” This statement demonstrates the complex constructs, i.e. attributes, of information quality. In [15, 14], information quality (IQ) is comprised of multiple dimensions (attributes) and they can be classified into the four categories: *Intrinsic IQ*, *Accessibility IQ*, *Contextual IQ*, and *Representational IQ*

as shown in Figure 1.1. Wang dubbed this classification the *Conceptual Framework of Data Quality*. The multidimensional aspect of information quality also can be seen in other studies [13, 16, 17, 18] as well. As in [18], this conceptualized framework can be extended to the Social Web with a special treatment to address the significant differences between traditional media and the Social Web. We will discuss the portable (important) attributes in detail later.

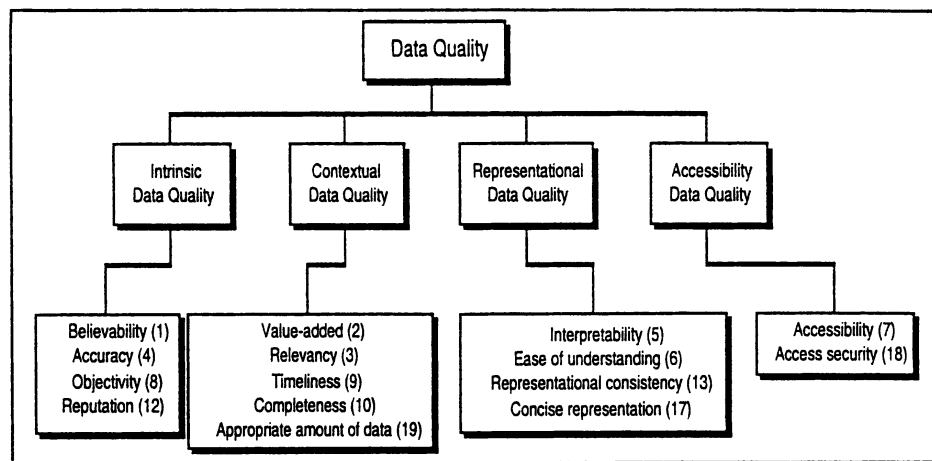


Figure 1.1: Wang and Strong’s conceptual framework of data quality in [14].

So far, we have discussed data or information as “product of an information manufacturing system”³. However, can we apply the same standards to the Social Web? In other words, is it possible to specify a general stage of information production, and who has the control for such quality assurance tasks (e.g. measurement, analysis or improvement)? Simply put, the answer is that it can not be interpreted exactly in the same way, but similarly.

Since a large portion of information production and dissemination has shifted from traditional or industrial sources to individual users, most (user-generated) content in the Social Web consists of products of unprofessional writing, such as prose and personal

³Please note that we use data and information interchangeably in this dissertation unless otherwise specified.

conversations that do not require a systematic quality control. Such information quality problem is more common in today's Social Web unlike traditional publishing environments [19], and we believe that this characteristics of the modern participatory Web needs to be carefully considered.

Adaptation of the Quality Framework What is the core difference between traditional information sources and social networking platforms? In the Social Web, assessment of information quality is up to the end users. Furthermore, most users play a dual role as an information provider and a consumer. Recently, some traditional news outlets and publishers (e.g. NYTimes, BBC, The New Yorker, etc.) have begun to harness social platforms for more exposure. In social networks, such carefully curated content from official sources is often considered more trustworthy and newsworthy information and also referred to as *ambient journalism* [20, 21]. However, this is not always true [22]. For example, a journalist working for one of the major news providers mistakenly posted a tweet about the Queen Elizabeth II's death without verification⁴. As another example, during the Boston Marathon Bombing coverage, a false report was announced on a live news channel⁵. These examples highlight not only the unique aspect of the Social Web but the importance of quality control and the seriousness of outcome when the problem is not properly addressed. Lastly, due to the brevity of information communicated through social platforms, *perceived* quality of information is another critical factor in assessing information; see Chapter 3. Taking these aspects into consideration, we adapt the information quality framework [15] to the Social Web for our studies.

In this dissertation, we focus on identifying reliable information on the Social Web by deconstructing information quality attributes into individual components and computing

⁴<http://www.telegraph.co.uk/news/bbc/11648109/BBC-journalist-apologies-after-accidentally-announcing-Queens-demise-on-Twitter.html>, last accessed on March 25, 2016.

⁵<http://www.hollywoodreporter.com/news/cnn-boston-marathon-bombing-mistake-441551>, last accessed on March 25, 2016.

them automatically. Based on Wang’s information quality framework, which captures the aspects of data quality in traditional environments, we propose our modified framework for the Social Web.

| IQ Category | IQ Attributes (Dimensions) |
|----------------------|---|
| Intrinsic IQ | Accuracy, Credibility |
| Consumer-oriented IQ | Relevance, Credibility, Accessibility, Security, Completeness |
| Producer-oriented IQ | Competence, Authoritativeness, Reputation, Influence |
| Consumer-Producer IQ | Credibility/Trust, Timeliness, Newsworthiness |

Figure 1.2: Proposed Information Quality Framework for the Social Web. We employ several attributes in this framework to describe information reliability in participatory Web environments; see Section 1.3.

Reliability According to Pierce [23], a general definition of information reliability is as “the extent to which we can rely on the source of the data and, therefore, the data itself.” Pierce also states that “Reliable data is dependable, trustworthy, unfailing, sure, authentic, genuine, reputable.” As can be seen in Figure 1.2, we employ *information reliability*⁶ as a more generic concept that underlies different (quality) attributes of information and their sources, such as credibility, accuracy, and competence of a source. There is no globally accepted scheme or framework that rationalizes (an exhaustive list of) information quality attributes and the relationships between them. We use the term *reliability* for both information and source. We use this term restrictively to conceptually integrate multidimensional attributes of information quality. The definition of information reliability we use in this dissertation is as follows.

Definition 1 *Information Reliability* *The degree to which an information consumer can depend on a piece of information or its source to make a decision or judgment in a*

⁶Please note that *information reliability* as discussed throughout this thesis is characterized by the definition here in Section 1.2.

particular context of task.

Using the term *reliability*, we refer to one or more attributes of information quality. The underlying rationale of the proposed framework is that reliability becomes a critical condition when an information-seeking task entails an action or a decision for which the information found is used. In other words, when a user performs an information-driven decision making, the result may vary depending on the quality of information on which the user relies. Furthermore, such example can be extended to automated, large-scale information-driven tasks such as collective intelligence and social marketing using algorithms and data on the Social Web. As a conceptual model of information quality on the Social Web, restricted in the context of information-driven actions such as a decision making, we discuss what is reliable information across different topics and applications and how we can identify such information by focusing on a variety of attributes that construct our information quality framework.

Reliability in the Literature Reliability of information or sources have long been studied by many researchers. A few studies [24, 25] employ “information reliability” as a superordinate concept of other subordinate attributes, which are similar to our approach in this thesis. On the other hand, there are approaches that build on different constructs, schemes, attributes or terms [13, 26] to model information quality on the Web. To the best of our knowledge effort to define and study an information quality framework for the Social Web.

Reliability in Information-Seeking Practices Let us assume that you want to search for information on a topic in Twitter. Regardless of the given topic of interest, you might assess the quality of the search results based on some attributes of information. For instance, if you search for news on a specific topic, you may consider “relevance”

and “timeliness” of information as well as “credibility” of both the source and the information. How users assess information quality in a search task often varies across users and contexts. However, recent studies [27, 28, 29, 30] found peculiar patterns such as more focus on visual representation, content-based cues and social credibility in search tasks. Others examined such patterns in a specific context, for example, browsing health information on the web [31, 32, 24]. To guide people performing such tasks, Metzger [33] integrated several findings and proposed guidelines for assessing information credibility on the Web. We will further discuss how humans assess online information in detail in the subsequent chapters (Chapter 2 and 3).

We will discuss how to model, measure and detect (Chapter 3 and 4) reliable information on the Social Web and how to validate (Chapter 5) our models in quantitative and qualitative ways.

1.3 Attributes of Information Reliability

We extended and modified one of the traditional information quality frameworks and proposed a narrow definition of *reliability* that features several attributes of information quality on the Social Web.

As shown in Figure 1.2, many attributes in the information quality framework construct the information quality space. Many studies on these attributes have employed different definitions and methods to identify reliable information across contexts and tasks. For instance, human-factor analysis approaches [20] and computational approaches [29] are used to evaluate information credibility on microblogs.

Information seekers may consider one or more attributes based on a specific given task. In an emergency-response situation, for example during an earthquake, users commonly search for the latest news on the Web. In social media sites such as microblogs, users

may begin their search with a keyword related to the event, seeking messages that report on experiences at the scene of the event. In this particular context, *credibility*, *accuracy* and *timeliness* of information play important roles.

In this section, we enumerate reliability attributes and discuss how they can be defined and measured based on the findings in different fields of study. The attributes presented in this section is the list of metrics that we cover in this thesis. We selected these attributes for our study since they have been recently understood by researchers as important quality measures on the Social Web, particularly in the context of data- or information-driven decision making. Please note that this framework do not form an exhaustive list of information reliability.

Credibility Credibility is a subjective attribute of information reliability that is difficult to measure quantitatively. This attribute is sometimes used interchangeably with *trust*. However, at the same time, credibility also can be defined as a concept differing from trust, and its meaning may vary depending on a given context. There is no a-priori assumption or definition that is widely accepted across disciplines. Fogg and Tseng [34] defined online information credibility as “a perceived quality made up of multiple dimensions” in their study and, based on their literature review, they found that the terms *believability* and *credibility* are interchangeable in most cases when it comes to information (sources) online.

Furthermore, both perceived and intrinsic credibility have been studied by researchers across different contexts. For example, credibility perception has been an important research topic in communications and computer science [16, 35, 20]. A typical approach used on this topic is taking an information consumers’ perspective and assuming perceived credibility as an independent attribute from the intrinsic, which resides within the information (or the source of information) [29, 30]. In this thesis, we focus on both

credibility metrics since the two indicators have different contributions to the course of reliability assessment on the Social Web.

Relevance Given a topic of interest, users can assess relevance of information retrieved in a search task. Many information retrieval algorithms apply first-round relevance and popularity filtering (e.g. Google’s PageRank) in search engines. There are different metrics by which we measure relevance of information with respect to the topic of interest. Some metrics can be derived statistically from search engine query logs, tags, or topical categories in a knowledge repository such as Wikipedia. In the field of information retrieval, relevance metrics have been developed and are widely used to assess the quality of search engine results. There are also variants of modern search algorithms that deviate from traditional formulations of information relevance. For example, PageRank takes bi-directional links between webpages as votes (proxy agents for popularity and/or relevance) given a search query, rather than directly employing the relevance metric. Furthermore, in exploratory search, topical relevance may not be the crucial factor since this type of task values serendipitous findings as long as information is interesting and minimally related.

Expertise (Competence) According to the Merriam-Webster dictionary, the term *expertise* is defined as “the skill or knowledge an expert has.” This implies that this is an intrinsic attribute that represents an entity’s ability or competence regarding a specific skill or topic. In the Social Web, an entity or user can be considered a source of information. The degree of expertise or competence of such a source may be assessed through the observed quality of information authored by the source. However, the time-variant aspect of a user’s expertise is also important. A user with varying quality in authored content exhibits a different level of expertise from another user who maintains

consistent quality. In the same vein, some users are less prolific than others and have been inactive for a long time. Expertise or competence of an information source can be measured differently across contexts. In many cases, expertise is used synonymously with competence. When a distinction is made, however, competence is generally considered a higher-level concept, encompassing (elements of) expertise.

Influence Often, microblog messages and blog posts “go viral” immediately through the sharing cascades on the respective social network. For example, features such as shares and likes in Facebook and favorites and retweets in Twitter make a message immediately visible on the timeline of friends or followers. In turn, a message could be seen by hundreds of thousands, or even millions of users within just a few hours, depending on the popularity of the content. Harnessing various metadata and features that can be extracted from postings, computer scientists now endeavor to predict the outcome of political elections [36, 37, 38] and customer feedback on a new product in the market [39] in real-time.

The concept of “influencer marketing” has recently emerged as a strategy to support effective social marketing in industry. Enterprises have started looking into the potential of the “influencers” on the social network. To understand and predict influential accounts or content in the practice of social marketing, computational algorithms [40] have been devised and deployed in web applications [41].

Newsworthiness Users want to find newsworthy information with ease on the Web. Identifying newsworthiness has developed into a major research problem over the years, which is difficult to solve due to the lack of an adequate quality control mechanisms. In the areas of communications, journalism and public relations, researchers have studied the elements (e.g. impact, timeliness, proximity, prominence and oddity) that constitute

newsworthiness across contexts and effective approaches to producing and predicting newsworthy content. Over recent years, computer scientists have developed automated algorithms that identify newsworthy information in social networks to help users find credible and newsworthy information [29, 42, 43]. In the field of computer science, the vast majority of work on newsworthy information on the Social Web makes use of data mining (e.g., association rule/sequence mining) and machine learning (e.g., via decision trees, boosting, or deep learning). For example, Castillo et al. [29] applied different machine learning algorithms to numerous features extracted from microblog messages and found that a decision tree classifier (J48) outperforms other classifiers in prediction accuracy.

Interestingness As a measure of information quality, interestingness has been studied for a few decades in the field of data mining [44], knowledge discovery in databases (e.g., under the heading of “automatic analysis of changes and deviations”) [45, 46] and information retrieval (unexpected information) [47]. Geng and Hamilton [44] proposed nine facets that construct information interestingness in a definition: “a broad concept that emphasizes conciseness, coverage, reliability, peculiarity, diversity, novelty, surprisingness, utility, and actionability.” As with information credibility, there is no globally accepted definition of interestingness of information. Nonetheless, many recent studies assume that interestingness can be understood as “unexpected relevant information or knowledge”. In information search or knowledge discovery, interestingness is considered an important quality measure in the context of serendipitous search. Furthermore, interestingness has become an indispensable factor for information filtering in the Social Web, since information seekers favour interesting or unique information that stands out among all relevant information returned by their search. In this thesis, we study interestingness of information from the user experience perspective in Chapters 4 and 6.

1.4 Opportunities and Challenges

Arguably, information overload itself is not a single major problem from an information consumers' perspective. This phenomenon can be understood as one of the by-products of contemporary information technology. Thousands of big corporations and government agencies have shown their interest in "Big Data", which has been a buzzword for years across all industries. They have decided to invest research efforts on this topic since many of them recognize this phenomenon as an opportunity in their business. For example, Amazon and Netflix developed recommender systems to provide collaborative filtering and recommendation services for customers. Such services have been very successful, and these companies plan to be even more proactive in big data technologies.

However, many challenges remain regarding the quality of available information on the Web. There are not enough established quality control mechanisms in the Social Web [16] due to the low barrier of publication to every user and the ease of information sharing in social networks. Users post or share content without a careful evaluation of its quality. At the heart of discussions on Big Data, we often find quality-related issues. For example, the familiar "Four V's" of Big Data are *Volume*, *Variety*, *Velocity*, and *Veracity*. In this dissertation, we consider all these characteristics, with a particular focus on *veracity*.

1.4.1 Intrinsic or Perceived Quality?

Many attributes of information quality in the participatory Web, including social networks, have been studied through different lenses. For example, Wang and Strong [14] define *believability* as "an integral part of intrinsic data quality." On the other hand, Fogg and Cheng [34] claim that "believability is a good synonym for *credibility* in most cases." They assert that it is a *perceived quality* comprised of the two key components, trustworthiness and expertise. As can be seen in this example, many quality attributes

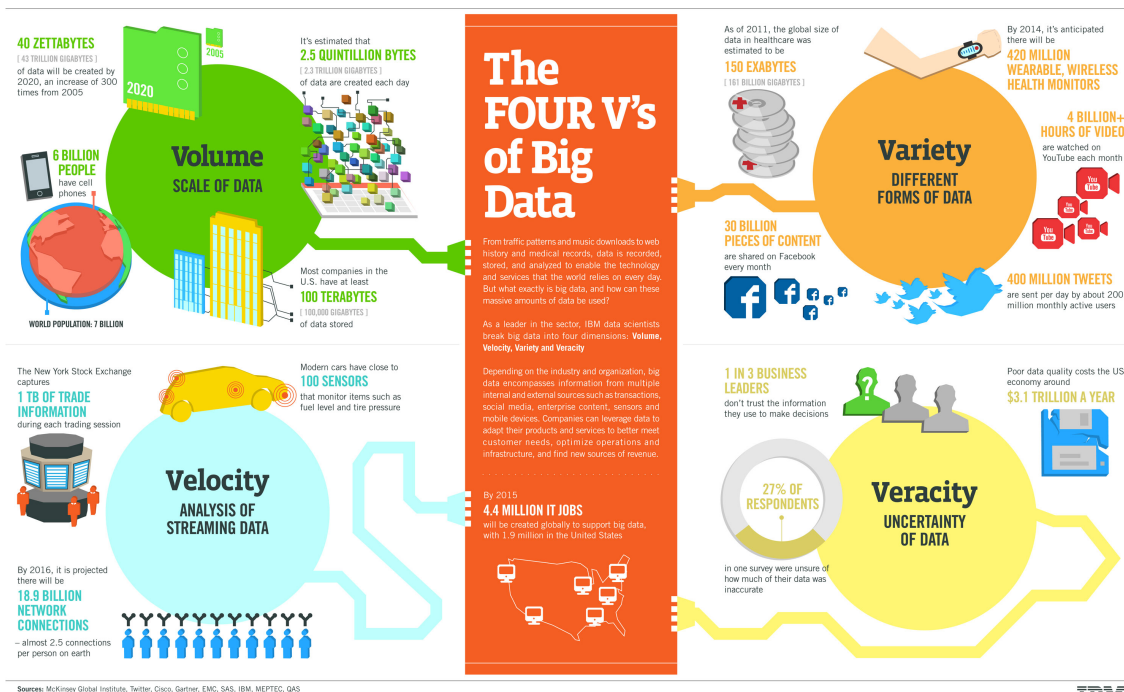


Figure 1.3: An infographic that explains the four dimensions (4V's) of Big Data: volume, variety, velocity and veracity. (Excerpt from <http://www.ibmbigdatahub.com/infographic/four-vs-big-data>).

need widely accepted definitions, even within a discipline. We discuss properties of individual quality attributes in detail in Chapter 2 and Chapter 3.

1.5 Contributions

In this thesis, we present three major contributions: 1) We developed and evaluated models for several important aspects of information reliability in today's Social Web; 2) We constructed more robust ground truth data as a combination of uncorrelated and noisy measurements for reliability studies; 3) Using our information reliability models, we designed and implemented novel and effective means of communications that help users better understand data and discover knowledge.

Below, we provide a breakdown of detailed contributions of the thesis.

- We provide a literature survey of information quality frameworks and a wide range of methods that filter, detect and predict information reliability.
- We propose a modified framework of high-quality (reliable) information based on one of the established frameworks of information quality.
- We provide different models of either intrinsic or perceived social web information reliability across tasks and contexts.
- In order to effectively communicate with extracted reliable information, we propose and implement several novel visualization and information management techniques.
- To validate information reliability models, we provide effective methods to find ground truth of information reliability.

1.6 Scope and Organization

This dissertation focuses on modeling, identifying and communicating reliable information on the Social Web. We approach our research problem from a holistic perspective since there are numerous factors that can explicitly or implicitly affect information reliability assessments by human users. In order to achieve our research goal, we conduct small and large-scale user experiments with human subjects, model different attributes of information reliability, and validate our models based on the ground truth discovered by investigating behaviors and feedback of real-world information seekers.

In this thesis, we structure the presentation of our work as follows. At the beginning, we introduce our motivations and research goals, review related studies in the literature and discuss the overlap and distinguishing characteristics of our research with this body of work. Subsequently, we discuss the perception of information reliability, theoretical and computational models, and ground truth to be used for validating computational reliability models. Lastly, we present our recent work on intelligent interfaces designed for effective communication of reliable information.

1.7 Permissions and Attributions

1. The content of Chapter 4.4 has been previously published at the International Conference on Social Computing (SocialCom) [48].
2. The content of Chapter 4.6 contains the result of my summer research internship at Adobe Research in 2014 with Dr. Nedim Lipka.
3. The content of Chapter 5 is the result of my collaboration with Professor Sibel Adalı's group at the Rensselaer Polytechnic Institute, and has previously appeared

in the Proceedings of IEEE SocialCom (2013) [49], and in the ASE Human Journal 2.1 (2013) [50]. It is included here with the permission of the co-authors.

4. The content of Chapter 6.4 is the result of my collaboration with Dr. John O'Donovan at UCSB and Dr. Nava Tintarev at the University of Aberdeen. This work has been submitted to the ACM International conference of Recommender Systems (ACM RecSys). Dr. Tintarev is currently working as an assistant professor at the Bournemouth University.
5. Special thanks to John O'Donovan for his valuable contributions to this thesis as a main collaborator. This thesis work has been completed under the supervision and guidance of Prof. Tobias Höllerer.

Chapter 2

Background Research

2.1 Introduction

In this chapter, we provide a comprehensive review on the studies from the literature that are highly relevant to our research topic. Reliability of online information has been studied over many years since the beginning of the Internet. However, yet, there is no clear or widely accepted definition of information reliability. In the previous chapter, we discussed one of the information quality frameworks in information systems and the difference between traditional information systems and information on the participatory Web. For our studies, we proposed *reliability* as an umbrella term to describe multiple related attributes to information quality in the Social Web. Particularly, we discuss definitions of the attributes, methods used by researchers for mining relevant data sets and techniques for modeling and predicting reliable information. Interesting approaches proposed and discoveries made by other researchers as well as what makes our approaches unique compared to the state-of-the-art methods are highlighted.

We first introduce general discussions and current research topics on the Social Web in Section 2.2. Related studies on information reliability and important subjects in the

study on perceived reliability are covered in Section 2.3 and 2.4. Since our research mainly focuses on a set of important attributes of information reliability (e.g. credibility, relevance, competence, newsworthiness), we discuss different information reliability metrics studied in computer science and other related disciplines in Section 2.5. This is followed by various topics on how to effectively communicate reliable information. In Section 2.6, we introduce case studies about novel visualization and intelligent user interfaces that improve user experience and satisfaction in information seeking and retrieval tasks and data analysis tasks.

2.2 The Social Web

In this section, we briefly introduce and discuss common research topics on the Social Web. According to the technical report published by W3C¹, the Social Web is “a set of relationships that link together people over the Web.” A large portion of social media content can be expanded to a variety of different information sources on the Internet by following embedded hyperlinks such as urls. By the definition, such information sources also belong to today’s Social Web, and thus, it is worth to study both the inside and the outside of social networks to understand and model information reliability. Accordingly, in this thesis, we extend our interest to the Social Web.

Information Reliability and the Social Web Experts in information technology keep warning that the advance of database and distributed computing technology may not be able to catch up the speed of information overload at a certain point and, eventually, humans will experience an era of stagnation in knowledge. Dan Ariely, a behavioral economist, recently said that *“Big data is like teenage sex: everyone talks about it, nobody*

¹World Wide Web Consortium <http://www.w3.org>

really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it."² This wise saying has been spread over the Social Web. Our research topic is on the edge of "the Big data epidemics" since the effective and efficient validation of information in the deluge of data nowadays is perhaps a necessary evil. There are a great deal of malicious information or software on the web to deceive users with false information or harmful intent. Thus, more effort is required to help people find reliable information online. As more Internet users utilize the Social Web such as microblogs or social networking platforms as rather information source than simply communication channel, we limited our research topic on social web data.

Information overload Over the last decade, "*Big Data*" has been one of the hottest buzzwords in both academia and industry. User-created content now account for the vast majority portion of the information in the Web. In 2004, O'Reilly media has coined the term "Web 2.0" and numerous social applications have been announced. Shortly after that the world wide web has been inundated with unprecedented volume of data. In the Social Web, both opportunities and challenges coexist due to the information overload, caused by low barrier of publication for users. Applying various data analytics methods to social media data, it is possible to gain insights using collective intelligence and predict human behaviors through network analysis.

However, the more data we produce and consume, the more powerful computing resources are required. Although recent progresses in lightening-fast multi-core processors and hundreds of petabytes storages address this issue in part, many challenges remain. For example, linear increase in the volume of data often requires orders of magnitude more resources. Furthermore, improved flexibility in data production has yielded high heterogeneity of information. Increased dimension and complexity of data make it more

²<http://www.philsimon.com/blog/featured/big-data-and-teenage-sex/>

difficult for users and other service providers on the Web to find, process and understand available data. To this end, more efficient and intelligent algorithms have been studied. To fill the gap that traditional filtering algorithms can not address, machine learning algorithms and information retrieval techniques are studied and applied to the Social Web data.

Information quality in the Social Web In the Social Web, the combination of information overload and quality problem makes users more challenging because it is extremely difficult to locate high quality information of interest promptly. To address the quality problem, the aforementioned quality measures (we use “attributes” in this dissertation) are individually, or together with others, investigated by information scientists. We provide related work on individual attributes in detail in Section 2.5.

Information seeking Before the Internet, people used to find information from traditional knowledge repositories such as books, periodicals and newspapers. Such process involves a large amount of effort and time. The advent of the Internet has made this process a lot easier for us. However, people begin to realize the need of new quality mechanisms to assess the quality of information online. In the contemporary participatory Web, by extension, the authenticity, or credibility, of information has become more important in quality assurance. Many Internet users now publish and consume information on social networking platforms such as Facebook, Twitter or Instagram as well as personal blogs. Therefore, the Social Web overflows with unnecessary or bad information. This problem is getting more serious as each user’s network becomes more complex and bigger than ever before.

To understand the underlying pattern of user communication on the Social Web, many researchers investigated social applications, for instance, personal blogs [51, 52, 53]

and microblogs [54, 55]. As mobile social applications that support multimedia contents become a new information outlet, users now curate and publish multimedia contents on microblogs such as Instagram³, Flickr⁴ or Vine⁵. This new paradigm has recently attracted researchers to focus on multimedia contents for information retrieval [56, 57, 58]. Major traditional news outlets such as ABC⁶ and BBC⁷ have recently started providing their news contents through multimedia-based social platforms. These fast-changing trends make it more difficult for data scientists to design algorithms to address information reliability issues.

2.3 Information Reliability

In Chapter 1, we introduced several information quality frameworks on which we base our information reliability model and proposed a definition of *reliability*. Although we repeatedly use the term “information” throughout this theses, better disambiguation on this term may be necessary since there are many different forms of information existing on the Web. Figure 2.1 depicts the Data - Information - Knowledge - Wisdom hierarchy (DIKW) [13], referred to as the *knowledge pyramid*. Our research aims to the levels of this knowledge hierarchy (except wisdom layer), albeit primarily focusing on both “data” and “information” layers. In Chapter 6, we discuss how to effectively transform data and information into knowledge through novel visualization techniques combined with intelligent user interfaces. This type of techniques is referred to *Visual Analytics*. We reserve an individual section in Chapter 6 to discuss the benefit that modern visual analytics techniques can provide so as to reliable information is communicated with users

³<https://www.instagram.com/>

⁴<https://www.flickr.com/>

⁵<https://vine.co/>

⁶<http://instagram.com/ABCNews>

⁷<http://instagram.com/BBCNews>

in better ways.

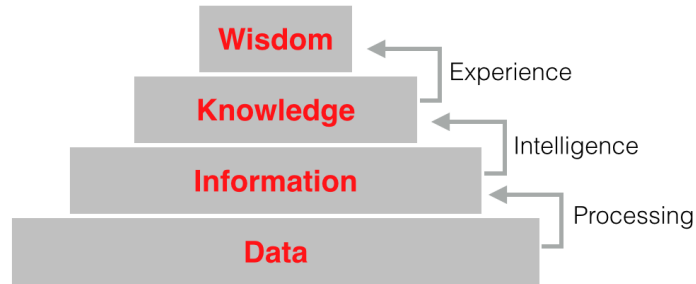


Figure 2.1: The Hierarchy of Data (Data–Information–Knowledge–Wisdom (DIKW) Model) [13]

DIKW Model Let us begin with the definition of the knowledge pyramid [59, 60]. According to Zins [60], “many scholars claim that data, information, and knowledge are part of a sequential order” and this knowledge pyramid maintains a hierarchical structure, comprised of individual elements (wisdom, knowledge, information and data). For example, data are the raw material that construct information and, in the same way, knowledge is created by synthesizing a group of information. Still, the meanings of individual elements and the nature of the relations among them is arguable. There are different versions of the DIKW pyramid which are represented in different structures such as a chain [61], a framework [62], and a continuum [63]. Although there has not been enough corroboration or consensus among scholars, many researchers agree that these elements construct a purported structural relationship between them.

Rowley [59] defines the relationship between the elements of the DIKW pyramid as follows:

- Typically information is defined in terms of data, knowledge in terms of information, and wisdom in terms of knowledge.

There are numerous definitions on each element of the hierarchy. We provide the

most frequently used definitions below, except “knowledge” since its definition is still controversial and have no enough consensus from scholars.

- *Data* is conceived of as symbols or signs, representing stimuli or signals [60] that are “of no use until...in a usable (that is, relevant) form” [59]
- *Information* is contained in descriptions and is differentiated from data in that it is “useful” [59]. Information is inferred from data, in the process of answering interrogative questions [59, 64] thereby making the data useful [65] for “decisions and/or action” [66].
- *Knowledge*⁸ is a fluid mix of framed experience, values, contextual information, expert insight and grounded intuition that provides an environment and framework for evaluating and incorporating new experiences and information. It originates and is applied in the minds of knowers. In organizations it often becomes embedded not only in documents and repositories but also in organizational routines, processes, practices and norms. [67, 68]

Please note that 1) we use both terms *data* and *information* interchangeably in most chapters, unless a distinction is provided and 2) we do not focus on the “wisdom” layer of the DIKW knowledge pyramid in this thesis.

Information reliability In general, information reliability is referred to as a measure by which information seekers evaluate quality of information for research purpose. For example, university libraries or research institutes provide guidebooks that help researchers or students find reliable information sources⁹. However, as we discussed earlier

⁸Knowledge is generally agreed to be an elusive concept which is difficult to define. [59]

⁹Example: <http://www.mhhe.com/mayfieldpub/webtutor/judging.htm>

with reliability attributes in Section 1.3, we refer to information reliability as a comprehensive, and rather abstract, metric which encompasses various information elements. These elements, namely attributes, are the quantitative metrics through which we can identify and recommend reliable information under a specific context. In other words, we interpret this term in a broad sense, not only focusing on trustworthy or authenticity of information but other metrics such as interestingness, influence, etc. Figure 2.2 shows a high-level overview of the identification process of reliable information. In order to understand a user’s goal of information seeking task, previous studies have focused on a manual query-log investigation approach. In a recent study, Lee et al. [69] proposed an automatic query-goal identification technique using user-click behavior and anchor-link distribution and their human subject study proved 90% of accuracy using this approach.

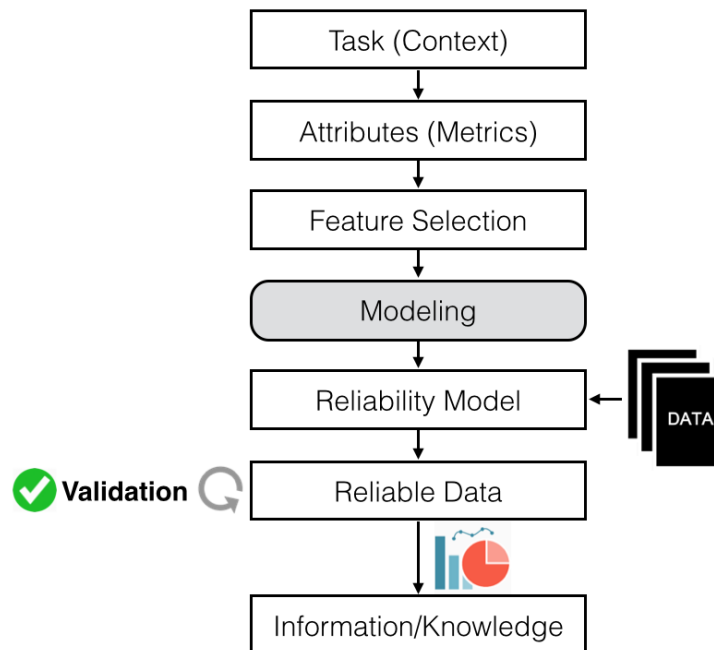


Figure 2.2: A flowchart of the reliable information identification process we use in this thesis.

In the narrow sense, reliability of information on the Web have been studied with different terminologies, mostly as the quality of information. For example, Strong et al. [70]

claimed that we can refer to high quality data as the data that fit for use by the data consumers. That is, the quality or usefulness of data depends on its users, including the intended use of the data. Therefore, according to this assertion, subjective factors must be considered when we assess information quality. Parker et al. [17] summarize different claims made by researchers and existing frameworks designed for assessing Internet information quality in their study.

Interestingly, information quality research in the context of medical or health information have frequently used the term “reliability” [35, 71, 72, 32]. One thing to note here is that many of these studies mainly focus on the credibility or reputation of the information sources for evaluation. This is because, not like the modern social web applications, the provenance of information is evident in traditional web sites. For example, in the context of medical information seeking tasks, users are generally assumed to be non-professionals who lack expertise on the topic of interest.

2.4 Perceived Reliability

In computer science, information reliability metrics are typically studied without a restriction on either intrinsic or extrinsic aspect of them. For this reason, sometimes it is difficult to understand which side of a given attribute is referred to by the study. Furthermore, each attribute of information is not mutually exclusive to others. To address the kind of confusion, for example, Fogg and Tseng differentiated credibility from trust by the following definition in their study [34].

- *credibility* \equiv *believability*
- *trust* \equiv *dependability*

We have a similar viewpoint on the both notions. However, we apply more holistic

perspective of *dependability* to the reliability of information as an embracing conception or aspect in the hierarchy that overlooks other information attributes (e.g. credibility, competence, newsworthiness, etc.); see Figure 1.2.

User perception experiment To understand how people perceive information reliability, various factors such as user interaction, visual cues and meta-information have been used by researchers. Based on the hypotheses (or null hypotheses) made, these factors are typically controlled in human subject experiments. Participants of the experiments are required to respond their perception, usually in Likert scale of predefined metrics, while they are shown with manufactured or real-world examples. If the experiments are *between subjects test*, the examples are selectively shown to each participant. Hypotheses are confirmed, or denied, through statistical analyses such as correlations between the applied treatments and their corresponding responses answered by the participants. Often, linear regression is used to verify underlying patterns of user behavior. More details about typical experimental designs are discussed in Chapter 3.

In this section, we review several studies from literature which investigate how humans perceive information reliability. We categorize the studies according to the approaches used and briefly discuss information attributes measured for the assessments.

2.4.1 User interaction

In many studies, user interaction or experience data are used as an important cue for perceived quality of information on the Web. Commonly, they are measured as user feedback information in different algorithms with the basis of assumption that particular user feedback is a signal of user preference or interest in the information to which a user interact. Before they apply a specific user feedback (or interaction), human perception experiment is performed to confirm the validity of the cue as a feature.

Xu et al. [73] provide a personalized web page ranking algorithm. Their approach is based on the attention time spent on a web page by a user. Using this metric as a feature, their algorithm produces a user-oriented web page ranking. This approach can be understood as a personalization of reliable information retrieval. Retrieved rankings from this algorithm were compared to the results from Google's PageRank algorithm for evaluation. However, cold start problem is still the downside of this approach since the algorithm adopts dwell time on individual page as training data. Other approaches [74, 75] use user clicks as a proxy of user preference of information on the web. Machine learning algorithms such as SVM [76] and singular value decomposition (SVD) [77] are also combined with click data to boost the accuracy of their approaches. The importance of user interaction data increase as this type of cue is considered as a good implicit user feedback, which can be easily obtained without any disruption in user experience. However, since there are various user interfaces developed by third-party service providers for social web applications, particularly for mobile devices, this approach has more issues to be solved in the context of social web information.

2.4.2 Visual cues

As more advanced web technologies are introduced, more sophisticated presentation techniques become the standard on the Web. This phenomenon seems more prevalent in many modern social web applications since this type of platforms requires more complicated functionality. For this reason, visual elements in social web platforms play a significant role in assessing the reliability of information.

Visual perception The human vision system provides powerful ability in processing visual stimuli. Photoreceptor cells in the retina of each eye accept incoming light. Afterwards, these stimuli are processed in primary visual cortex. Subsequently, the cerebral

cortex, which is responsible for cognition, understands the pre-processed data as information. During the course of information perception, the visual cortex performs its duty extremely fast. Due to the efficiency of visual data processing in the human brain, we can see and recognize things with very little effort. That is, much less cognitive load is required when we interpret visual elements of information presented on web pages compared to textual information.

Related work To understand how humans perceive visual information, Schmidt et al [78]. studied the correlation between web page aesthetics and performance in use in terms of the perceived usability of a web page. Their experimental results show that both technical performance and aesthetic factors are relevant to web page design considerations from their recent study. One drawback of this study is that there is no precise definition nor quantitative metrics of aesthetic factors. More effort is still necessary to establish a standard of measurement for aesthetic factors.

Limited research has been carried out on the impacts of visual elements on information credibility. For instance, Kensicki [79] examined visual factors that affect perceived credibility of non-profit organizations web pages through their simulated user experiments. They found that structured symmetrical design has less influence on consistent credibility perception compared to organic and asymmetrical design selection. Interestingly, the experiments show that photographs and bright or warm colors have a contribution to building more credibility for non-profit web sites. This study shows two potential factors from their observation. First, the visual layout of a web site has less impact on perceived credibility. Second, image and color have not significant, but meaningful, impact on credibility perception. However, their findings are limited to a particular context: non-profit organization web site. This study provides a useful guideline for human subject experiment design for credibility perception.

A more recent study by Can et al. [80] examine the role of images in information diffusion on microblogs by analyzing collected Twitter dataset. They focus on predicting retweet counts using visual cues, which is the linked images to individual tweets. The results show that baseline features (content and structure-based features) can be improved by adding visual cues in terms of its accuracy in prediction of retweet counts. In this experiment, three machine learning algorithms—SVM, linear and random forest regressions—are used. This study implies that visual cues are a potentially useful indicator of information reliability when the power of influence (information diffusion) is considered as a key attribute of quality assessment. It is also a good example of reliable information on the Social Web.

2.5 Reliability Metrics

In Section 1.3, we provided an overview of different attributes of information reliability. Sometimes these attributes are deemed as elements of information in general or credibility [34]. In this dissertation, however, these elements play a role as information reliability metrics. As we discussed different aspects of information examined by a number of studies in Chapter 1, information on the Web have multi-faceted unique characteristics varying across contexts encountered by a user. Different criteria such as the type of information seeking task, user’s objective and personal background must be defined to retrieve most reliable information that satisfy one’s preference. Understanding a user’s intent aside, we need to understand some essential elements of information, called *reliability metrics*. From many studies that attempted to achieve this goal, we discuss (1) what they are and (2) how they can be measured or modeled, and (3) effective methods that predict or recommend particular metrics of information reliability.

On the popularity and importance of social web application as well as the upsurge of

its popularity, a huge amount of research effort has been made on modeling and predicting information reliability in the Social Web. A lot of research effort has focused on a popular microblog service, Twitter, due to both the popularity of the service and the ease of access to datasets. We selected a handful of relevant studies (e.g. [81, 29, 30, 50, 82]) below to discuss various techniques and methodologies introduced by many researchers. We closely look into how these studies measured different metrics and the relevant proxies (such as metadata of microblog posts) revealed from experiments.

2.5.1 Credibility

Information credibility is a concept that has received research attention from a variety of disciplines over many decades. Recently, much research has been done in modeling information credibility in the Social Web. As users are allowed to easily publish or share information on social platforms, malicious or false information have also increased. Accordingly, the importance of identifying credible information has emerged from research on information quality.

Credibility Definition There has been many conflicting definitions of information credibility. A notable recent study on the elements of computer credibility by Fogg and Tseng [34] provided a definition of credibility that has been frequently adopted by subsequent research and is defined as *believability*. Fogg and Tseng defined credibility as:

- a perceived quality
- made up of multiple dimensions, mostly *trustworthiness* and *expertise*

In agreement with many credibility scholars (e.g. [83]) They claim that credibility does not reside in an object, a person or a piece of information. Thus, we consider

credibility as the *perception* of credibility. In addition to that, the majority of the studies to date find credibility is comprised of two primary dimensions—*trustworthiness* and *expertise*.

Credibility Models Fogg et al. [84] studied how people evaluate information credibility online and proposed the idea of *Prominence-interpretation theory* which is comprised of the two values: “prominence” and “interpretation” ($Prominence \times Interpretation = Credibility Impact$). Based on such theories, many researchers proposed models or algorithms that either measure or predict credibility of users in different contexts. For example, Kumar et. al [81] build a model of location and topic affinity to identify credible, relevant users in crisis situations. Wagner [85] found that network effects such as retweets are quite ineffective at capturing topic expertise. Castillo et al. [29] and our previous work [30] both propose models for identifying credible sources of news information based on computational models. Other researchers focus on evaluation of credibility and the problem of ground truth. Our work in the study [50] proposed a two pronged approach to gathering ground truth information on the credibility of microblog data by combining manually annotated scores with observed network statistics (e.g: retweets) from the data to achieve a “more stable” estimate of credibility. Collections and clusters of information such as the curated news collections in [82] have not been studied with respect to credibility, and this would be a interesting avenue for further analysis.

Perceived Credibility Most researchers agree that “credibility” that is inherent to an entity and perceived credibility of that entity are not necessarily equivalent. The latter could be viewed as a subjective function of the former. Perceived credibility can fluctuate from (inherent) credibility based on the way in which the entity is represented, and based on the whims of the person making the credibility assessment. Accordingly,

numerous researchers from different disciplines have attempted to identify a set of salient factors that contribute to our perception process. Visual and textual components have also been studied in the same vein [27, 86, 87] to reveal complex relationships between data metadata and context that inform our perception of information credibility.

In recent years, features or cues that affect perceived credibility of information in microblogs have been studied [88, 20]. The recent studies, in general, first select numerous candidate features that are likely to contribute to the assessment processes, and analyze them in both qualitative and quantitative ways through online surveys or user studies. For example, Morris et al. [20] found the disparity between the features considered for evaluating information credibility between search engines and Twitter. Morris also reported a controlled experiment that revealed the features through which users assess information credibility on Twitter. Their study also provided insight and suggestions for interface design to improve credibility perception from the end-user perspective. Perceived credibility has also been compared across different cultural settings by Yang et al. [88] Their study reports on experimental and survey data that compares and contrasts the impact of several features of microblog updates (authors gender, name style, profile image, location, and degree of network overlap with the reader) on credibility perceptions among U.S.(Twitter) and Chinese(Weibo) audiences. Their goal was to design new user experiences which can maximize both *credibility* (as a property of entity) and *perception of credibility* (an end-user subjective function, to which the entity is an argument) of the contents on social media. Perceived credibility can be impacted by personality characteristics such as those modeled by Mahmud in [89].

Sundar [90] conducted a within-subjects experiments ($N = 48$) to investigate the effect of quoted sources in online news stories on how humans perceive information credibility. Their result shows that the participants rated news stories including quoted sources with higher in quality and credibility than the others. Interestingly, they also

found that there is no effect of quoted source presence to the personal preference or representativeness (newsworthiness) of the participants from the experiment. This is one of the early experiments about human perception on reliability of online information.

Credibility for Persuasion Fogg and Tseng’s definition of credibility [34] has been adopted by many studies. For instance, Castillo et al [29] focused on newsworthiness of information based on the definition. Fogg and his colleagues extended the previous theoretical study towards more practical and context-specific aspects of perceived credibility (e.g. website credibility perceived by users” [87]). Their large-scale online user survey involved more than 2,500 subjects and analyzed their self-reported comments on the assessment of the credibility of two live websites. Quantitative factors of information credibility were extracted from the comments of the respondents. The implication of the result was that the half of the participants (46.1%) prioritize design look of the website when they assess credibility of websites. Fogg et al. investigated credibility perception of website users in order to better understand how *persuasive systems* work as the aim of their study. About the same time, Fogg studied further about how people assess credibility and established his own theory, namely *Prominence-Interpretation Theory*. According to this theory, there are at least five factors that affect prominence of a user: *Involvement*, *Content*, *Task*, *Experience*, and *Individual differences*. Interestingly, some of these factors are highly correlated or identical to the attributes that we will further discuss in the subsequent chapters. For instance, *Involvement* is often interchangeable with *user engagement* and, *Experience* is also an element of *expertise* (see Dreyfus model [91]). Likewise, most of the attributes or factors studied by researchers are considered as high-level element in the system hierarchy or a sub-component of the others.

Credibility and Trust on the Web Research on trust and credibility in a social context has been popular for many decades, from Kochen & Poole’s experiments [92] and Milgram’s famous small worlds experiment [93], trust has been shown to play an important role in social dynamics of a network. With social web API’s, researchers now have many orders of magnitude more data at our fingertips, and we can experiment and evaluate new concepts far more easily. This is evident across a variety of fields, for example, social web search [94], semantic web [95] [96], online auctions [97] [98, 99], personality and behavior prediction [100, 101], political predictions [36] and many others.

Credibility on Twitter Scale, network complexity and rich content make twitter an ideal forum for research on trust and credibility. Some approaches, for example [102] rely on content classifiers or the social network individually, while others harness information from both sources. Canini et al. [103] present a good example of the latter, to source credible information in Twitter. As with the methods in this paper, they concentrate on topic-specific credibility, defining a ranking strategy for users based on their relevance and expertise within a target topic. Based on user evaluations they conclude that there is “a great potential for automatically identifying and ranking credible users for any given topic”. Canini et al. also evaluate the effect of context variance on perceived credibility. Later in this paper, we provide a brief overview of a similar study performed on our data, correlating with the findings in [103] that both network structure and topical content of a tweet have a bearing on perceived credibility.

Twitter has been studied extensively from a media perspective as a news distribution mechanism, both for regular news and for emergency situations such as natural disasters for example [29, 22, 55]. Castillo et. al. [29] describe a very recent study of information credibility, with a particular focus on news content, which they define as a statistically mined topic based on word co-occurrence from crawled “bursts” (short peaks in tweet-

ing about specific topics). They define a complex set of features over messages, topics, propagation and users, which trained a classifier that predicted at the 70-80% level for precision/recall against manually labeled credibility data. While the three models presented in this paper differ, our evaluation mechanism is similar to that in [29], and we add a brief comparison of findings in our result analysis. Mendoza et. al [22] also evaluate trust in news dissemination on Twitter, focusing on the Chilean earthquake of 2010. They statistically evaluate data from the emergency situation and show that rumors can be successfully detected using aggregate analysis of Tweets. Our evaluation of Follower / Following relations from our crawled data (shown in Figures 4.1 and 4.2 yields a very similar pattern to their result.

Credibility in Recommendation Recommender systems have been the focus of research attention for many years, and reputation metrics (such as credibility) [104] have been shown to play an important role in the process of content prediction. They can be applied in social filtering to augment user similarity metrics in the recommendation process. [104]. They have also been shown to increase robustness of prediction algorithms in cases where bad (malicious / erroneous) ratings exist [105, 106]. Models that include explicit distrust have recently been shown to produce better predictions, for example, Victor et. al [107] highlight the advantage of combining trust and distrust metrics to compute predictions over multiple network paths, while a recent study by Golbeck shows that distrust metrics can be used to predict hidden trust edges in a network with very high accuracy [108]. In this paper, we are not propagating credibility values around the network, or computing direct interpersonal trust at the dyadic level, however, the authors believe that distrust metrics can potentially improve credibility predictions in Twitter.

2.5.2 Competence

A brief summary of human competence was provided in Section 1.3. Recently, a handful of research effort has been made to identify unique features for social media analytics and building models to predict various facets of human behavior. Twitter has a unique combination of text content and underlying social link structure, in addition to a variety of dynamic or ad-hoc structures, making it ideal for the study of information credibility and competence of an information provider. Common methods for data mining in Twitter can be loosely classified by the type of data that they operate on.

- *Content-based Methods* generally rely on the text and other metadata in a message to make assertions about information or users. For example, trust, credibility, competence of the author etc. These methods can be quite scalable, since they require only a single API query per assertion. Examples include Canini et al. [103] Kang et al. [30] and Castillo et al. [29]
- *Network-based Methods* generally rely on analysis of the underlying network structure to make decisions about information quality. Examples include Zamal et al. [109]. Network based methods can be slower and less scalable since they potentially require many API queries to make assertions about a single user or message. *Dynamic* network analysis methods, such as retweet analysis can be even more computationally expensive and less scalable, since they focus on information flowing through an already complex network.
- *Hybrid Methods* combine facets from content and network-based approaches. Examples include Sikdar et al. [49], O'Donovan et al. [110] and Kang et al. [30].

Canini et al. [103] present a good example of content-based analysis of messages in Twitter, they concentrate on modeling topic-specific credibility, defining a ranking

strategy for users based on their relevance and expertise within a target topic, using Latent Dirichlet Analysis. Based on user evaluations they conclude that there is “a great potential for automatically identifying and ranking credible users for any given topic”. Canini et al. also evaluate the effect of context variance on perceived credibility.

Twitter has been studied extensively from a media perspective as a news distribution mechanism, both for regular news and for emergency situations such as natural disasters, and other high-impact situations [29, 22, 111]. For example, Thomson et al. [111] model the credibility of different tweet sources during the Fukushima Daiichi nuclear disaster in Japan. They found that proximity to the crisis seemed to moderate an increased tendency to share information from highly credible sources, which is further evidence for our earlier argument that credibility models in Twitter need to account for and adapt to changes in context. Castillo et. al. [29] describe a study of information credibility, with a particular focus on news content, which they define as a statistically mined topic based on word co-occurrence from crawled “bursts” (short peaks in tweeting about specific topics). They define a complex set of features over messages, topics, propagation and users, which trained a classifier that predicted at the 70-80% level for precision/recall against manually labeled credibility data. While the three models presented in this paper differ, our evaluation mechanism is similar to that in [29], and we add a brief comparison of findings in our result analysis. Mendoza et. al [22] also evaluate trust in news dissemination on Twitter, focusing on the Chilean earthquake of 2010. They statistically evaluate data from the emergency situation and show that rumors can be successfully detected using aggregate analysis of Tweets.

While identification of indicators of human-behavioral features such as competence and credibility is an important task, it is also important to consider the end-user’s *perception* of them. Morris et al. [20] performed a study to address users’ perceptions of the credibility of individual tweets in a variety of contexts, for example, from socially

| Attribute | Feature | Example |
|-----------------------|---|---|
| gender | language use (stylistic features: pronouns, determiners, prepositions, quantifiers, conjunctions, etc.) | <i>traditional text</i> [112, 113], <i>blog</i> [114], <i>email</i> [115], <i>user search query</i> [116, 117], <i>review</i> [118], <i>Twitter</i> [119, 109], <i>Facebook</i> [120] |
| message location | message/web content, search query, | [121, 116, 122] |
| regional origin | message text, user behavior, network structure | [119] |
| profile age | search query, profile description | [116, 119, 109] |
| political orientation | message text | [109, 123, 119] |

Table 2.1: Common demographic attributes used in Twitter mining algorithms.

connected and unconnected sources, e.g., in blogs [114], email[115] and search [116, 117]. From the results, Morris et al. derive a set of design recommendations for the visual representation of social search results.

Demographics play an important role in understanding information quality in Twitter. Table 2.1 presents an overview of key user-based attributes that researchers tend to rely on. In this table, attributes are shown on the left, example features for each are shown in the middle column, and the research papers that employ the features/attributes are given in the right column. For example, [124] conducted a simple survey on the application of features which can be used for analyzing people’s profiles on the style, patterns and content of their communication streams. Herring [112] investigate the language/gender/genre relationship in web blogs and show gender-related stylistic features from diary and filter entries. Incorporating occurrence of words and special characters based on pre-defined corpora is another type of feature selection. For example, [125] use simple nominal or binary binary features to classify tweets into different categories such as news, temporal events, opinions, deals or conversations. [121] propose a proba-

bilistic framework for content-based location estimation using microblog messages. The framework estimates each user’s city-level location based purely on the message text without any geospatial coordinates, while [119] apply stacked-SVM-based classification algorithms for their classification task on a Twitter dataset. Since we are interested in creating mappings between existing models of human behavior and the Twitter network, understanding these different features, methods and their performances is a critical first-step.

2.5.3 Influence

Only a few users in a social network have an important role with regards to the information flows. Social marketers like to target these high influential users and harness their impact in the network.

To date, many researchers have been working on modeling user influence in social network in order to achieve effective and efficient influencer marketing. For example, Klout¹⁰ developed their own PageRank-like algorithm which computes user influence metric for their service.

However, most of the existing influence models or algorithms rely on user popularity metrics or activity logs such as URL clicks and followings. Furthermore, these models yield a single-dimension feature (score) that is difficult to reflect different user behaviors in information flow. Another limitation of these approaches is that the aforementioned algorithms require to obtain a snapshot of the entire network before evaluating each user’s influence.

Based on an investigation of real-world datasets that we obtained from Twitter®[®], we observe different roles among influencers in social networks. We compile a schema that comprises three major roles that are relevant for common influence marketing strategies

¹⁰<https://klout.com/home>

and model each role with significant features identified in our analysis.

Influence have been studied for decades in many disciplines including psychology [126] and political sciences [127]. However, in the past few years, as people get engaged more and more in microblogging services, this topic has received unparalleled amount of attention from both academia and industry due to its significant impact on the society beyond cyberspace.

The recent studies [128, 129] report on the findings derived from a large scale data analysis. For example, Cha et al. [128] provide an in-depth comparison across three influence measures: indegree, retweets and mentions using a collection of Twitter data. One of their interesting observations is that users who have high indegree (equivalent to number of followers in bi-directional network such as Twitter) are not always influential in terms of information flow. However, Bakshy et al. [129] shows that the largest cascades in information flow tend to be generated by influential users and they are likely followed by many users. This result contradicts the previous claim.

There have been several influence models focused on modern social network. For example, Bao and Chang [130] proposed AdHeat algorithm which considers user influence and relevance to match ads for targeted users on social network. Their approach harnesses hint words and influence propagation based on the assumption that individual user's level of activity and authority has impact on information dissemination. The proposed model is a combinatory metric of the two factors, however, this algorithm does not capture context-specific influence. Furthermore, this algorithm can only be applied to a snapshot of overall network structure of individual user, which makes it difficult to incorporate it into a real-time detection/monitoring system.

In the recent survey about influential user detection algorithms by Singh et al. [131], different techniques for identifying influential users are provided. Various approaches such as Markov random fields, random walk, network topology and simple features like

number of followers or posting time on which the prior art base their approaches are introduced in the article. However, again, only small number of the approaches support context-specific or real-time computation settings and this can be deemed as a downside of those approaches.

2.5.4 Relevance

In the context of information seeking, relevance of information is a good indicator for different information quality metrics such as timeliness, authoritativeness or trustworthiness of items. In many cases, reliability of information also implies reliability of items, particularly if a task is built on a specific topic of interest. For this reason, relevance has been studied for many years by researchers as one of the primary factors to evaluate information search processes.

Relevance of information has been widely studied as an important metric in entity search task. As opposed to web search in which results are web pages, entity search provides a more semantically cohesive view of information with the results being people, organizations, places, etc. The problem of discovering interesting relations from unstructured text has led to a surge in research on entity search [132, 133, 134, 135, 136]. To extract entities from raw text, the common approach is to map text to a Wikipedia page, which signifies an entity. Now, given a search query, we retrieve other entities relevant to it by first building an entity network [137, 138] based on a pairwise entity relevance score [139, 140] and by then applying random-walk computations on this network [139, 140].

A work closely related to ours is the one of [141], who studied composite retrieval in the context of aggregated search – where results from different verticals available on the Web (image, video, news) are returned to users. They develop several algorithms,

treating relevance as their main criteria to construct bundles, and cohesion and diversity as secondary. To tackle the challenges arising from the heterogeneous nature of the data, they exploit entities to link relevant results across verticals. They also incorporate query intent into the formation of bundles.

Our work complements theirs, as we focus on entity search and investigate how composite retrieval promotes exploratory search in this context. We also do not put relevance as our first criterion to form the bundles, as we know from our previous work that relevance and interestingness are different criteria.

2.5.5 Newsworthiness

More recent and rather widely accepted definitions have been proposed by Shoemaker (2006) [142]. According to her earlier study with Cohen in [143], newsworthiness and news do not refer to the same notion, unlike one of the common assumptions that are widely accepted by people. She asserts that “news is a social construct or a thing whereas newsworthiness is a cognitive construct.” Many studies in journalism and communication link newsworthiness to the practice of news gatekeeping in traditional media. For example, Shoemaker et al. [144] and Diakopoulos et al. [42] employ a rather simple definition of news—what’s worthy of sharing—in their studies. A more recent study by Sundar et al. studied perceived newsworthiness entangled with credibility on newsbot services such as Google News. They restricted their study in the context of information overload and how different news cues (e.g. name of the primary source, elapsed time and number of related articles) affect users during news consumption.

2.5.6 Interestingness

Many disciplines such as data mining [44], knowledge discovery in databases (often referred to as “changes and deviations”) [45, 46] and information retrieval [47] have studied *interestingness* of information over the last few decades. Naveed et al. [47] found that users likely to retweet “when they find a message particularly interesting and worth sharing with others.” Following the reasoning, they assume that retweet is a proxy of interestingness, and thus, it can be used as a function of interestingness to build a model that bases on the content-centric characteristics of retweets. The study provides insights into the factor that affects retweeting behaviors that reflect interestingness of content. In microblogs, content metadata that represent associated topics (hashtags) have also been used to model interestingness of information [145].

2.6 Communicating Reliable Information

Several visualization frameworks have been designed and implemented for the purpose of analyzing social media information. However, most visualization tools provide visual information based on post-hoc data analysis, in particular, statistics or rankings on off-line datasets, previously collected by another process. In this section, we introduce relevant works from the literature in order to compare them with our proposed visualization systems, see Chapter 6, (*TweetProbe*, *GeoProbe* and *HopTopics*). In this section, we review the literature from two different perspectives: Social Stream Filtering and Analysis on the one hand, and Real-time Visualization on the other. As a literature review, we discuss key contributions made by recent related studies and highlight differences between our approaches and these studies. We will discuss our work more in detail in Chapter 6 later.

2.6.1 Social Stream Filtering and Analysis

As our visualization technique, particularly *TweetProbe* (see Chapter 6) is based on the real-time social data stream, similar approaches to ours have been proposed by a few researchers. For instance, [146] develop a framework which collects microposts that contain media items, shared on social platforms like Twitter, Facebook or Instagram. As a result of a query, this framework returns the resulting images or video clips that are relevant to the query in various ways such as timeline, graph and narrative visualizations. Particularly, they take a storyboard approach which automatically curates shared information about a specific social event.

An interactive visualization based on Twitter streaming data was also proposed by [147]. They present a system called “TwitterMonitor” which performs trend detection over the Twitter stream using the Twitter Streaming API. This is a web-based framework which heavily relies on user interaction such as manual ranking or user-provided description for each trend. However, they only provide a simple chart showing topic popularity over time for each trend and it is mainly targeted as a text-based search framework. Another example of Twitter stream filtering is [148] which apply a user profiling approach based on a user’s posted URLs using topical categorization. The topics obtained from this algorithm are then used to filter tweet streams for extracting more relevant information from their followers.

Social stream filtering can also be performed on a collaborative environment. For instance, [149] propose an intranet system that shows the results of faceted search tasks in real-time. Their system takes the enterprise activity stream as input data and returns relevant results via a small visualization module on the web page. In this work, both sentiment and topical visualization approaches are also used along with tag clouds.

2.6.2 Real-time Visualization

Most of the real-time visualization techniques in the literature have been focusing on network intrusion detection (IDS) [150, 151, 152] or infrastructure monitoring¹¹. IDS is one of the representative systems in the field of Cyber-Security Situational Awareness. Since timely alerts are a crucial factor in an intrusion detection system, real-time visualization is an essential feature in this application. However, all of these visualizations lack of aesthetic factor, simply visualize the entire topology of a network in real-time. Although none of the systems employ major design consideration on visual components of their visualization, simple interactive interfaces are supported in general.

Another work “We Feel fine” [153] should be noted here although this work is not fully based on real-time data streams. This work shows various emotions emerging through an emotional search engine, which can be seen as web-based artwork. The authors categorize each web content crawled from various information sources such as blogs and web sites into pre-defined emotion classes and combine them with corresponding meta-data (location, demographic information etc.). Each content element is mapped to a color-coded particle and users can filter them through an interactive web interface. The authors carefully considered aesthetic factors in their visualization.

2.7 Conclusion

In this chapter, we have provided related work on modeling and communicating information reliability. In particular, we have discussed a range of definitions, different methods and algorithms, and various applications used by other researchers. From the next chapter, we will dive into our main studies on understanding and modeling information reliability. For example, in Chapter 3, how humans perceive information credibility

¹¹<http://www.francastillo.net/>

on the Social Web, particularly on microblogging platforms, will be discussed in detail through a range of interlinked studies.

Chapter 3

Perception

3.1 Introduction

In Chapter 1, we discussed the information quality framework (Wang [14, 15]) on which our study builds; how we adapt the framework to the Social Web; and the mapping between the framework and the conceptualization of information reliability. Wang [14] described the chain of information quality assessment process as “High-quality data should be intrinsically good, contextually appropriate for the task, clearly represented, and accessible to the data consumer.” This also means how humans perceive information can be partially attributed to the intrinsic factors of the information. However, separate from the intrinsic aspects of information, how a human perceives it is another significant factor that defines the reliability of information. For example, presentation quality of information on the web is an important factor in the context of health [154] and other search practices [33]. Furthermore, particularly on the Social Web, credibility of source must be considered along with the intrinsic and extrinsic factors [35].

In this chapter, we study how humans perceive information reliability to understand the underlying patterns and mechanisms. To this end, we look into the dominating

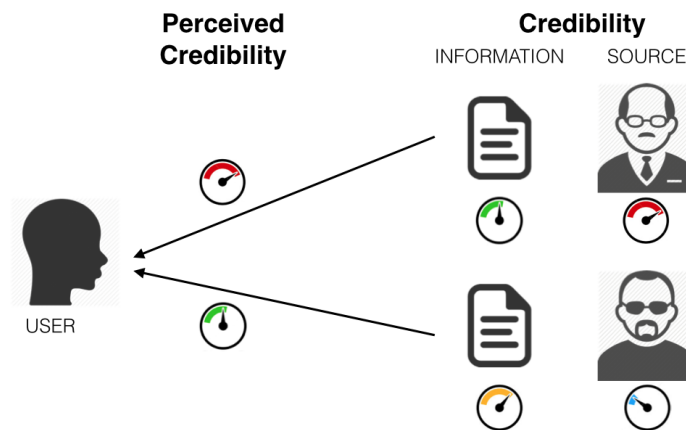


Figure 3.1: Perceived credibility directly originates from the synthesized credibility of source and information

factors that affect users' information perception. An initial user study (N=81) and two online experiments (N=102, N=646) are presented to show the common pattern and discrepancies across different social platforms and topics.

3.2 Human Perception

Imagine you are searching for information on the Web. Whatever the motivation is, you need to assess one or more factors that affect reliability of information of your interest such as source, content and how the content is presented on these web sites you sift through the search engine. The purpose of this search task decides how carefully you select the information amongst them. However, can you always guarantee the reliability of your assessment and final choice in your search task? If not, what do you think are the factors that made you to judge so? There are several factors that help or interfere with your assessment. In this section, we discuss these factors and how they affect human assessment procedure in both negative and positive ways.

3.2.1 Information source

Information *quality* can imply different aspects of information, depending on the context. For example, if you are searching for information for your news story as a journalist, newsworthiness and credibility might be the central criteria of your judgment. If you need to write a story in entertainment section, the importance of credibility might be lowered and interestingness will be added. However, there is another element which captures your attention in your task. That is the “source” of information.

Referral trust In trust study, information source is considered as a good or bad agent acting as referral trust [155, 156, 157]. In other words, reputation of the source which has been established by other users in the network or community affect your credibility judgment (trust) on the information. Detailed reviews on trust studies can be found in the previous chapter in Section 2.5.1.

As shown in Figure 3.1, in the case of information credibility assessment, the information will be evaluated with the perceived credibility which is a synthesized metric reflecting both information and source credibility. For example, the first information in Figure 3.1 has moderate level of credibility, however the perceived credibility of this information is boosted with the source credibility. Please note that Figure 3.1 is a conceptual diagram and shows a possible example of credibility assessment on online information.

Personal experience As with referral trust, which is provided by other users or agents in the network, there is another type of reputation-based trust: trust based on past personal experience with the information source (in both direct and indirect ways.) If a user has face-to-face interaction with the source, this experience will provide a significant impact on the credibility assessment task. Indirect interactions or experiences such as prior evaluation made by the user on the other information authored by the source

will be a good indicator as well. Jonker et al. [158] studied how negative and positive trusts are turned into positive and negative trusts from their experiments with human subjects. This study shows that prior experiences can change how humans perceive source reliability. Since people exhibit a tendency to equate source with contents authored by the source, information source is a non-negligible part in information reliability assessment.

Source reliability In many cases users do not have face-to-face personal experience with information sources in search tasks. This makes them difficult to evaluate legitimacy of information and its source. If the source is an individual person, a user can look for the distributed opinions on the reputation of the source. To make this validation process easier, some online systems provide popularity or reputation scores of users or contents online. For instance, Google Scholar¹ provide citation counts of academic articles and many researchers utilize this as a quality or popularity metrics for their search task. Moreover, Klout score² has been developed and is being served as a integrated popularity metric for social network accounts.

Organizations and authoritativeness Organizational source is another type of information provenance. News media, big and small business enterprises, non-profit organizations and governments are good example. Typically, this type of information source publish plenty of information on a regular basis, and thus, it is comparatively easier to assess reliability of the information. For this reason, there are some malicious attempts online that mimic these organizational information providers in order to deceive other users and disseminate unwanted information such as advertisements, malware or false rumors. This kind of practice is exponentially increasing in the Social Web such as microblogs (e.g. Twitter, Facebook). To avoid fraudulent schemes and spams, users

¹<http://scholar.google.com>

²<https://klout.com>

often assess authoritativeness of the source. This type of assessments mostly involve inspections on the legitimacy of contents and tackiness of visual elements such as logo, color, fonts, etc. In brief, representation of information is a crucial factor in assessing authoritativeness, i.e. credibility of organizational information sources (also pertain to individual sources.) Role of representation in information reliability assessment as well as some important (visual) elements are discussed in detail in the following section.

3.2.2 Visual cues and perception

Final assessment of information reliability is made when we actually perceive the information through our sensory system, as can be seen in Figure 3.1. The majority of information entering into our brain is processed by human visual system. For this reason, visual attributes (stimuli) play an important role in assessing information reliability on the Web. In fact, visual perception, which is operated by the rear part of human brain, is faster and more efficient than thinking (cognition) process handled by the front part, namely cerebral cortex [159].

Humans acquire visual patterns of reliable information on the Web through empirical learning and this has been investigated by many researchers in the recent years [32]. For example, websites presented in a balanced layout with clear font faces such as Helvetica, Arial and Times New Roman are perceived more credible [27].

Figure 3.2 shows how presentation of online information contributes to perceived information reliability. For example, a viewer perceives *information A* in Figure 3.2 as moderately credible since it is awkwardly presented on the web, albeit the source is considered highly credible. On the other hand, *information B* is compensated with high credibility in presentation although its source has comparatively low credibility. Again, Figure 3.2 is a conceptual example of perceived credibility as a composite attribute based

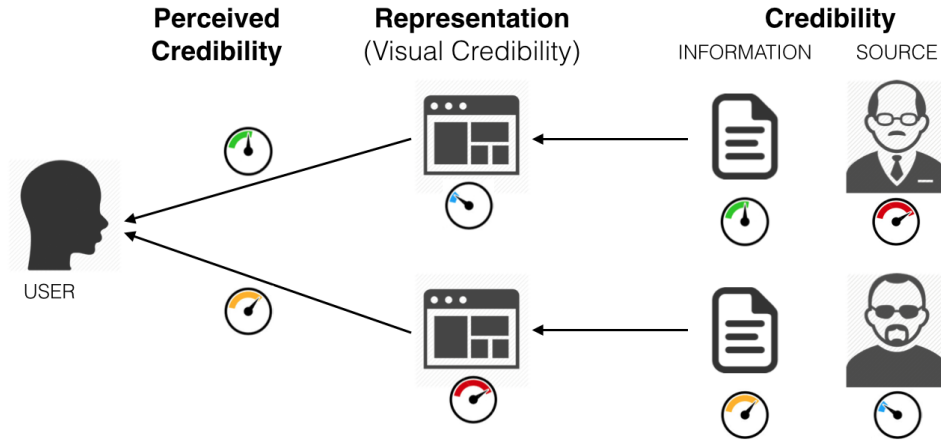


Figure 3.2: Perceived credibility through different visual representations of source and information

on multiple credibility factors on the Social Web.

How visual cues affect human perception on information reliability Importance of visual cues such as design look [87] and features (detected object and color histogram) from the images linked in tweets [80] have been studied in recent years. Then, how these cues affect perceived credibility? To answer this research question, particularly with information on the Social Web, we conducted a large-scale online user experiments. In brief, from this study, we found that there are some independent factors that contribute to human credibility perception in the context of social web.

We will further discuss how human users perceive information reliability on the Social Web from our recent study based on a preliminary online survey and two subsequent human experiments in Section 3.3.

3.2.3 Impact of user interface design

Traditional web pages

Both psychologists and computer scientists who study about human perception or human-computer interaction have focused on perceived quality of information. The quality here can be deconstructed into multiple attributes such as credibility or authoritativeness of information, or source of information. The research has been active to date because there are lots of factors and features we can evaluate on the web page interface. For instance, as with the quality of textual information, visual attributes such as layout of the web page, color usage, and font are also considered as the most significant factors that contribute to the perceived quality of information.

BJ Fogg et al. [87] reported on their large-scale experiment which involved 2,684 internet users. This study analyzed participants' comments to find important features users notice when a Web site is evaluated for credibility. In this study, two live websites on a similar topic have been shown to the participants. The results imply that design look (46.1% participants) has the highest impact on their credibility assessment, followed by information structure and information focus. However, this study simply relied on self-reported comments and, thus, lacks quantitative evaluation.

As we mentioned before, to measure both quantitative and qualitative reliability of information, we take into account all of these important attributes. This will be discussed in detail in the subsequent chapters.

The Social Web

We have shown some recent studies about how users perceive information reliability on traditional web pages, focusing on specific contexts or topics. Since we primarily focus on information on the Social Web, we briefly discuss the different nature of perceived

reliability between the traditional web pages and the social platforms such as microblogs and social networking services.

Arguably, social platforms exhibit different aspects in content, source, and visual representation from traditional web pages. Basically, these discrepancies originate from the difference of information provenance between them. First, information on traditional web sites are represented by the source of information, namely the owners of such web sites. Thus, typical users assume that both contents and visual representation, for instance aesthetic design, layout, user interfaces, etc., are provided by the owner of the web site they visit. On the other hand, normally, social web pages provide the whole infrastructure to their users, including standardized framework for representation of information. In other words, users of such social network or platform do not have a control on the representation of contents that they create or share. For this reason, when other users assess information reliability on the Social Web, their evaluation task is inevitably limited to certain number of features or cues. For instance, when a user assess information credibility on Twitter, they often focus on content, profile image, metadata such as retweet symbol, mentioned user name, or included multimedia contents (e.g. video clip or still image).

Interestingly, this characteristic of social platform³ is a double-edge sword for evaluating information reliability. Since the Social Web platforms limit both content (e.g. 140 character limit on Twitter) and visual features, it is comparatively trivial to apply them to computational models and machine learning algorithms. However, since the provenance of individual content is not evident, users (or computer algorithms) need to track the origin of information by following retweet chains.

We have studied about the discrepancy between traditional and social web pages. In

³In this dissertation, social platform is interchangeably used with social application, social media and social network.

order to deal with the difficulties caused by the ambiguity of information provenance in social applications, we need to reduce the dimension of the feature space before applying any computational models. In brief, various dimensionality reduction techniques can be applied to solve this type of issue. We will discuss this in more detail in Chapter 4.

3.3 Experiment: Credibility Perception on Microblog Contents

In this section, we present an experimental study which investigates the impact of individual attributes on *credibility perception in microblogs*. Specifically, we report on a demographic survey (N=81) followed by the two user experiments (N=102, N=646) in order to answer the following research questions: (1) What are the important cues that contribute to information being perceived as credible. (2) Can we separate these cues from the content and quantify their contribution?, and (3) To what extent is such a quantification portable across different microblogging platforms? To answer the third question in particular, we use data from *Reddit* and *Twitter*. Key results include that significant effects of individual factors can be isolated, are portable, and that links, profile pictures and image content are the strongest influencing factors in credibility assessment.

3.3.1 Introduction

Microblogging platforms such as Twitter and Reddit are increasingly relied upon for real-time news and a broad range of other information. However, all platforms that support user-provided content, including microblogs contain a large amount of noisy and unreliable content. Microblogs have emerged as serious sources for news at a global level, as indicators and early informers of natural phenomena such as earthquakes and severe

weather, and even as mechanisms for financial transactions. For instance, tweet-to-pay functionality was recently introduced by several banks⁴. Accordingly, the importance of identifying credible information and information sources on microblog platforms continuously increases.

History Traditionally used as online journals or specialized peer-communication platforms [160], microblogs have transformed and proliferated into powerful online information sources operating at a global scale in every aspect of society. This is due in part to the advance of mobile technologies, including audio and video streaming, speech recognition and others. These technologies enable people on-location at an event or incident to serve as news reporters [161]. Several recent studies [22, 162, 163] show how user-provided microblog content is an effective mechanism for understanding crisis situations such as earthquakes, hurricanes or political conflicts. In fact, a recent study of traditional media journalist practice [164] shows that they rely heavily on social media for their information. Another study [12] reported that 53.8% of all U.S. journalists use microblogs to collect information and to report their stories, which of course raises the potential issues of cyclic dependency and rapid propagation of misinformation. These are issues which bolster the need for better approaches to information credibility on the web.

Evolution The mass proliferation of microblog usage also brought about a shift in the interaction mechanisms and information flow within the platforms themselves. In particular, a 2013 PEW research report [165] shows that an increasing number of users search microblogs by keyword or hashtag as opposed to the traditional content stream or message exchange practices. This means that a larger portion of information is coming from complete strangers, accessed via keyword matching than from sources that a

⁴Example available at <http://www.paywithatweet.com/>

user is actively following and are likely to be known by the user. This reduced window of information about the source presents a difficult challenge in assessing credibility of information, and requires a more comprehensive understanding of the components of a microblog message and their potential impact on the information consumer’s assessment of information credibility. This is bolstered by the fact that the majority of users in Twitter (52%) and Reddit (60%) treat the system as their primary source of news information [165].

Understanding Recently there have been many efforts to study information credibility in microblogs, ranging from automated algorithms to model and predict credibility of users [81, 85] and messages [29, 30] at general [29, 49] and topic-specific [30] levels, to visualization and interaction applications such as [166, 167, 168]. However, with the exception of [110, 88], little research has focused on isolating the impact of individual microblog features such as profile images, links and other available metadata on perceptions of credibility—a problem that is increasingly important as a growing portion of news information gets produced by people the information consumer does not know.

Contributions Now that we have discussed a high level motivation of the problem, including history and common methods to evaluate information credibility in microblogs, we turn to the key contributions of our work, within the above context. In particular, we propose the following three research questions, noting that the first question has been partially answered through our earlier study in [169]. We include the research question from [169] and a short discussion of their experiment since they are critical to understanding the user experiment [N=646] in this paper.

1. What are the important cues that contribute to information being perceived as credible on microblogs?

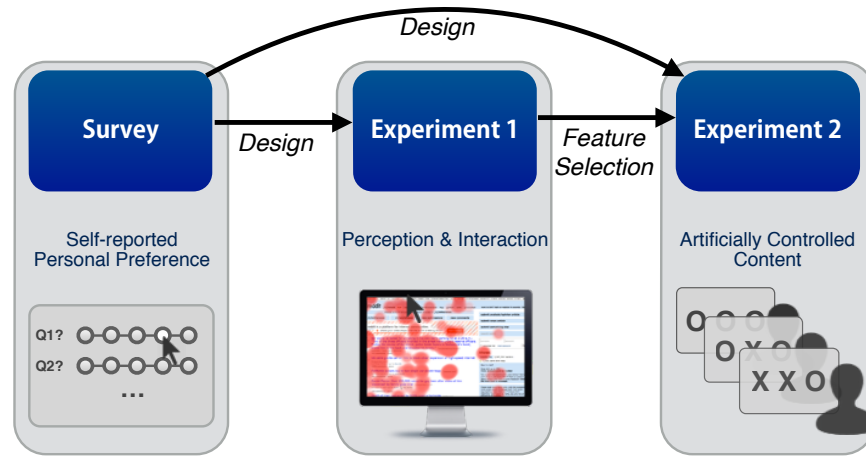


Figure 3.3: Overview and dependencies between the initial survey (N=81) [169], Experiment 1 (N=102) [169] and Experiment 2 (N=646) reported in this paper.

2. Can we separate them from the content and quantify their impact on a credibility assessment?
3. To what extent is such a quantification portable across different microblogging platforms?

A high level overview of the three key phases in our studies and the dependencies between them is shown in Figure 3.3. First, a crowd sourced survey (from [169]) of 81 microblog users was performed on Amazon Mechanical Turk (MTurk) to assess patterns in user assessments of information credibility across different topics and for demographic groupings. In this survey, the participants answered to 13 demographic and 17 main questions regarding their microblog usage and perceived credibility on microblog messages. By analyzing a large set of microblog features, a set of important factors was identified for further evaluation.

Second, a user experiment (N=102, from [169]) was designed to place users in context across two microblog platforms and elicit a more refined set of salient factors that

influenced their judgement of information credibility. To do this, heatmaps were computed from mouse click behavior in the microblog interfaces. To address the question of portability, we perform experiments across two of the most popular microblogs for news-consumption: *Reddit* and *Twitter*.

Third and finally, we conduct a novel experiment (N=646) in which variables from [169] are experimentally controlled. This allows us to assess the impact of the each individual variable on credibility assessments in a range of contexts. The main findings in this study is described in Section 4.5.2. In summary, by artificially controlling metadata such as number of friends, or the profile image type of an information provider, and gathering human-provided assessments of associated content, we observe a significant shift in reported “information credibility” between the treatments. An example of a high and low value treatment for the “number of friends” feature is the average number of friends in the 5th and 95th percentiles for a randomly sampled batch of 250,000 users from Twitter and Reddit. By ranking each evaluated feature by the absolute distance between mean reported credibility scores on treatments sampled from large samples of real data, we are able to produce a top- n list explaining the relative influence of metadata features on human assessments of information credibility. Surprisingly, our results show that features that can easily be user controlled, such as profile image or a company logo produce a larger shift in credibility assessment than features that are network-controlled, such as number of shares or number of friends, despite the obvious fact that these features are much more difficult to fake.

3.3.2 Credibility Perception Survey

To gather a fair assessment of candidate features to evaluate in our main experiments, and to gather insight about credibility decisions in microblogs, we conducted a crowd

Table 3.1: Primary use of information on Twitter

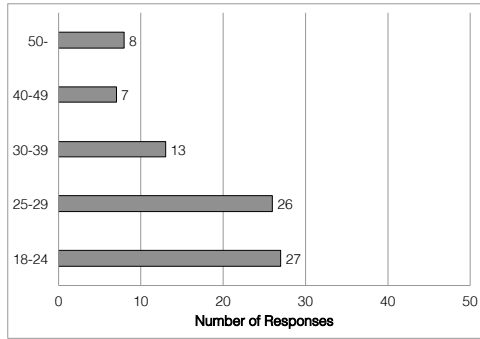
| | | | |
|----------------------|-----------|---------------------|-----------|
| Business | 22(27.2%) | Information Sharing | 21(25.9%) |
| Social Purpose | 16(20.0%) | Information Search | 15(18.5%) |
| Serendipitous Search | 4(4.9%) | Other | 3(3.7%) |

sourced study targeting Twitter users in late 2013 using Amazon’s Mechanical Turk (MTurk) platform. A total of 81 respondents were asked a series of questions to explore what information from microblogs they mostly consider when they need to search for credible information about particular events. The 59 male and 22 female participants were from different parts of the world, with a majority from the United States and India. Participant age ranged between 18 and 60 with an average of 28. 60% of the subjects used microblog on a weekly (22%) or daily (38%) basis.

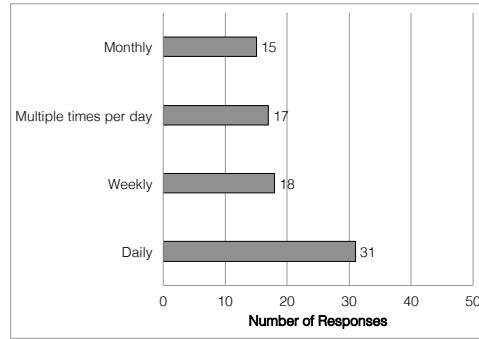
In the demographic questionnaire, participants reported basic information such as Twitter usage, educational and cultural backgrounds and yearly income. For example, 82% of the subjects reported that they hold Bachelor’s degree or equivalent for educational background. The majority of the participants reside in either medium-sized or large cities. In terms of the marital status, the subjects are equally distributed between single and married groups. The results are shown in Figure 3.4.

The overarching goal of the survey was to explore the following general hypothesis through self-reported metrics and to identify the set of Twitter features (E.g: links, profile images etc.) reported as most influential in credibility assessment. The two additional experiments in this paper expand the general hypothesis into 6 additional hypotheses, and were both designed based on analysis of the results from this survey. In the survey, 20% (16/81) of the participants reported that they consider visual cues as a major factor that affects their credibility assessments. Details of the resulting design decisions are discussed in Section 4.5.2.

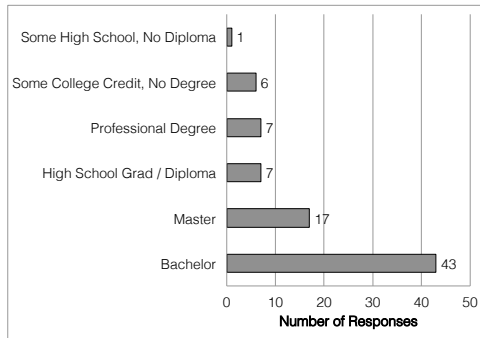
Research Question (RQ) 1: Does the display of metadata in microblogs influence



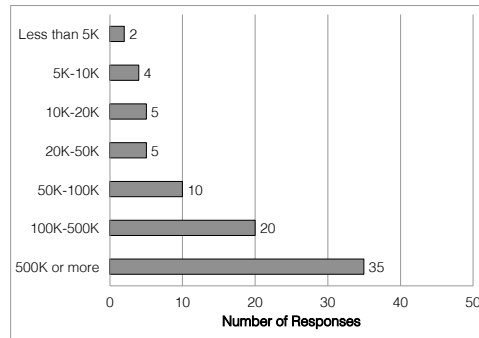
(a) Age



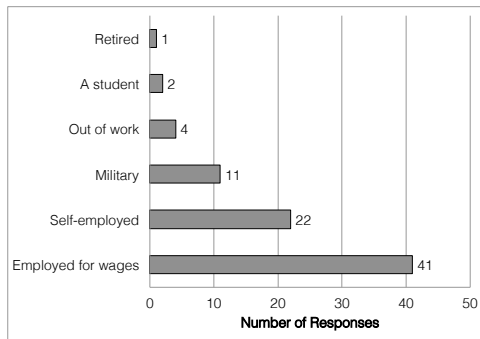
(b) Microblog Usage



(c) Educational Background



(d) Population of Locality



(e) Employment Status



(f) Marital Status

Figure 3.4: Self-reporting responses on the demographic questionnaire from Twitter users

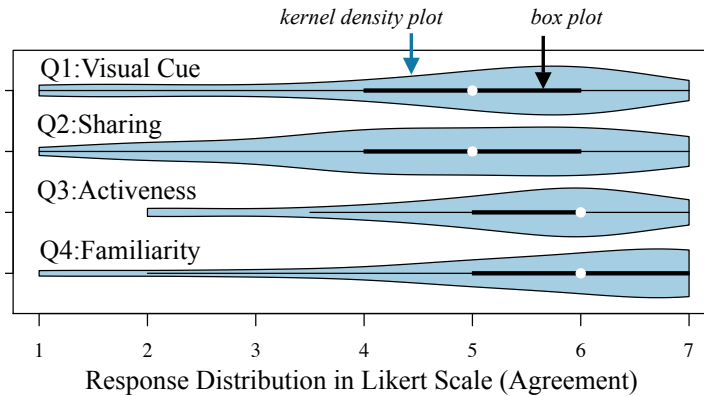


Figure 3.5: Survey responses for key questions (marked with * in Table 3.2. Each line and enclosed curve shows the response distribution for questions in (1) a box-whisker-chart indicating medians and quartiles and (2) a kernel density plot, respectively. Responses are provided on a Likert scale (Strongly disagree (1) – Strongly agree (7)).

perceived credibility of the associated content.

HYPOTHESIS. Metadata display (textual or visual) influences perceived credibility of microblog content. The direction of influence is dependent on the specific content displayed.

To gauge usage patterns and credibility perception, participants were asked 17 questions in a web survey, covering aspects such as activity rate, perceived impact of visual cues, and sharing frequency, among others. A selection of these survey questions are shown in Table 3.2, and response distributions are shown in Figure 3.5. The results indicate that the majority of the participants consider themselves active information consumers on microblogs (79% for 'Activeness'). These users also share their own content frequently with their followers (57%, 'Sharing'). The most common usage reasons (Table 3.1) were reported as business (27%) including online marketing, information sharing (26%) and social use (20%). Interestingly, 68% of participants reported that visual elements have significant impact on their credibility assessment in the microblog, as indicated by Q1 in Figure 3.5. Our population exhibited reasonably heavy use of Twit-

Table 3.2: Survey Questions. * denotes further detail in Figure 3.5

| Label | Questions |
|-------------------|--|
| Activeness* | Do you consider yourself as an active online information consumer? |
| Sharing Freq* | Do you frequently share your information with the people in your network (followers)? |
| Primary Usage | What is your primary usage of information on microblogs? |
| Familiarity* | Are you familiar with microblog services? |
| 1st Cred Factor | Which do you consider as a primary factor for measuring information credibility? |
| 2nd Cred Factor | Which do you consider as a secondary factor for measuring information credibility? |
| Visual Cues * | Do you think that visual cues are important for judging credibility? |
| Url Relevance | Do you think that the presence of URLs in a tweet, which point to an external information source, can enhance information credibility? |
| In-Person Friends | About how many of your "friends" on Twitter have you met in person? |
| Non-Human Friends | About how many companies or organizations do you currently follow on Twitter? |
| Celebrity Friends | About how many celebrities do you currently follow on Twitter? |
| Time-On-Others | On Twitter, about how much time do you spend looking at what other users have posted? |
| Time-On-Me | On Twitter, about how much time do you spend posting tweets about yourself? |

ter (38% daily use) and a good standard of education across participants, with 67/81 possessing at least a bachelor's degree.

Credibility Factors The main section of the survey investigated what kinds of attributes participants consider as a primary factor when they search for credible information on microblogs. We can intuitively expect that both content and information source would be highly ranked and the results indeed support this. However, it is interesting to note that *visual cues* such as design and layout were also reported as influential in the

process of credibility assessment. 10% of participants responded that design/layout was the primary factor (20% elected it a major factor) in their credibility assessments.

Correlation Analysis People consider many different factors during credibility assessment with microblog information. Numerous researchers concluded that, ultimately, credibility can be perceived or measured in different ways based on the given context, cultural background, language, etc. [84]. We also find that many microblog users agree with this statement from pre-study offline interviews. Thus, we designed our questionnaire to find underlying correlation, if any, between demographic background and the question responses. Results are shown in Figure 3.6. Table 3.2 provides a full description of each element in the correlation plot of Figure 3.6. Some notable correlations include Twitter use and general information use. There was a positive correlation between employment type and content use –this may have been a result of the number of users who said they used the microblogs for marketing purposes. There was a strong correlation between locality (size of city lived in) and amount of information shared on microblogs. People in larger cities shared more information than those in small cities and towns. Predictably, employment type was positively correlated with primary usage of the microblog. We also find a correlation between gross income / ethnic origin and microblog usage, complementary to [170], who found strong correlation between these factors and browsing behavior. Demographic factors (both age and cultural background) correlated with usage rate, and with the impression of visual cues as an information credibility factor –younger people had higher usage rates and were more influenced by visual cues.

In summary, the initial analysis from the survey highlights visual cues as a useful factor for further study, incorporating aspects of content, and metadata about the source/provenance of microblog messages.

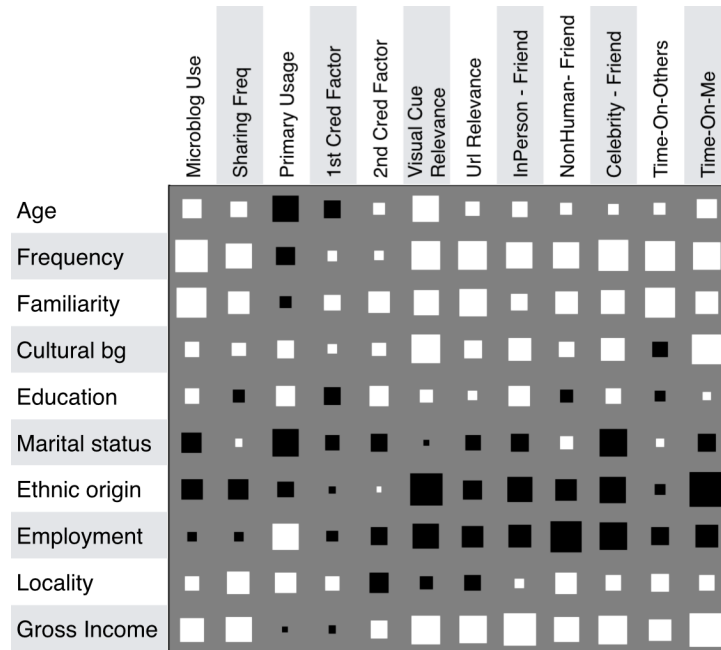


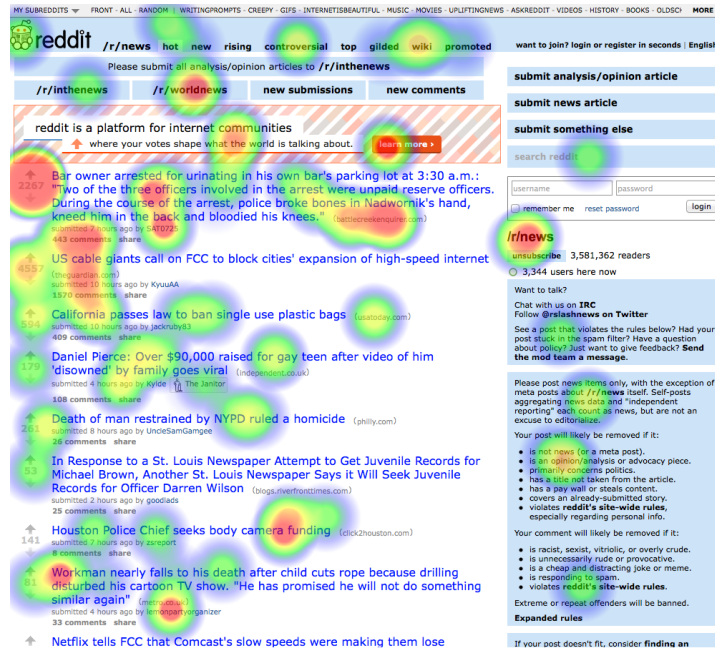
Figure 3.6: Correlation matrix between the demographic information of the participants and the responses to the survey. The matrix is visualized in a Hinton Map (white and black squares represent positive and negative correlation, respectively. Square size is proportional to the absolute value of the score [0–1].)

3.3.3 Experimental Setup

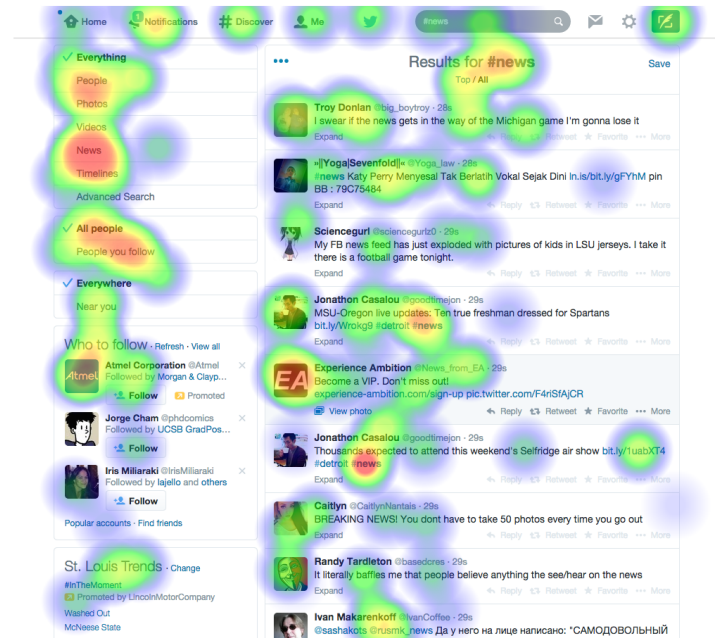
To further explore what factors influence credibility perception in microblogs, we designed and conducted two different user experiments guided by insights from the credibility perception survey. Both experiments were conducted on MTurk. In this section, we detail the design of both experiments (Exp1 and Exp2) and discuss a refined set of experimental hypotheses.

Exp1: Perception and Interaction

The initial survey highlighted that content (meaning) and sources (origin) of microblog posts are the most influential factors in credibility assessment. However, the representation of information such as metadata also plays a significant role in user perception of information credibility. According to these findings, we refine the initial research



(a) News subreddit page in Reddit



(b) A search result page with 'news' keyword in Twitter

Figure 3.7: Heatmap visualization of the user annotations for Exp1, where clicks on an entity indicate perceived credibility.

questions to include the following three questions/hypotheses.

RQ2: Do different features influence credibility by different amounts?

HYPOTHESIS. *Features have varying degree of influence over credibility perception*

RQ3: What are the effects of different classes of microblog features on perceived credibility?

HYPOTHESIS. *Visual factors will have the most influence, followed by network and content-based (text) factors.*

RQ4: Can our models of feature influence be ported successfully to different microblogs?

HYPOTHESIS. *Influence of features is consistent across platforms*

To test these hypotheses, a study was designed to place users in a familiar/typical microblog context and provide them with a simple mechanism to highlight the specific features that they felt had an influence on their perception of content. To address the last hypothesis, the study was designed to be cross-platform, comparing features from Twitter and Reddit. Figure 3.7 shows two example interfaces from the study (N=102). Users were requested to click on or close to items that they felt had *any* impact on their perception of information, regardless of positive or negative direction, which is evaluated separately in our third experiment. They were given no a-priori information on specific feature lists. This mechanism for identifying influential features was used in an effort to avoid bias from manual or expert selection of a feature set (intended for detailed analysis in Experiment 2). Domains (Twitter and Reddit) were a between-subjects variable, and only participants with significant prior experience with a domain were allowed perform the task.

First, participants were asked 6 general questions about their microblogging practice. Then they were shown 3-4 screenshot images of the microblog (3 for Reddit and 4 for Twitter). To capture the aforementioned features avoiding possible bias, we let participants select three visual elements instead of having them rank an arbitrary selection of features we provide. On each click, a slider selector was shown below the image to record the amount of impact the element that the user clicked on has on her credibility assessment. We collected coordinates of each click and its corresponding score in likert scale (0 for no effect to 5 for major effect).

Feature extraction In order to extract meaningful features from this experiment, we analyzed the results from a heatmap visualization (Figure 3.7) and statistical analysis using five-number summaries (Figure 3.8). As can be seen in Figure 3.8, there is overall similarity in credibility ratings on different elements for both Reddit and Twitter users. However, users of Reddit express higher priority on both information sources and textual elements for their credibility assessments. This observation may be due to small differences between two social platforms: For example, most of the posts in Reddit are directly connected to external webpages and this makes the source (URLs) more important during credibility assessments. Additionally, posts in Reddit are longer than in Twitter, which could account for the higher text credibility score for Reddit. Although the metadata scores are similar in Figure 3.8 for both platforms. Twitter does provide a richer set of metadata (e.g. classifications, hashtags, retweet counts etc.) on their page layouts, and this is evident from the increased number of clicks on metadata components in the heatmap (Figure 3.7(b)).



Figure 3.8: Click frequency/perceived credibility by UI component type for (a) Twitter and (b) Reddit (Exp1)

Exp2: Artificially Controlled Content

From the previous experiment, we selected a set of features on which to base our experimental evaluation for Exp2. Once again, our research question and hypothesis was refined based on information from the previous studies. By artificially controlling values for each target feature, we can assess the directional effect of metadata content on credibility perception. Furthermore, to avoid topic-specific biases in assessing the stability of feature influence on credibility perception across topics, we incorporated a variety of common topics (e.g.: World, Health, Politics, Entertainment) into the evaluation.

RQ5: How do different treatments of metadata variables influence credibility perception?

HYPOTHESIS. By applying artificially controlled values for metadata from the 5th and 95th percentiles of sampled real world data, we will observe differences in perceived credibility of the associated information.

RQ6: Is the influence of displayed metadata on credibility perception consistent across different topics?

HYPOTHESIS. *Feature influence varies across topics.*

Study design

In this experiment, we aim to test hypotheses 1 through 6 by artificially controlling metadata values for each of 12 salient factors identified in the previous experiment, and eliciting credibility assessments from participants. To achieve this, we construct 12 different formulated lists of microblog postings, controlling one independent variable on each list to capture how much impact that individual factor (e.g: profile image, link, number of friends) has on perceived credibility of information. Figure 3.10 shows a screen shot of the interface used in the study. The treatment in this case is a default profile image, which is the only controlled variable in this example. A five point Likert scale for feedback on perceived credibility is shown beneath the blog post.

A larger user experiment (N=646) was deployed on MTurk to evaluate the effects of artificially controlled treatments of each feature. In order to determine in which direction each factor impacts on the perceived credibility, we designed the study with two treatments and one baseline for each feature. *Treatment 1* uses ‘feature present’ in case of binary features and 95th percentiles (high values) in case of numeric features. Correspondingly, *Treatment 2* exhibits ‘feature absent’ in case of binary features and 5th percentiles (low values) for numeric features. The exception to this choice of percentiles are our two “sentiment” related features. Since we assume that low sentiment indicates more credibility, in view of the fact that objectivity is linearly correlated with credibility, we mapped the 5th percentiles to Treatment 1 and 95th percentile to Treatment 2 here.

Parameter Selection To estimate reasonable values for the treatments listed in Table 3.3, 1,727,556 Twitter messages and 4,000 Reddit posts including user profiles were crawled using the Twitter Streaming API and Reddit API. A distribution analysis of feature values in this data allowed us to find reasonable thresholds (and extremes) to select values for our experimental treatments. From this dataset, we extracted 5th and

Table 3.3: List of features and their independent variables in the second user experiment. (For each treatment, posts from both outlets, NYTimes and The Onion, are presented to the participants.)

| Type | Feature | Treatment 1 | Treatment 2 |
|-------------|-----------------------|------------------------|-----------------------|
| Visual | Embedded image | Present | Not present |
| Visual | Profile image(person) | Present | Not present |
| Visual | Profile image(logo) | Professional | Unprofessional |
| Network | # of friends | 95th percentile(7,524) | 5th percentile(8) |
| Network | Classification | News | Non-news |
| Network | # of comments | 95th percentile(9,565) | 5th percentile(0) |
| Network | # of shares | 95th percentile(933) | 5th percentile(0) |
| Network | Age of message | 95th percentile | 5th percentile |
| Content | Sentiment degree | No sentiment | High sentiment(95%) |
| Content | Sentiment polarity | Negative value(-0.95) | Positive value(+0.95) |
| Content | Tags | Tags present | No tags |
| Content | Links | Links present | No links |

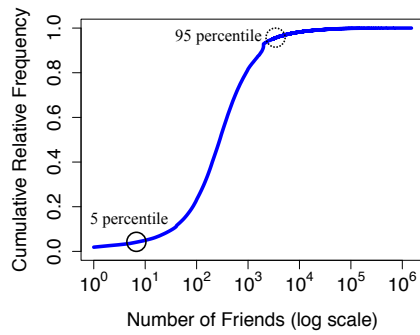
95th percentiles from the feature distributions.

The complete list of features tested in this experiment (Exp 2) are shown in Table 3.3.

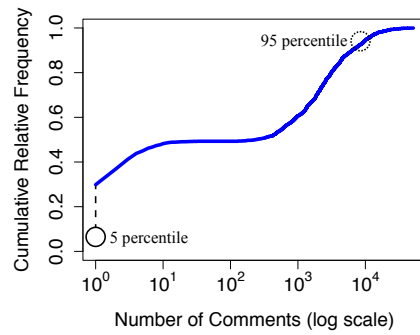
The leftmost column shows a categorization of each feature into one of three classes:

- *Visual* This is the set of highly visual factors in the microblog, including profile pictures, attached images, and photos.
- *Network* This is the set of network-based factors, including static features such as number of friends or followers, and dynamic/conversational features such as retweets, votes or mentions.
- *Content* This is the class of features solely based on text, including sentiment terms, hashtags, and links.

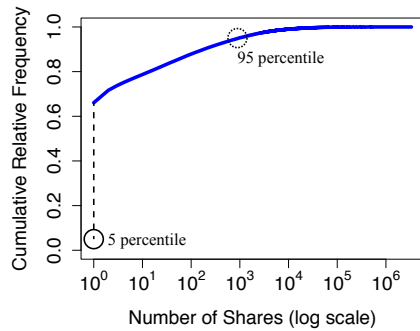
Table 3.3 describes our selection of treatments for each target feature. While we cannot exhaustively evaluate all values for a particular feature, our aim was to construct a reasonably diverse set of values for each, based in some cases on analysis of real world



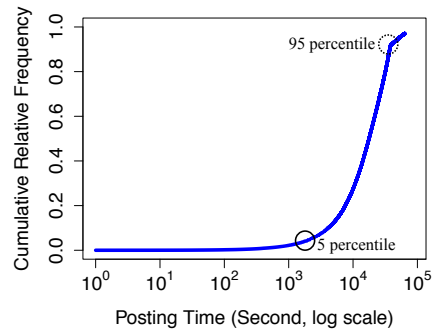
(a) Number of Friends



(b) Number of Comments



(c) Number of Shares



(d) Age of Message


Figure 3.9: Cumulative Relative Frequency (CRF) on numeric features on which we base to select conditions for the two treatments in Exp2. Please note that we use Twitter dataset (a sample set of 1,727,556 messages) for (a), (c) and (d) and Reddit dataset (a sample set of 4,000 posts) for (b).

distributions (e.g., for numeric attributes such as number of friends), and on manual selection for others (e.g: profile image content). The two rightmost columns give examples of controlled metadata treatments for each variable in our study. For instance, in the case of profile images, one treatment was selected from a popular satirical website. A second treatment (showing an image of President Obama giving a speech), was selected from a major US news outlet. The main goal of Table 3.3 is to show that differences can be produced in perceived credibility compared to a baseline, by manipulating a single visual feature, and our results show that this is indeed the case. For data with a range, such as number of friends, scores were taken based on large-scale distribution analysis for each feature, as described in the Random Sampling section below. The majority of examples were straightforward to construct, for example, network features such as number of friends or retweets were represented as low or high values. Recency was evaluated by grouping messages by age of message and displaying 5th or 95th percentile values of the resulting distribution. Sentiment features were more difficult however: for polarity and sentiment degree, a negative value (negative sentiment used) and low value (little sentiment used) were assigned respectively. These were indicated using “score bars”, shown in the example column of Table 3.3. For content/text features such as tags and links, a binary value (present or not present) was used. A baseline condition was also used. The baseline consisted of raw text with no associated metadata. Participants were asked to answer the same type of general questions about their demographic information and experience with microblogs as in **Exp1**. To avoid topic-specific biases and to explore our earlier hypothesis, 5 topics (World, Health, Politics, Entertainment, Business) and 10 posts from both *New York Times* and *The Onion* accounts on Twitter were manually collected. We purposely sampled two outlets that represent two different aspects of online journalism (objectivity and satire) in order to reflect real-world contexts.

GM To Pour All Resources Into Single Car That Can Be Safely Driven Down Street And Back

How believable do you think this post is?

Not at all believable
 Slightly believable
 Moderately believable
 Very believable
 Extremely believable



hours ago by @amony@exp2

A defective sense of smell appears to be a good predictor of how long you'll live

comments ## shares ## favourites sentiment

friends ## followers (www.thepirate.com)

How believable do you think this post is?

Not at all believable
 Slightly believable
 Moderately believable
 Very believable
 Extremely believable

At a higher level, what types of associated information do you feel influences your credibility assessment on the underlying text the most?

Visual components, such as profile pictures, embedded images,

Network-based information such as 'number of friends, shares or comments'

Features of the text such as sentiment, links/urls or tags.

Please explain your top choice.

Figure 3.10: Screenshot showing part of the microblog interface from Exp2.

3.3.4 Results and Discussion

In this section, we provide the findings from our two main user experiments. This section provides an overview of participant statistics for both studies, and following that, is organized around the six research questions and hypotheses posed earlier.

Study Participants

Exp1 had 102 participants. The average interaction time for both Twitter and Reddit users was 5 minutes. Users annotated three items each for a total of 306 annotations. 646 users participated in Exp2. Most reported that they were active daily on Twitter and had been active for more than a year. Most used the the official application, on a combination of mobile and desktop platforms. Most users did NOT guess that content was sourced from either the New York Times or the Onion. Facebook and Twitter was the most common response for the two source platforms. Participants spent an average of 9 minutes completing the survey and were paid \$0.40. 55% were male and 45% female. They ranged in age between 18 and 60, with the majority between 18 and 29.

Influence of Metadata

RQ1: Does the display of metadata in microblogs influence perceived credibility of the associated content?

HYPOTHESIS. Metadata display (textual or visual) does influence perceived credibility of microblog content. The direction of influence is dependent on the specific content displayed.

All three experiments produced results that reveal an impact of metadata on perceived credibility in microblogs. The subjective results in Survey 1 show a strong (but self-reported) indication that Content of a message and Origin (author) of a message are the

strongest influencing factors. This is followed by visual features, including design of the UI and visual components such as profile pictures and other metadata. The discussions that follow here further illuminate and reinforce this basic result.

Cross-Feature Analysis

RQ2: Do different features influence credibility by different amounts?

HYPOTHESIS. Features have varying degree of influence over credibility perception

Figure 3.11 shows a list of the 12 evaluated features ranked according to the observed difference in influence between the treatments from Table 3.3. The distribution clearly supports our initial hypothesis on RQ2. From this list, the profile picture/logo is the most influential factor on credibility perception, showing a significant increase over other features in terms of the difference in credibility rating between the treatments (ANOVA $p = 7.79e-9$ and $p < 0.05$ in Tukey post-hoc tests). Factors reinforced by the underlying network (Number of Friends –static, and Number of Shares –dynamic) are next-most influential. On the opposite end of the scale, sentiment polarity and age of message were the only two features where the lower value treatment achieved a better score than the higher value treatment. It seems that users in this study were not too concerned with recency. The more sentiment a message contained, the more likely it was to be deemed not credible.

Figure 3.11 shows the individual differences between the treatment 1 (white, top bar), baseline (dark gray, middle bar) and treatment 2 (light gray, bottom bar) for each of the 12 features. Overall, it is clear that the treatment 1 had a much stronger influence compared to the baseline (no controlled metadata) for almost all features. Effects were significant between the treatment 1 and the baseline in most cases but not significant

between the treatment 2 and the baseline.

Influence of Feature Classes

RQ3: What are the effects of different classes of microblog features on perceived credibility?

HYPOTHESIS. *Visual factors will have the most influence, followed by network and content-based (text) factors.*

Many existing credibility models perform feature classification to arrive at a credibility prediction for a given message. To evaluate our simple classification of (Network, Visual and Content)-based factors, we examined the overall group-wide credibility ranking and the results are shown in Figure 3.12 (a). Our initial hypothesis was that visual factors would be most influential, and it appears from the left column in Figure 3.12 (a) that this is in fact the case, at least for the high value treatments of the features in the group. No significant effect was shown for the lower value feature differences. The visual features produced a 10% increase over the baseline, compared with 7% for the Network and 3% for the Content group. This result further supports the notion that manipulating visual components can have a large effect in terms of how strangers perceive a microblog profile.

Cross-Platform Analysis

RQ4: Can our models of feature influence be ported successfully to different microblogs?

HYPOTHESIS. *Influence of features is consistent across platforms*

To recap, **Exp1** evaluated self-reported importance of microblog features across two platforms: Reddit and Twitter, allowing users to place simple clicks in context of the ac-

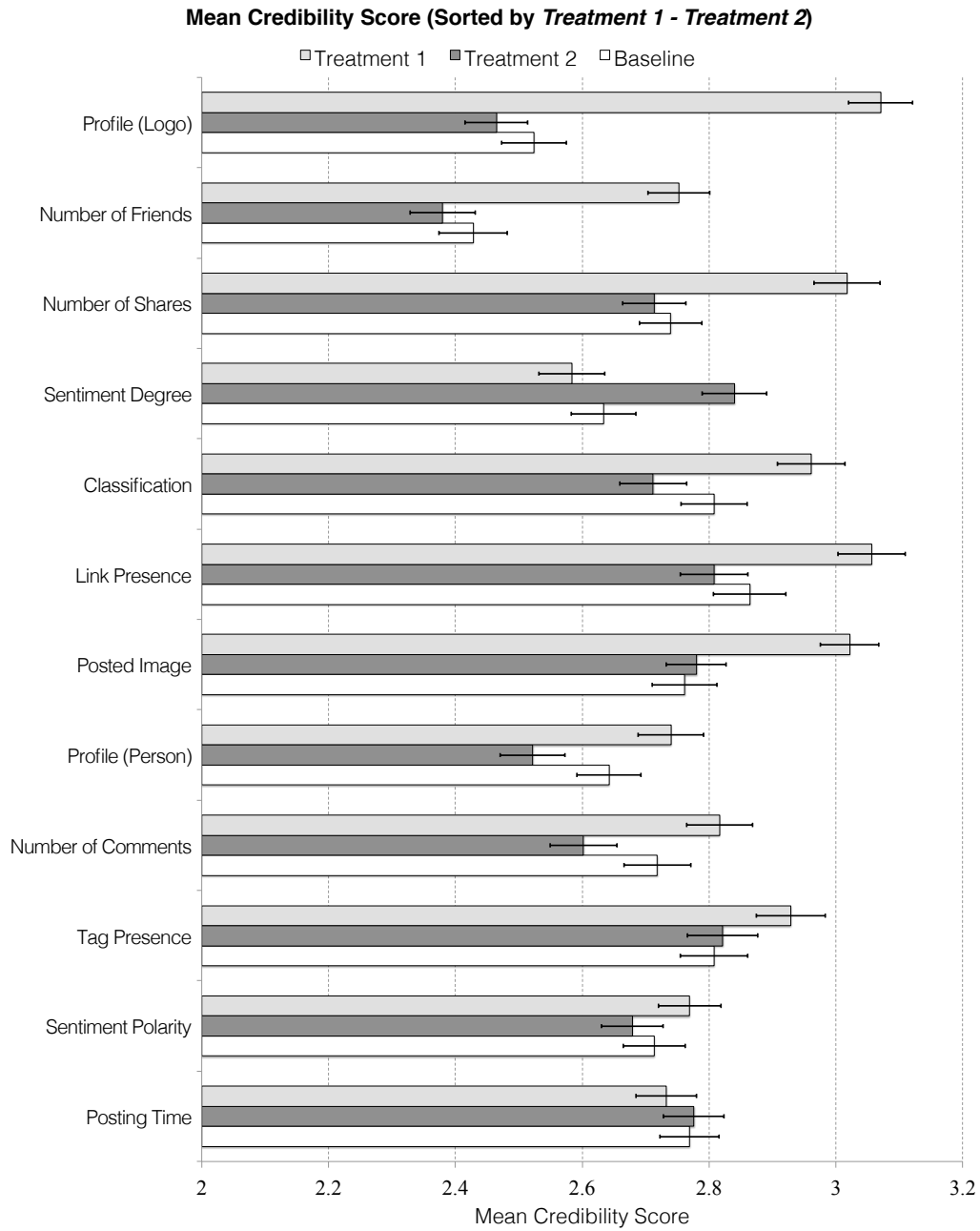


Figure 3.11: Mean credibility scores across different features for the three treatments including baseline. Groups are sorted from top to bottom by the difference between the treatment 1 and treatment 2 scores.

tual interface. Figure 3.8(a) and (b) show the results for Reddit and Twitter respectively. Our results disprove our initial hypothesis that feature ratings are invariant, since the text features achieved a higher credibility rating in Reddit. This is likely to be related to the fact that there is significantly more text per post allowed in Reddit. Image features appeared to garner similar ratings across the two platforms, which is a meaningful finding, especially coupled with the fact that the Visual/Image-based features are the strongest influences on credibility perception.

Impact of Different Treatments

RQ5: How do high and low values of metadata contents influence perceived credibility?

HYPOTHESIS. *Low values for metadata have a stronger effect than high values*

Figure 3.12(d) shows a small (7%) improvement across all features and topics for treatment 1 over treatment 2. There is also a significant improvement shown for treatment 1 over the baseline (6%), Specifically, a single factor ANOVA test over the treatments shows the result is statistically significant (F value:87.54, $Pr(> F) < 2e - 16$).

Cross-Topic Analysis

RQ6: Is the influence of displayed metadata on credibility perception consistent across different topics within a domain?

HYPOTHESIS. *Influence of the treatments varies across topics.*

To determine whether or not credibility ratings for our feature sets vary across different topics, and also to avoid introducing topic-specific biases in our other evaluations, we examined a collection of 5 diverse topics (World, Health, Politics, Entertainment and

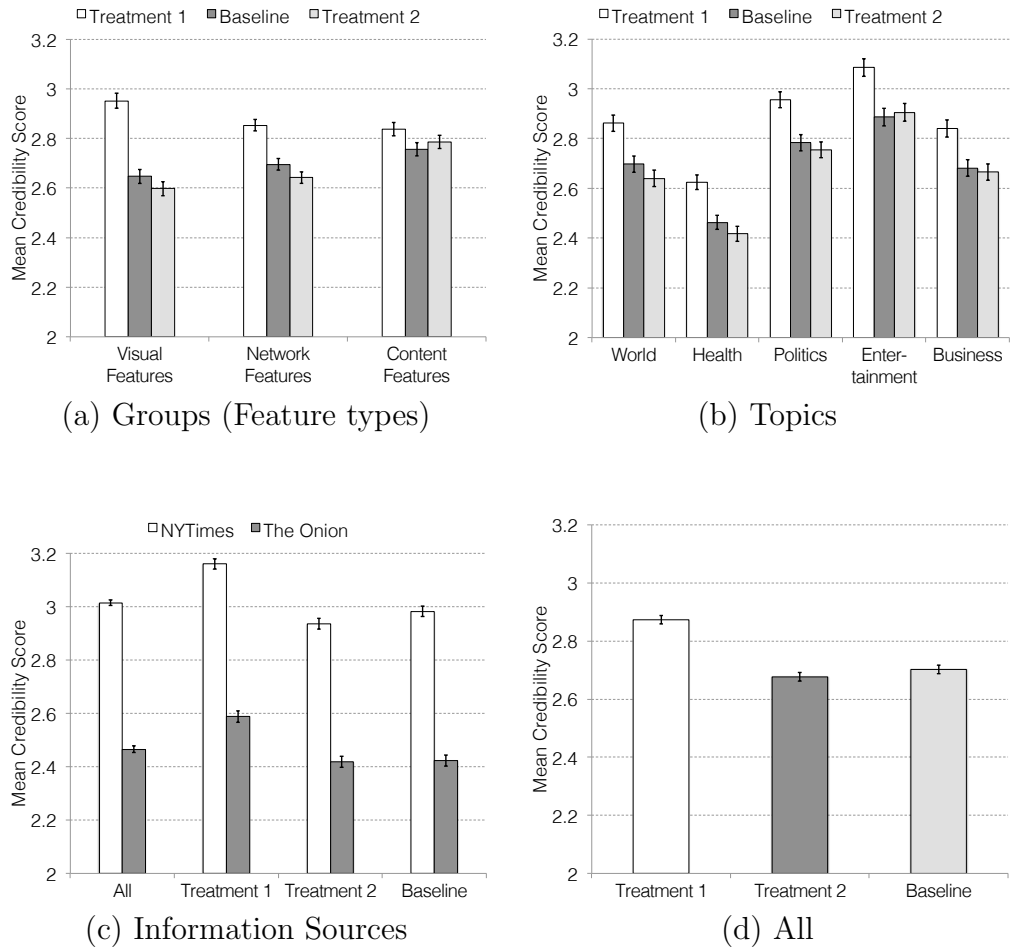


Figure 3.12: Mean credibility scores across different treatments

Business). This approach could be relevant for a system such as “CatStream” [171] which classifies Twitter feeds automatically based on topic interests. Messages from two well-known websites were sampled for each topic on a specific date in September 2014. The two websites were the New York Times (News) and The Onion (Satire/Comedy). Figure 3.12 (b) shows the results of an analysis of our different treatments across each of the five topics individually. From the results, it is immediately clear that the recurring trend of treatment 1 having a stronger effect than treatment 2 is invariant across all topics. ANOVA tests for differences showed $p < 2e - 16$. Another interesting result is that the overall ratings for the health-related topic was significantly lower than all other topics, across all treatments. This might be a reflection of participants’ cautious behavior when approaching a serious topic such as health. Conversely, the Entertainment topic produced the highest credibility ratings. Perhaps it is easier to appear credible in a domain that has far fewer constraints.

In addition to the cross-topic analysis, we also examined the overall ratings of articles from “New York Times (NYT)” and “The Onion” across every rating context, broken down by the three treatment conditions. Figure 3.12 (c) shows the results of this analysis. There is a significant improvement of 25% (ANOVA, $p < 2e - 16$) in credibility rating between the two sources, with NYT at the upper end. Over all features, treatment 1 of metadata produces an equivalent increase in both NYT and Onion ratings, keeping the difference at approximately 25%. When comparing treatment 2 to the baseline treatment, we did not observe any significant change.

3.3.5 Summary

To conclude, in this set of three interlinked studies we evaluated ways in which individual components of microblogs can influence end-user perceptions of information cred-

ibility. In particular, an initial survey (N=81) provided general insight into a candidate set of factors that influence credibility most; a second study (N=102) examined these factors in the context of two microblog domains, Reddit and Twitter, to allow real users to communally identify the most influential factors. A set of 12 factors were evaluated in detail in a third study that artificially controlled values for each feature and assessed credibility opinions from 646 participants in a crowd sourced evaluation.

Six hypotheses related to the impact of different microblog elements and treatments on human-provided assessments of credibility were tested and the results were discussed in detail. Key findings from the study show that 1) metadata from the high end of observed real-world distributions (e.g. high number of friends, high number of hashtags) have a far stronger effect on credibility ratings than treatments from the lower end of the distribution. 2) Visual factors, in particular, display of a Profile Picture Logo had the most positive impact on reported credibility. 3) Participants in the study did not view recency of posts as an important factor in credibility assessment. 4) Factors that influence perceived credibility did not remain constant across platforms. Text-based features scored higher on Reddit while Visual features did remain constant.

In follow-up studies, the authors plan to further examine some of the findings from this work. For example, what are the exact reasons for ‘negative’ traits, such as low numbers of friends, comments, and shares, not exhibiting as much of an effect on the mean credibility score as the positive ones across most features and topics? Future work will also apply findings from this experiment to improve prediction models from real world data such as [29, 30], predicting newsworthiness, credibility, or actions such as retweets, up-votes or shares.

3.4 Conclusion

In this chapter, we have investigated into important factors that affect users' perception of information reliability, particularly focusing on information credibility on microblogs (Twitter and Reddit) through different user studies. Different metrics such as self-reported credibility ratings and interaction data provided by human users have been used to shape the online user study with the artificially controlled 12 credibility factors. The results and implication of the study (Exp 2, N=646) show that there are several important (mainly visual and metadata) features that can be harnessed to detect reliable information on the Social Web. Another finding is that such features are selectively portable across platforms. Furthermore, credibility perception significantly varies across different topics, although exhibits consistent variation in terms of the type of features. Based on these findings, in the next chapter, we will demonstrate our studies on modeling different attributes of information reliability, such as information credibility, competence (expertise) and influence of information source, and newsworthiness of information.

Chapter 4

Modeling

In Section 1.2, we provided a definition of information reliability, which is based on, and extended from, an established information quality framework by Wang and Strong [14]. We also discussed current issues and research topics on the Social Web. Since our study aims to *automatically* identify and recommend reliable information on the Social Web, followed by the study on the perception of information reliability, in this chapter, we take a closer look into how we can model different attributes of information reliability across contexts through our recent studies. This chapter will be followed by the validation of our reliability models we study in this chapter (Chapter 5) and communicating information reliability (Chapter 6).

4.1 Introduction

We aim to find the feasibility of modeling information reliability and, taking a step forward, design computational algorithms by which we identify and recommend the best reliable information for users. Our user studies and experiments on perceived credibility revealed some important factors and conditional attributes that affect information relia-

bility assessment. However, current understanding of human perception of information is still in its infancy since there are a number of variables and conditions that lead to different reliability perception. For example, different user intents in information seeking tasks may result in diametrically opposed perception. Exploratory or serendipitous information search does not value relevance to the query used, and thus, differ from an advanced search task with several options and conditions seeking for a particular result in mind. When a user is seeking information to confirm a non-verified testimony or to disprove an existing claim, interestingness may not be taken into account. In brief, from the disparity of tastes or views between individuals to difference in demographic background, various constraints must be considered in modeling information reliability. Furthermore, people hold different perspectives on the definition of information reliability. In this chapter, we provide a general insight on modeling information quality and report on our recent studies on the same topic.

In the rest of this chapter, we first discuss modeling information reliability in general. Second, we provide several studies on different information attributes: credibility (Section 4.3); user competence (Section 4.4); newsworthiness (Section 4.5); and influence (Section 4.6).

4.2 Modeling Information Reliability

To quantify and gauge information reliability, we need to identify relevant aspects or attributes of information. Having relevant information quality attribute(s) in mind, we can conceptualize how humans evaluate and consume information in the real-world. We already discussed important reliability attributes in the previous chapters. Depending on the task or context, attributes such as credibility, newsworthiness, topical relevance, and expertise may be considered as the indicators of information reliability. We selected

some attributes and briefly discussed them in section 1.3.

4.2.1 Reliability Metrics

We selected four reliability attributes that are commonly studied in the context of the Social Web. Table 4.1 lists them and summarizes with brief description of each element. Since users prioritize different aspects of information across tasks or topics, we need to specify a particular context to model information reliability. O’Donovan et al. [110] reported on their feature analysis on microblog data and found that feature distribution varies across different topics and content types. In a more recent study, Sikdar et al. [49] revealed that the difference in a user’s search intent on the same topic changes not only the content produced by the community but also dynamics of communication in the social network. They discuss the results with the portability of ground truth in credibility assessment on microblog contents. We need to take these findings into consideration when we model attributes of information reliability. Below, we provide guidelines for modeling information reliability regarding individual factors to be considered.

Attribute We need to specify one or more attributes to model and specific task or context from users’ perspective. For instance, if a journalist searches for information on a recent earthquake occurred in the local area in social media, s/he may prioritize credibility of information source and content, timeliness of information (how recent the information is), and expertise or authoritativeness of the source during the assessment.

Platform Our study on credibility perception in microblogs (Twitter and Reddit) in Chapter 3 confirmed that patterns in feature behaviors and perceived credibility *selectively* differ across social platforms. As shown in Figure 3.8, some of the features were evenly distributed, which means consistently perceived, across platforms. The result

| Attribute | Property | Correlated Attributes | Example Features |
|------------------------|----------|--|--|
| Credibility | I&E | Trustworthiness, Believability, Expertise, Newsworthiness, Relevance | url (external links), popularity score, citation count, number of friends, etc. |
| Competence (Expertise) | E | Credibility, Expertise, Relevance, Popularity, Experience, Influence | popularity score, citation count, topical diversity, topical keywords, number of friends, etc. |
| Newsworthiness | I | Relevance, Credibility, Timeliness, Influence, Impact, Interest, Proximity, Prominence, Oddity [172] | topic, keyword, content similarity, publication date, citation count, popularity score, uniqueness score, etc. |
| Influence | E | Popularity, Credibility, Expertise, Experience, Prolificity | number of endorsements, recommendation count, number of citations, number of mentions, etc. |

Table 4.1: List of attributes of information reliability (covered in this chapter) for the Social Web (e.g. blogs, microblogs, social networks). Entries of the property column are either I (Information metric) or E (Entity metric), or both. Please note that “correlated attributes” column contains only a set of selected attributes studied in the literature (not an exhaustive list).

indicates that the consistently behaving features can be generalized independently of platform types. However, others such as information source and textual features vary across the platforms. Thus, features and other aspects must be carefully selected by taking into account their portability across platforms.

Pertinent features Once we select candidate features considering their portability across platforms, for the given task or context, useful features on which users can rely on must be chosen. It is not a trivial task since we need to have a thorough understanding of feature behavior, profound background knowledge and wealth of experience in the given context for successful feature selection. For a computational modeling approach, we can apply feature selection algorithms to find the best feature set. However, in some cases, general feature selection algorithms are not sufficient to expect the best result in the assessment. Sometimes, different approaches such as mathematical modeling that

involve synthesis of multiple features and weighting schemes are needed; see Section 4.4. We will discuss modeling procedure more in detail in this section.

Ground truth When we develop a new information reliability model, we need to validate the model to make sure it properly functions under the expected conditions. Using the real-world data, we can examine how the model address the aspects (e.g. selected quality attribute(s), the desired tasks, users' search intent, etc.) in the wild. In computational modeling, the most typical approach is evaluating the model with labeled ground truth data. Using the labeled datasets, we split the data into training and test sets and apply them to appropriate machine learning algorithms. K-fold cross-validation approach is typically used (e.g. $K=10$). For computational modeling, different metrics such as precision, recall, F-measure, G-measure or mean absolute error (MAE/MAPE) can be used to evaluate the model. For statistical modeling, correlations between the ground truth and feature behaviors can be tested. In fact, different statistical measures If multivariate analysis is used, other techniques such as inverse or pairwise correlations, covariance matrices, and principal components. Regarding ground truth and validation, various approaches in literature and more discussions are provided in Chapter 5.

4.3 Modeling Credibility in Twitter

This section summarizes our previous study [30] on modeling information credibility in microblogs. We provide important findings and implications from the results of the study in which we developed three computational credibility models for microblog contents.

We believe that this study provides a useful guideline for future study in identifying information credibility in microblogs. In particular, the study builds on the state-of-the-art approach performed by Castillo et al. [29] but proposes a different method for assessing the credibility of individual microblog messages, not an aggregated set of posts regarding newsworthiness. In this study, we utilize the three computational models as an apparatus in the context of topic-specific information seeking tasks on microblogs.

In addition to this study [30], our recent study [169] tackles the same research question by starting from the perceived information credibility, based on the perception experiments discussed in Chapter 3. The findings from the user experiments in Chapter 3 show us that we can harness both global features and context-dependent features to model information credibility in different circumstances. Furthermore, there are specific goals or motivations of use for which each social platform is designed. For instance, Reddit¹, a popular social bookmarking site, has its main purpose of use: information sharing, mostly for news and entertainment. The entertaining factor of its use exists in community activity. *Microreddit* communities and its original feature referred to as *up/down votes* are the key value of the service. With such considerations in feature selection and modeling processes, a close imitation of reliability assessment to human judgments in the real-world can be achieved by computational algorithms.

The main focus and contribution of this study are on an evaluation and comparison of three novel approaches for predicting “credible” information for specific topics on Twitter

¹<http://www.reddit.com/>

—an important challenge given the deluge of noise and misinformation in the network. We first model social credibility, then focus on content-based (message-only) credibility, and lastly on a hybrid of features from both strategies. This approach aims to maximize the available window of information from the social networking platform. This study focuses on two main research questions.

1. How well can we assess credibility in Twitter using our proposed models?
2. How do social, content-based and hybrid models perform at identifying credible information?

We evaluate our methods on a range of metrics, from credibility-based predictions of simple features from available metadata, to prediction on thousands of tweets with manually labeled credibility assessments. The results from our user study assessing credibility on a set of tweets with varying credibility indicators (context) are also presented and discussed.

4.3.1 Experimental Setup

The first model focuses on credibility at the user level, harnessing multiple dynamics of information flow in the underlying social graph to compute the degree of credibility. The second model employs a content-based approach to compute a finer-grained credibility rating for individual microblog messages. Finally, we discuss the third model that combines aspects from the two models in a hybrid method. To evaluate the proposed credibility models, we evaluate on seven topic-specific data sets collected from Twitter, with specific focus on a data set of 37K users who tweeted about the topic “Libya”. The results show that the social model outperforms hybrid and content-based models in terms of predictive accuracy over a set of manually collected credibility ratings on the “Libya” dataset.

Defining Credibility. This study defines two types of “credibility” in the context of a target topic of interest:

Definition 1 *Tweet-Level Credibility:* A degree of believability that can be assigned to a tweet about a target topic, i.e.: an indication that the tweet contains believable information.

Definition 2 *Social Credibility:* The expected believability imparted on a user as a result of their standing in the social network, based on any and all available metadata.

Tweet-level credibility is akin to Castillo’s definition in [29], with the addition of the topic level constraint. Tweet level credibility can also be summed and propagated to the user level by averaging over a profile of tweets. Conversely, a user’s social credibility is attached to all of the tweets on her timeline.

4.3.2 Modeling Credibility

Traditional recommendation strategies such as content-based [173] or collaborative filtering [174, 175] typically compute a *personalized* set of recommendations for a target user based on some derivation from that user’s profile of item preferences. An important distinction between these techniques and the approaches presented here is that personalization is only performed at the topic level in our algorithms. While we believe that traditional personalization does play an important role for predicting credible content, the focus here is on predicting credible information within a target group centered around a topic of interest.

Given these goals and constraints, we present three computational models for assessing information credibility within a specific topic. We begin with by defining nomenclature for the domain:

Definition 3 *The Twitter domain can be represented as a quintuple (U, F_o, F_e, T, X) , where F_o and F_e are two $U \times U$ matrices representing binary mappings $f \in F_o, F_e \mapsto 0, 1$ between users in U (termed “follower” and “following” groups, respectively). T is the set of tweets, distributed over U , and X is the set of topics in T .*

By this definition, Twitter is rich in both text content and social network links. Research in recommender systems has long argued the benefits of combining content-based and collaborative approaches to recommendation to maximize information gain in the prediction process [173, 174, 175]. For example, while content-based methods tend to predict narrowly, in that they must match a text description of an item already in a target user’s profile, collaborative techniques can provide more serendipitous predictions since they are based on subjective opinions of groups of similar users.

Since our domain is rich in both content and network links, we propose the following three approaches for identifying credible information, borrowing from the content and collaborative synergies identified by the recommender system community.

- *Social Model*: A weighted combination of positive credibility indicators from the underlying social network.
- *Content Model*: A probabilistic language-based approach identifying patterns of terms and other tweet properties that tend to lead to positive feedback such as retweeting and/or credible user ratings.
- *Hybrid Model*: A combination of the above, firstly by simple weighting, and secondly through cascading/filtering of output.

Social Model

Complex network structure and feed-based information flow make dissemination of information on Twitter extremely dynamic and ephemeral. Therefore, detecting credi-

bility factors is inherently difficult. Moreover, outliers such as celebrity accounts (e.g. Oprah Winfrey follows 245 users, but is followed by 31.3 million accounts in Twitter), fake or malicious accounts (e.g. bots) and accounts used for social marketing campaigns all don't behave as "regular" nodes in the network. Our social model attempts to mitigate these problems by weighting a diverse range of *positive credibility indicators* within a target topic.

We first consider the "retweet" as a proxy of credibility. Equation 4.1 gives a value for credibility based on the deviation of a user $u \in U$'s retweet rate RT_u from the average retweet rate $\overline{RT_x}$ in a target topic $x \in X$. In practice, values from the following equations are mapped to a log-log scale to handle large outliers in the data.

$$Cred_{RT}(u, x) = |RT_u - \overline{RT_x}| \quad (4.1)$$

Keeping with retweet analysis, Equation 4.2 considers retweet rate but factors in usage rate and number of followers F_o , in other words, a utility metric from the potential number of retweets.

$$Utility_{RT}(u, x) = \left| \frac{RT_{u,x} \times F_o(u)}{t_{u,x}} - \frac{\overline{RT_x} \times \overline{F_{o,x}}}{t_x} \right| \quad (4.2)$$

Retweet metrics function over both the content of a set of tweets and the underlying network. We believe that the network topology itself can also provide insights into credibility of a user. Equation 4.3 computes a social credibility score as the deviation in the number of user u 's followers from the mean number of followers in the domain, normalized by number of tweets.

$$Cred_{social}(u) = \left| \frac{F_o(u)}{t_u} - \frac{\overline{F_o}}{t} \right| \quad (4.3)$$

Assuming that a “follow” request is usually an indication of credibility, we can now also weight Equation 4.3 by taking account of the friends to followers ratio as a deviation from the norm for a given topic. For instance, an information gathering agent is likely to follow many profiles, but have few followers. Equation 4.4 describes the social balance of a user u as the ratio of follower (F_o) to following (F_e) group size.

$$Balance_{social}(u) = \left| \frac{F_o(u)}{F_e(u)} - \frac{\overline{F_o}}{\overline{F_e}} \right| \quad (4.4)$$

There are cases where the opposite is true however, for example, a popularity-hungry politician may pay to have automated agents create accounts and follow his profile, but these profiles are not likely to have strong social connectivity, and can be discounted by other filters in this model, such as Equation 4.2 for example.

We also consider social connections within a given topic as a positive indication of credibility, both in the F_o and F_e groups. Consider a user who has tweeted frequently about a topic, lets say “#androidgames”. If that user has few or no followers with associations to that topic, this should raise suspicion about the user’s credibility in the topic. Our findings indicate that network data is frequently too sparse within a specific topic for this metric to yield useful results, but we include it in the model because it leverages social connections in a potentially useful way.

$$Cred_{social}(u, x) = \left| \frac{F_o(u, x)}{t_{u,x}} - \frac{\overline{F_{o,x}}}{\overline{t_x}} \right| \quad (4.5)$$

The final metric in our social credibility model addresses the focus of a target user within a given topic space as a function of their global profile. For example, many people have set up Twitter accounts solely for business or research purposes, and thereby have a more constrained number of topics that they tweet about, potentially indicating an

increased level of credibility, since the likelihood of recurring topics is higher. Equation 4.6 computes this metric as the sum of the tweets for a user u on topic x as a percentage of their total number of tweets t_u .

$$Focus(u, x) = \left| \frac{\sum_{t \in T} t_{u,x}}{\sum_{t \in T} t_u} \right| \quad (4.6)$$

Content-based Model

We have described how the social provenance of a piece of information can have a bearing on its credibility. However, credibility can be assigned both to the information source, and to the information itself in an intrinsic way. Accordingly, our second credibility model focuses on message content, isolated from the underlying social network. We begin by representing all tweets in our topic-specific datasets as a set of salient credibility indicators (12 numeric and 7 binary).

Numeric Indicators:

1. *Positive Sentiment Factor*: Number of positive words (matching our lexicon)
2. *Negative Sentiment Factor*: Number of negative words
3. *Sentiment Polarity*: Sum of sentiment words with intensifier weighting (x2) ('very', 'extremely' etc)
4. *Number of intensifiers*: 'very', 'extremely' etc., based on our lexicon.
5. *Number of swearwords*: Simple count, based on lexicon.
6. *Number of popular topic-specific terms*: Simple count, based on lexicon.
7. *Number of Uppercase Chars*: Simple Count

8. *Number of Urls*: Simple Count
9. *Number of Topics*: Number of topics ‘#’ (All have at least 1)
10. *Number of Mentions*: Number of user’s mentioned with ‘@’
11. *Length of Tweet (Chars)*: simple count.
12. *Length of Tweet (Words)*: simple count.

Binary Indicators:

1. *Is Only Urls*: No text, only links.
2. *Is a Retweet*: From metadata
3. *Has a Question Mark*: ‘?’ or contains any of Who/What/Where/Why/When/How
4. *Has an Exclamation Mark*: ‘!’
5. *Has multiple Questions/Exclamations*: ‘??’ ‘???’ ‘!!’ ‘!!!’ etc.
6. *Has a positive emoticon*: :) :-) ;-) ;)
7. *Has a negative emoticon*: :(:-(-) ;-(

To evaluate the utility of this model for predicting credible information, we train a range of classifiers using 5,000 manually annotated tweets from a user evaluation. Details and results of this analysis are presented in Section 4.3.3.

Hybrid Model

So far we have focused our discussion on credibility indicators at the user level and the tweet level individually. A logical progression is to combine aspects from both methods to maximize the information upon which we can base credibility decisions. We now

present four novel methods for combining aspects from the earlier models to better predict credible information and sources. Since our earlier models compute credibility at different levels of granularity (user and information level), so also do the following hybrid strategies.

Content-based Ranking This strategy predicts credibility at both the user and tweet levels. The hybrid algorithm first performs a filtering step based on the user level (social) credibility score from model 1, passing profiles with a credibility score above a threshold S_{min} to the second model. The content based model extract features from each tweet and computes a credibility score which is used to re-rank tweets from the set of credible users. $u \in U$ where $S_u < S_{min}$.

Weighted Combination This simple combination of output from the two earlier models predicts at the user level only. Credibility scores from the content-based model are aggregated over each $u \in U_t$. The resultant user-level score C_u is combined with the social credibility S_u using a harmonic mean weighting strategy to minimize outlier values:

$$C_{weighted} = \frac{2}{C_u + S_u}.$$

Feature Combination This strategy computes a credibility at the tweet level, and is designed to use all available data to generate a prediction. Feature lists from both the social model, the content based model, and a collection of other user metadata obtained by using the Twitter API are taken to train a J-48 decision tree² to generate a prediction model.

Content-boosted Social Credibility The final hybrid method predicts at the user level, incorporating the aggregated content-based score for a user into the social model.

²J-48 is another name of C4.5 classifier, which is an extension of Quinlan's earlier ID3 algorithm.

| <i>Set Name</i> | <i>Core Tweepers</i> | <i>Core Tweets</i> | <i>F_o and F_e (overlapped)</i> | <i>F_o and F_e (distinct)</i> |
|--------------------|----------------------|--------------------|---|---|
| <i>Libya</i> | 37K | 126K | 94M | 28M |
| <i>Facebook</i> | 433K | 217K | 62M | 37M |
| <i>Obama</i> | 162K | 358K | 24M | 5M |
| <i>Japanquake</i> | 67K | 131K | 25M | 4M |
| <i>LondonRiots</i> | 26K | 52K | 30M | 4M |
| <i>Hurricane</i> | 32K | 116K | 35M | 5M |
| <i>Egypt</i> | 49K | 217K | 73M | 36M |

Table 4.2: Overview of 7 topic-specific data collections mined from the Twitter streaming API.

This approach is similar to the Weighted Combination with the exception that the content-based credibility factor is considered at the same level as the $Cred_{RT}$ and $Cred_{social}$ scores from Equations 4.1 and 4.2 respectively.

4.3.3 Evaluation

Now that we have presented our models for predicting credibility, and our ground truth collection process, we must assess and compare the performance of each model. Given our available resources, substantial credibility assessment data could only be collected on the Libya data set, so we focus on that set for most of the following evaluation.

Data Analysis

Before we describe our evaluation of predictive accuracy, we first take a broader statistical view of the collected data sets to gain insights about trends, clusters and any interesting anomalies about the data sets. Figure 4.4 shows a comparative analysis of a selection of features from both social and content-based models. In this figure, lighter (red) nodes indicate positive credibility and darker (blue) nodes indicate negative credibility assessments from the user study. Clusters appear in some of the scatter plots, indicating that the feature does have some bearing on assessed credibility. For example,

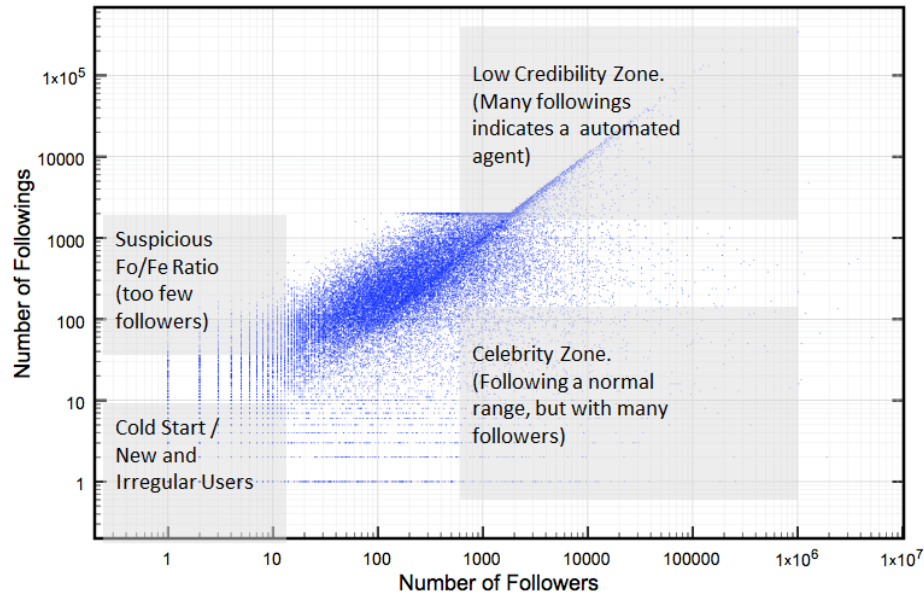
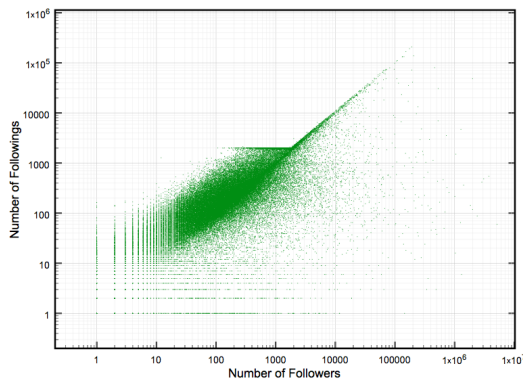


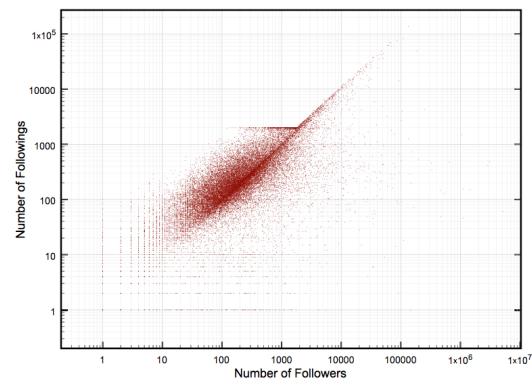
Figure 4.1: Plot showing number of followers to number of following profiles for the Libya data set. Areas of particular interested are shaded in grey and labeled accordingly.

looking at the features for “char” and “word”, it is clear that longer tweets tend to be assigned more credibility than shorter ones. Number of tweets (status-count) and number of listings (listed-count) also align well with reported credibility.

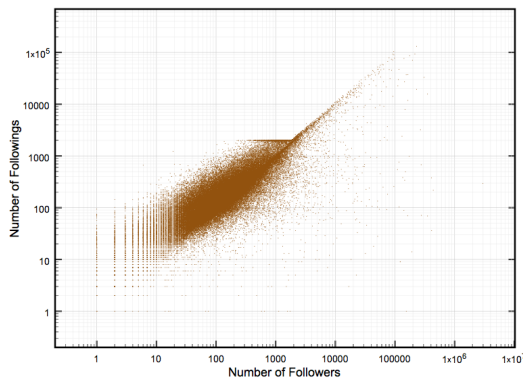
Friend to Follower Analysis Figure 4.1 shows a comparison of the number of followers F_o to the number of following users F_e over the 37K core users in our Libya data set. The shaded areas in the resulting distribution reveal areas that have a potentially negative impact on credibility. For example, at a threshold where users are following more than approximately 5,000 accounts, the data appears to form a straight line. However, on the log-log scale, this is evidence of the long tail of a power law distribution. This is highlighted as low credibility zone since the group size is abnormally large for a human user, and we must, therefore, assume that the profiles are based on automated agents or



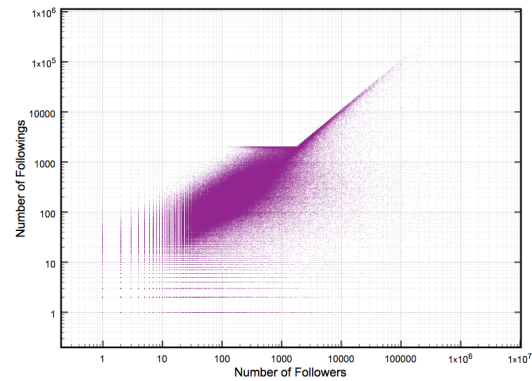
(a) Japan Quake



(b) Hurricane Irene



(c) Enough is Enough



(d) Facebook

Figure 4.2: Friend to Follower patterns across four of our topic-specific data sets. All other sets exhibited a similar distribution.

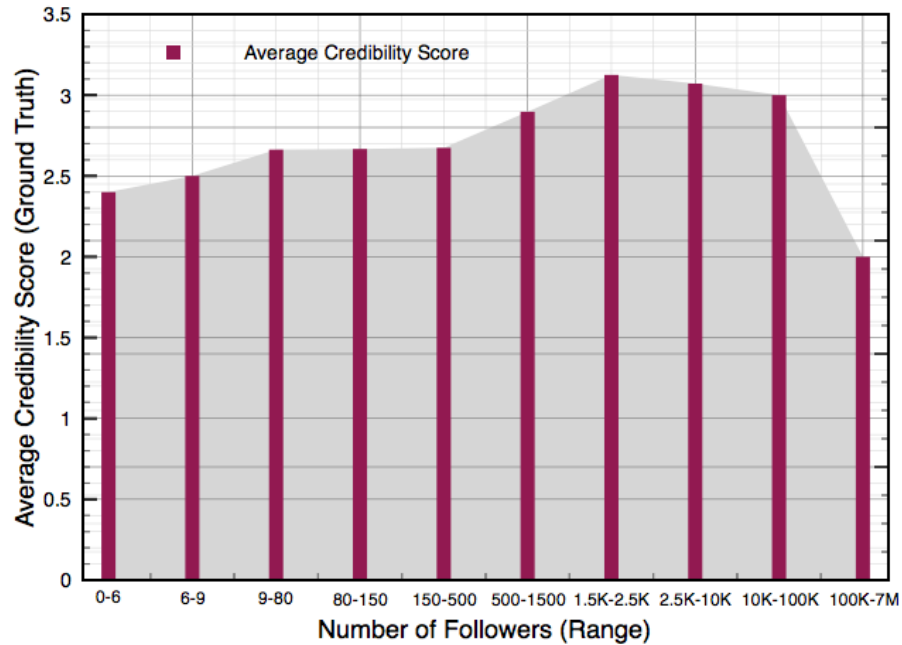


Figure 4.3: Average credibility rating from the web survey versus number of followers for the tweet authors (binned).

“bots”. Conversely, small sized F_o and F_e groups indicate new or inactive users. This is a low credibility group because we do not have sufficient social/content information to perform a reasonable credibility assessment. Groups along the other extremities of this graph are also interesting. Those with very few followers but larger following groups are likely to tweet less and be leaf nodes in retweet chains, while conversely, the “celebrity” group (high F_o and low F_e) tend to be higher in retweet chains, and have many retweets. The latter two groupings do not necessarily bear on credibility, but the other shaded areas of Figure 4.1 do indicate negative credibility. Accordingly, the “balance” component of our social credibility model, shown in Equation 4.4 is weighted to penalize these groups. Figure 4.2 shows similar distributions of follower to following groups across other topics. We found similar distributions for all of the other sets in Table 4.2.

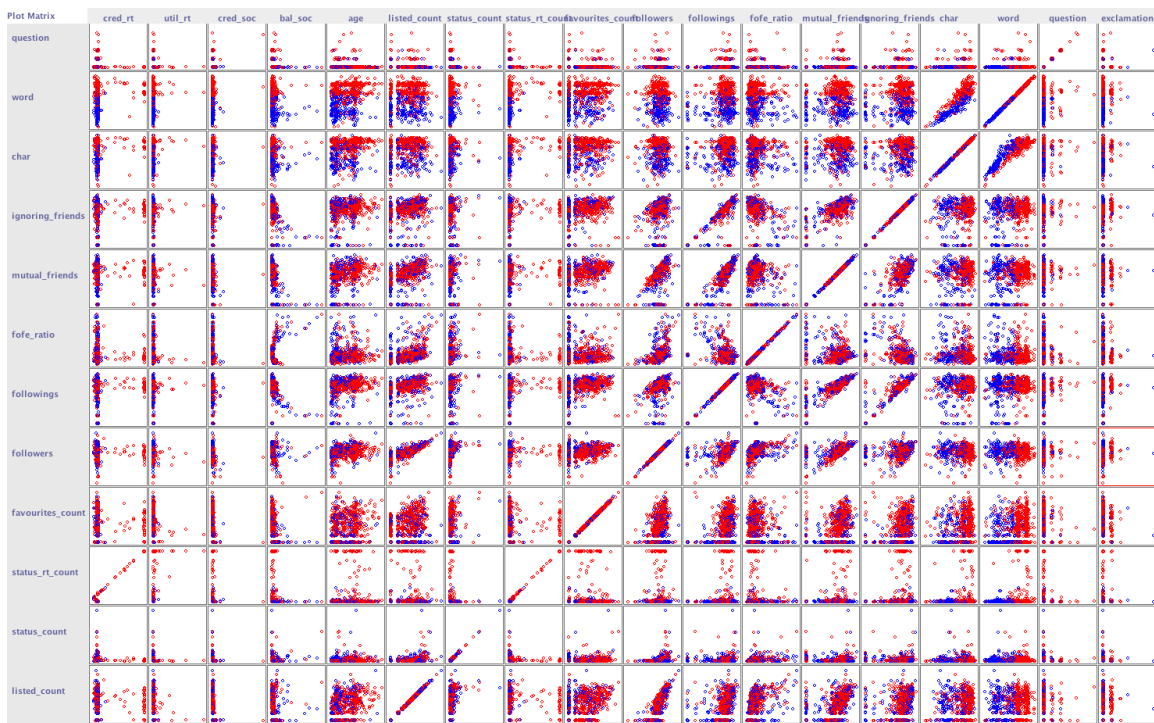


Figure 4.4: Comparisons of each feature used in computing the social credibility model. In this plot, lighter (red) areas indicate high credibility and darker (blue) areas indicate low credibility assessments.

Followers and Credibility Figure 4.3 shows an analysis of the average credibility reported for tweets in the user study compared with number of followers. Binning of followers was applied to highlight the distribution. There is a significant correlation between reported credibility and number of followers up to a network size of approximately 1,500 followers, after which, reported credibility drops off steeply. This result aligns well with our earlier analysis of follower to following groups. The reported drop in credibility past this threshold is likely due to the “bot” effect shown in Figures 4.1 and 4.2.

Retweets, Links and Other Credibility Indicators The hybrid approaches rely on a variety of different features from the network, messages (content) and from user metadata. Once again a discussion of the benefits and merits of all features tested is not possible here, as a sample, we now discuss interesting findings from our analysis. The distribution graph in Figure 4.4 shows an overview of a subset of features. Links (presence of urls in tweets) were found to be a very positive indicator of credibility and were more frequently used in older profiles. Users who provided links frequently tended to be listed more often, and added to other users’ “favorite” groups. Retweets were generally reported as credible in our study, and were also more frequent in older user profiles. Interestingly, emoticons (both positive and negative) were found to be an indicator of retweeting. Additionally, longer tweets were retweeted more frequently than shorter ones.

Predicting Credibility

Treating each hybrid strategy independently, a total of 6 credibility prediction strategies were evaluated. Each strategy was represented as a set of weighted features and loaded as an input file to WEKA³ machine learning toolkit. Our goal was to accurately

³www.cs.waikato.ac.nz/ml/weka/

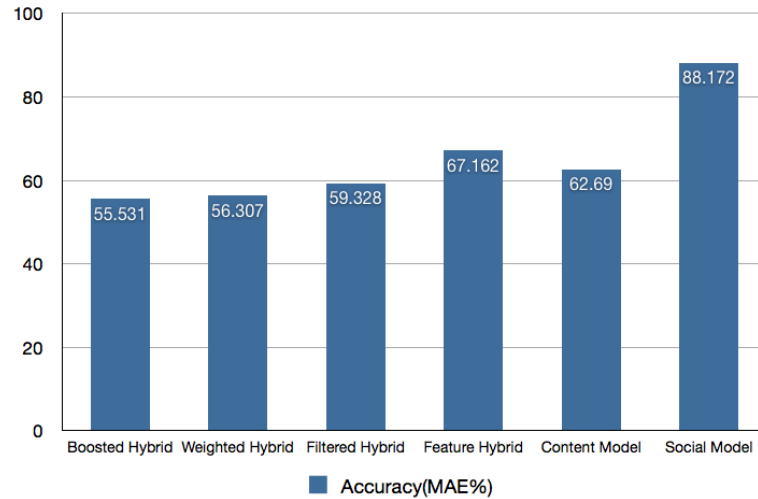


Figure 4.5: Comparison of predictive accuracy over the manually collected credibility ratings for each model.

predict the user-provided credibility scores from the online study. Preliminary experiments were performed using Bayesian classifiers (and a range of others) to learn a model based on the features of each prediction strategy. For the full experiment a J48 tree-based learning algorithm was used, firstly since it performed well in preliminary tests, and secondly to allow for comparison of results with Castillo et al.’s evaluation in [29]. Predictions were run on a training set of 591 tweets with annotated credibility scores. A 10-fold cross-validation was applied, and training sets were separate from test sets. The algorithm classified each test instance into one of two credibility classes. To clarify, we note that all predictions were made at the tweet level, that is, if a strategy (such as the standalone social model) predicts credibility at the user level, the evaluation applied this approach to predicting credibility of a tweet by that user. Class instances were evenly distributed in the training sets. For each strategy, the mean absolute error between the predicted rating and the user provided rating was recorded.

Figure 4.5 shows the results of this evaluation for each strategy. The content-based and hybrid models performed reasonably at the prediction task, but were far outper-

formed by the social model, which achieved an accuracy of 88.17%, an improvement of 11% over the feature hybrid which was the next best performer (statistical analysis shown in Figure 4.6). The content-based approach scored an accuracy of 63% while the hybrid approaches ranged from 56% to 67%. Our initial expectations were that the simple rule of “more features, better prediction” would apply across this study, but in this case our findings have indicated otherwise, since the social model outperform the hybrid and content-based methods significantly. The relatively poor performance of the content based model (67%) can perhaps be attributed to the fact that tweet text is short and does not always contain sufficient information to make a credibility judgement. The feature-hybrid method exhibited a small improvement in accuracy ($\tilde{10}\%$) over the next best hybrid strategy, which was the filtered approach. An overview of the statistical output from the J48 learner process is provided in Figure 4.6 for our best performing method, showing a correct classification of 902 of the 1023 instances, yielding 88.172% accuracy. The content-based (and therefore, hybrid) approaches rely on tweet text, whereas our social model relies on rich interconnections in the twitter network, including dynamic information flow metrics (retweets). Our findings indicate that the underlying network and dynamics of information flow are better indicators of credibility than text content.

Castillo et. al’s examination of credibility in Twitter produced similar accuracy scores to the above (8% less accurate than our best performing social model result), with “precision and recall in the range of 70-80%”. Several key differences make it infeasible to perform a fair comparison of classification accuracy however. For example, [29] analyses groups of “newsworthy” tweets, whereas our analysis focuses on “credible” individual tweets or users, as per our earlier definition. Furthermore, our analysis are focused in a topic-specific domain consisting of a different set of users and tweets.

```

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      902          88.172 %
Incorrectly Classified Instances    121          11.828 %
Kappa statistic                    0.8222
Mean absolute error                 0.0771
Root mean squared error             0.2162
Relative absolute error             22.8584 %
Root relative squared error         52.6661 %
Total Number of Instances          1023

=== Detailed Accuracy By Class ===
                TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
                0.901   0.009   0.966     0.901   0.932     0.966    1.0
                0.946   0.061   0.924     0.946   0.935     0.965    2.0
                0.896   0.083   0.816     0.896   0.854     0.936    4.0
                0.214   0.02   0.387     0.214   0.276     0.876    5.0
Weighted Avg.   0.882   0.054   0.872     0.882   0.875     0.952

=== Confusion Matrix ===
  a  b  c  d  <-- classified as
200 21  1  0  | a = 1.0
 5 424 17  2  | b = 2.0
 2 12 266 17  | c = 4.0
 0  2  42 12  | d = 5.0

```

Figure 4.6: Statistical results from a J48 tree learner for the best performing credibility prediction strategy (Social Model)

4.3.4 Discussion

The study of credibility models presented here is by no means exhaustive, and we believe that there are still better prediction strategies to be found. It appears from our evaluations that while stand alone social or content-based approaches fair reasonably well at predicting user provided credibility ratings, they are outperformed by hybrid methods which combine features from both, ultimately basing credibility assumptions on a larger window of information. Our evaluation answers the research question posed in the introduction, that credible tweets can be automatically detected with high accuracy (88% for our social model). Accurate detection of credible information in Twitter has many practical implications. For example, automatic filtering/ranking of twitter feeds based on credibility, spam detection, automatic recommendation of credible information [104] and identification of key players in information dissemination, which can be useful in assessing/predicting situations of social unrest such as the “occupy” movements and the London riots for instance.

There are a significant number of possible next-steps for this research. We believe

that improvements could be made by incorporating interpersonal factors such as the credibility that exists between users, also known as a trust relation. Such metrics can be incorporated both at the positive and negative levels (distrust), and have been shown to be useful for finding credible information in microblog domain. Furthermore, we are interested in evaluating the mechanisms presented here in a real world system, to elicit significant user feedback on the credibility of information in a real-world information consumer context, as opposed to the simple user survey approach presented here. This includes considering the role of interfaces and interactions that communicate credibility to, and elicit credibility data from real users.

4.3.5 Summary

As with most interactions on the Social Web, the window of information upon which we can make credibility judgement on Twitter is limited. As this forum becomes more popular, it becomes increasingly important to investigate new models for assessing credibility of the information it distributes. In this study, we presented three computational models for assessing such credibility, using social, content-based and hybrid strategies. The models were evaluated on 6 collections of tweets about current topics, including the associated social network information for each tweeter, as provided by the Twitter streaming API. An automated analysis of the predictive ability of each model was performed, predicting on both empirical “retweet” data, and on a collection of manually assessed tweets collected in an online user survey. Results showed that the social model outperformed both content-based and hybrid models, achieving a predictive accuracy of 88.17%, compared with 62% and 69% for content-based and the next best performing hybrid (weighted strategy) respectively.

4.4 Modeling User Competence by Mapping Theory to Practice

Psychologists and social scientists have developed and established models of human competence, credibility, trust and skill over many years. Currently, much research is being conducted by computer scientists to evaluate these human-behavioral aspects using real-world data from Twitter and other sources. It is a well known fact that where user-generated content exists, there is always a large amount of noisy or otherwise useless data. A key challenge to harnessing Twitter as a information source, is the ability to find relevant, reliable and trustworthy users to follow. Computer scientists in the fields of search and information retrieval (e.g.: recommender systems) have attempted to address this problem in other domains for several decades [176], while Behavioral scientists (Psychologists, Cognitive scientists, Social scientists) have studied the concepts of trust, reliability and competence for a far longer period of time, and have developed established theory for identifying and classifying these characteristics, both at the human level and the information level. [91, 177] While many studies of Twitter in the computer science literature attempt to model and mine for these characteristics [29, 50, 110], their models and algorithms tend to be formulated in an ad-hoc manner, without strong grounding in established theory from the human behavioral sciences.

This study presents a framework for mapping existing models of human competence and skill onto a real world streaming data from a social network. An example mapping is described using the Dreyfus model of skill acquisition, and an analysis and discussion of resulting feature distributions is presented on four topic-specific data collections from Twitter, including one on the 2014 Winter Olympics in Sochi, Russia. The mapping is evaluated using human assessments of competence through a crowdsourced study of 150 participants.

4.4.1 Mapping Theory to Practice

This study describes an experimental framework to map and validate established models of human behavior with the Twitter network and the information that flows within it. If applied successfully, such a framework has three clear benefits. First, it can serve as a form of validation for existing theoretical models by applying them at scales that were previously unattainable. Second, it can help analysts to constructively reason about observed phenomenon in the real world data. Lastly, it can be used as a guide to improve design of search and recommendation applications that attempt to relieve the information quality and overload problem.

Mapping of complex theoretical models of human behavior to observed behaviors in Twitter is clearly not a trivial task. The examples shown in the following sections all require a level of interpretation and a common sense reasoning about the links between factors in the model, and features and indicators in the Twitter information network. For the purpose of generalization we highlight the following steps for integrating an arbitrary human behavioral model with the network and associated data from Twitter, and follow this with an example implementation of the general process.

- *Task Identification and Analysis* What are the information requirements? What data elements from Twitter API can provide insight?
- *Model Selection* Is there a model in the behavioral/social science literature that is relevant to the task?
- *Feature Selection* What are the best features in the social network that may be useful indicators to the model?
- *Interpretation and Mapping* How should the features be related to the model itself?

- *Model Building and Validation* Train a prediction model using the mapped feature set and validate against a test set of annotations, or other available ground truth data.

In the remainder of this study we detail the above mapping procedure using an example task and an established theoretical model over four large current event data sets crawled from Twitter. Since identification of reliable information is such a critical aspect of today’s social web, we have chosen the following as an example task: *can we predict that a Twitter user will provide information about a target topic in a competent way*. Since Twitter is still a relatively young platform, and many users are still unfamiliar with the full scope of its operation and use, we have borrowed a model of competence from educational psychology known as the “Dreyfus Model of Skill Acquisition” [91] as a working example that to our knowledge has not previously been applied to social web data.

4.4.2 Setup and Data Collection

In this section, we will describe the experimental setup for our evaluation, particularly the crawling process and the collected data. Table 4.3 shows a summary of all data used in our evaluation, and Figure 4.7 shows an overview of the crawling process for users and topics. The larger circle denotes a set of messages gathered during a retroactive crawl using keywords that emerged after a period of time had elapsed since the initial crawl, but were still deemed to be a part of the core topic.

Data Collection

To allow for comparison of feature and model behavior, three different data sets are used in our evaluation. The first data collection is centered around the 2014 winter

| Collection | # Users | # Msgs | Keywords | Hashtags | Example tweet |
|----------------------|-----------|-----------|--|--|--|
| <i>boston</i> | 357,152 | 460,945 | marathon, pray, suspect, victims, bomb, police, hit, shrapnel, doctor, pellet, running, die, affected, rip, explosion, swat, blood, bombings, fbi, tragedy, donate, watertown, arrest, kill, injured, runner, hurt, donors, dead, identified | #bostonmarathon, #prayforboston, #boston, #prayersforboston, #watertown, #bruins | RT @Channel4News: There have been no arrests made yet after the bombings at the #BostonMarathon - US sources. #c4news |
| <i>boston strong</i> | 62,461 | 120,442 | affected, bostonisback, bostonstrong, boylston, charitymiles, donate, fbi, flyers, fund, help, honor, hope, marathon, memorial, oneboston, onefundboston, police, silence, spell, strength, strong, support, donors, tribute, victims, blood, bomb, doctor, tragedy, dead, rip, pray, hurt | #bostonstrong, #oneboston, #copley, #bostonisback, #prayforboston | @Nicolette_O Thank you for your support of the original #BOSTONSTRONG campaign, Nicolette! Nearing \$400K raised for The One Fund Boston! xxx |
| <i>sochi</i> | 4,305,508 | 9,521,089 | sochi, olympic, winter, female-olympians, games, gold, team, russia, hockey, medal, opening, usa, athletes, figure, canada, win, men's, ceremony, skating, ice, stray, putin, women's, gay, sport, won, ski, live, slope, skater, world | #sochi, #olympics, #sochi2014, #sochiproblems, #wearewinter, #sougofollow, #olympics2014 | RT @Bobby_Brown1: In air shot on the #Olympic slope course. Jumps are huge. Gonna be fun http://t.coXCQz90k1Eb |

Table 4.3: Overview of three data collections used to evaluate the mapping framework.

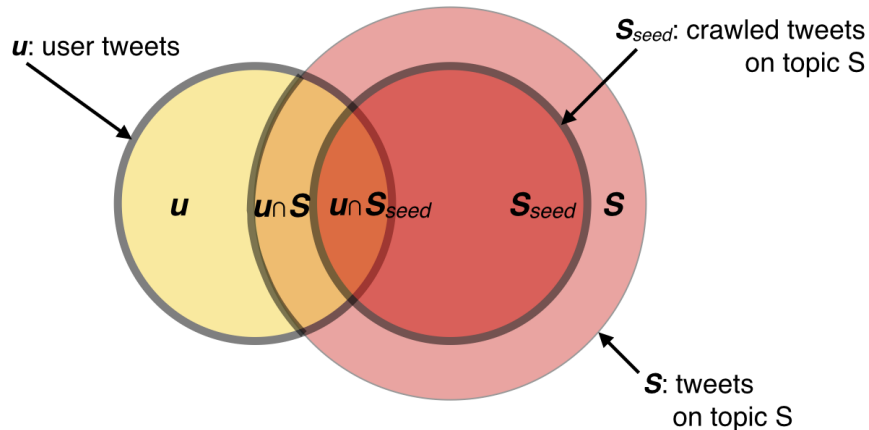


Figure 4.7: Overview of the crawled set of users and topics. Set S_{seed} represents the initial seed crawl from a key hashtag. Set S represents an expanded topic crawl to incorporate additional hashtags that evolve over the course of the event. Set u represents the set of all tweets from users who exist in S

| SKILL LEVEL Mental Function | NOVICE | COMPETENT | PROFICIENT | EXPERT | MASTER |
|---------------------------------------|------------------------|--------------------|-------------------|-------------------|-----------------|
| Recollection | Non-situational | Situational | Situational | Situational | Situational |
| Recognition | Decomposed | Decomposed | Holistic | Holistic | Holistic |
| Decision | Analytical | Analytical | Analytical | Intuitive | Intuitive |
| Awareness | Monitoring | Monitoring | Monitoring | Monitoring | Absorbed |

Figure 4.8: Overview of the Dreyfus model of skill acquisition. A component mental function is represented on each row and associated skill levels are shown on the columns. The horizontal arrows on each row represent the change in an observed mental function that facilitates an increase in the skill level represented in the model.

| <i>Function</i> | <i>Non-competent State</i> | <i>Competent State</i> | <i>Corresponding features</i> | <i>Other possibilities for features</i> | <i>Description</i> |
|-----------------|----------------------------|------------------------|--------------------------------|---|---------------------------------------|
| Recollection | Non-situational | Situational | $S[(u, t_0)] - avg(S)$ | specific #ht ↔ non-specific. writing of content | Adaptation to context (time specific) |
| Recognition | Decomposed | Holistic | Fraction of T that is in u | – | Coverage of topic T by user u |
| Decision | Analytical | Intuitive | $Opin(u, T)/U_{Opin}$ | – | Opinion and Sentiment of u in T |
| Awareness | Monitoring | Absorbed | Fraction of u that is in t | – | Involvement/Immersion in a topic T |

Table 4.4: Interpreted mappings between the Dreyfus model and a set of Twitter features

olympic games in Sochi, Russia. Data was crawled for approximately three weeks using a variety of keywords shown in Table 4.3. Sochi was chosen as a potentially interesting data set because of the diversity of cultures involved, and because of the associated excitement, politics and availability of concrete ground truth data in the form of event results.

Our second and third data sets are related to the terrorist attack that occurred during the 2013 Boston Marathon. The larger of the two collections was collected about the event itself, using the popular hashtag “#boston”. In this case, the data crawling began an hour after the event occurred and continued for two weeks. The second data collection was about the aftermath and recovery movement, crawled using the keyword “#bostonstrong” This was also crawled for approximately two weeks.

Theoretical Foundation

To exemplify the mapping process, we have chosen to borrow a model from the field of educational psychology known as “the Dreyfus model of skill acquisition” [91]. Since Twitter is a relatively new phenomenon, many of its users are still learning about the complex information, information flow, and network structure that Twitter supports, so we deemed this competence-based model of skill acquisition to be a reasonable example. Ideally, the generalizable framework we are describing will support many other established models of credibility, competence, trust or other factors that influence human decision-making, provided that appropriate mapping steps can be performed.

Dreyfus model of skill acquisition The Dreyfus model of skill acquisition describes the process of human skill acquisition in 5 different levels. This model was first introduced by the brothers Stuart and Hubert Dreyfus [91], and is established in the fields of education and operations research. The model is based on the four different transitions that define boundaries between five binary states of mental function during human learning. The original model, as can be seen in Table 4.4, is based on the three scenarios that show progression of a through each of the transitions, respectively. Table 4.4 suggests one of many possible mappings to a set of observable features in the Twitter based on expert interpretation of both.

Mapping

Now that we have selected a model, the next step is to study the meaning of each component within it, and formulate a reasonable analog in the behavior of an available set of Twitter features. A discussion of all such features is not possible here. The feature sets that we consider are discussed in Sikdar et. al [49], especially in Tables I and II of [49]. First it is necessary to define the network, topic, users and associated features

more concretely: For the following discussion, we view the Twitter domain as a triple (S, U, T) , where $S = (s_1, s_2, \dots, s_n)$ is a set of tweets crawled about a target topic. U is the set of users (u_1, u_2, \dots, u_n) who have at least one tweet in S . Additionally we define T a vector of event timestamps representing when messages in S were posted. This is given by $T = (t_1, t_2, \dots, t_n)$. Furthermore, each topic S can be represented by its component hashtags, $S_{hash} = (h_1, h_2, \dots, h_n)$. A notable property of S_{hash} is that the vector emerges over the values in T . Last, we define S_{seed} as the subset of S , gathered from the earliest emergent hashtags in S_{hash} .

Importantly, the mapping procedure we discuss here is simply an example to demonstrate the process. Mappings between a complex network and a complex behavioral model obviously require a degree of manual interpretation. Figure 4.8 illustrates a general form of the Dreyfus model, highlighting four key mental functions and the related competence levels. Table 4.4 shows the mental function on the leftmost column, followed by the associated indicators of competence or non-competence. The third row is the critical component, showing the analog feature combinations in Twitter. This is followed by other notable analogs and a text description of each feature. Our approach first looks at behavioral features in Twitter that could *potentially* serve as an indicator of each state. First we will describe the reasoning behind each mapping, and in the following section we present an evaluation of the behavior of each mapped feature, further indicating its potential to measure competence.

To recap, we are interested in evaluating the competence of information providers in Twitter with respect to a target topic. This covers both authorship and information propagation alike. Within this context, we interpret recollection in a topic as the ability to think back into the topic history, in the sense of maximizing ones posterity in the target topic. To approach this computationally, we consider the sequence of event times T of topic S from our earlier definition, and attempt to gauge where individual users

reside with respect to the normal for the topic. For example, if Alice’s history goes farther back than Bob’s, she has a greater degree of posterity, and perhaps this can be an indicator of competence. We compute this for every user simply as the earliest timestamp of a tweet that they have made in topic S . This is compared against the average timestamp of all users’ first tweets (s_0 within the topic, as shown in Equation 4.7 below). In a perfect mapping, we could simply examine the distribution graph of this feature over all users and segment it using a threshold value to determine the boundary between the competent and non-competent state. In this case, the boundary between non-situational (general) and situational (specific, detailed) recollection. The following section describes evaluations of this type for all features on all three data collections.

$$recollection(u, S) = T[(s_0, u)] - \frac{\sum_{i=1}^n T[(s_0, u_i)]}{n} \quad (4.7)$$

The next function of the Dreyfus model in Table 4.4 is “recognition”. Assessing whether a human’s recognition of a topic is in a decomposed or holistic state can be very difficult, depending on the complexity of the topic being analyzed. For our simple computational model, we treat recognition of a topic S by user u as the degree of *coverage* of S by u . This could be simply computed as the sum of all messages in u that are related to S , divided by the total number of messages in S . However, sparsity, irrelevant messages and other noise in the topic can weaken the link to the user profile. A better way to approach this mapping could leverage a) the set of hashtags in S_{hash} that describe the topic, or b) the set most frequently occurring terms as a more well-defined descriptor of the topic. We compute the hashtag-based coverage as Equation 4.8 below.

$$recognition(u, S) = \frac{S_{hash}(u)}{S_{hash}(all)} \quad (4.8)$$

The “decision” function in the Dreyfus model is treated differently in our mapping. Dreyfus categorizes this into analytical decision-making and intuitive decision-making, with the latter being an indicator of expertise within the topic (see Figure 4.8). Deciding whether an individual is making analytical or intuitive choices has been the subject of many research papers in itself, e.g. [178], so again, we will need to simplify here for the purposes of discussion. Our computational model looks to *sentiment* as an indicator of decision making potential. This approach has been studied and validated by many researchers, For example, O’Connor et al [179] found that decisions to purchase products (consumer confidence) and decisions about elections [180, 179] can be predicted by examining frequency of sentiment-related word usage in Twitter posts.

In particular, we examine three aspects of sentiment:

- *Degree of Subjectivity* If a user demonstrates the ability to form subjective opinion on a given topic, it *may* point towards a higher level of competence. To assess this, we borrow a subjectivity lexicon from the Opinion Finder tool described by Wilson et al. in [181]. Each user u is represented as a bag of terms and a count is performed for terms that occur in the lexicon. The resulting value is our subjectivity score for that user. At a finer grained level, we focus on words that imply personal preference (e.g. cool, excellent, awesome, etc.), and on expressions / idioms that imply opinion (e.g. I think, I suppose, I believe etc.).
- *Sentiment Intensity* Intensity of sentiment is a good indicator of knowledge about a topic [179]. In our model, this is measured as a simple count against the sentiment lexicon from NLTK [182].
- *Sentiment Polarity* Our third sentiment metric examines sentiment of user u as a polarized scalar $sp = [-1\ 1]$ by comparison against negative and positive sentiment lexicons from NLTK.

While the Dreyfus model from Figure 4.8 shows a single factor for “Decision”, we choose to analyze the three sentiment factors separately in the analysis that follows, in case varying behaviors can be observed. After the initial feature behavior analysis they can be pruned or combined in some way to produce a single attribute.

The final function listed in Table 4.4 is the concept of awareness. According to the model shown in Figure 4.8, when a human’s awareness transitions from persistent monitoring to an absorbed level, it is an indication of mastery of a particular skill. Put another way, this transition occurs when actions become “second nature” instead of as a result of careful fine-grained analysis of rules and inputs. Again, this is a potentially difficult concept to map onto a simple computational model, since one essentially needs to be at the mastery level in a given topic to recognize such intuitive actions. In this example, our goal is to evaluate competence of an information provider in a target topic. As a simple proxy for detecting the transition in awareness between monitoring and absorbed, our computational model focuses on the degree of *immersion* of a user in a topic. That is, the percentage of the user u ’s profile that is dedicated to a topic S . One problem with this proxy is that it does not facilitate fair comparison between users—a property that is required for the feature behavior analysis that follows. Consider our Sochi Olympics dataset for example: If the official winter olympic feed has 1,000 tweets all about the event, and a random user (Joe) has 10 tweets that are also about the event, this metric would produce the same score for both profiles. To control for this, we introduce a weight w based on the number of tweets in the profile, shown here as Equation 3:

$$awareness(u, S) = \frac{u(S_{hash})}{u(all)} \times w. \quad (4.9)$$

This concludes the interpretation and mapping phase of the framework. Now, we

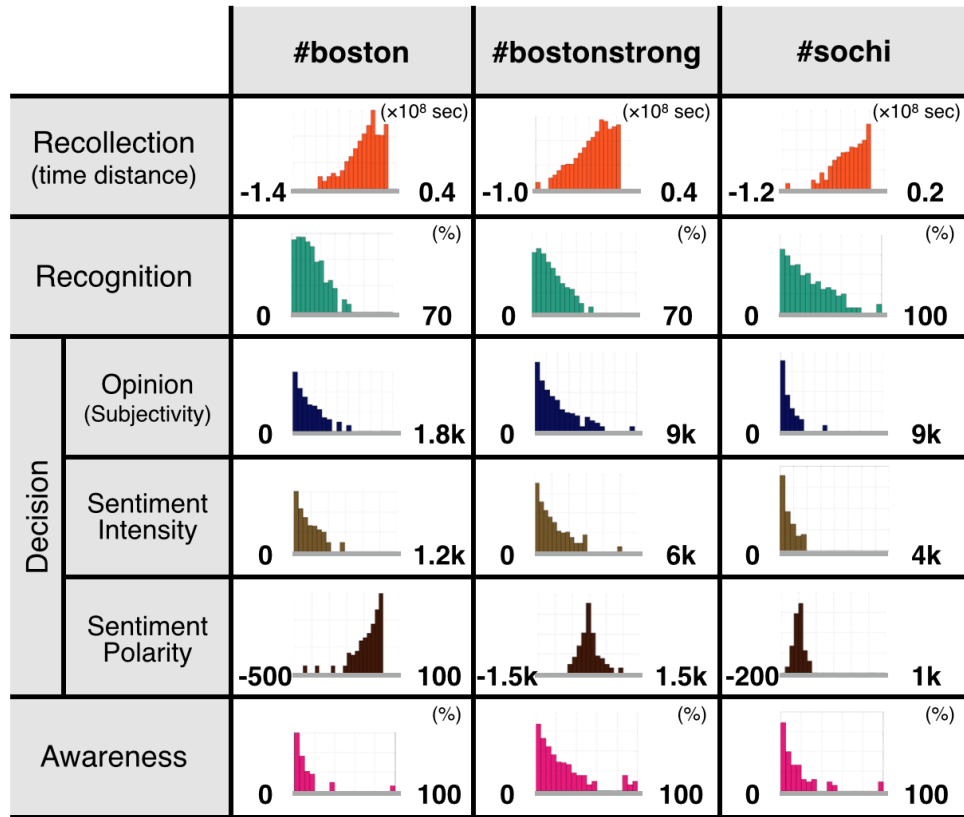


Figure 4.9: Analysis of behavior for the mapped feature set (Dreyfus model representation). Each row represents an individual feature, and each column represents a data set. The “decision” feature has been broken into three sub-features: opinion, sentiment intensity and sentiment polarity, shown on rows 3-5. All values are shown in percentages with the exception of the first row, which is a time-based value (seconds).

arrive at a computational model in the form of a set of observable features that maps, albeit loosely, to the theoretical model in Figure 4.8. The next step in the procedure is to evaluate the behavior of these features to determine distribution curves and see if we can identify reasonable thresholds that can correspond with the phase transitions of the Dreyfus model, shown in Figure 4.8.

Feature Analysis

Now that we have described the computational model we must assess its potential to predict human behavior in real world Twitter data. To achieve this we compute the 6 individual features described in the previous section on each of the three data collections (Boston, BostonStrong and Sochi). All of the features described can be considered user-based features, that is, they are attached to a single user, as opposed to a single message (see [30, 110, 49] for a discussion on user and message-based features). In order to examine potential of a feature for predicting competence of a user as a provider of information about a topic, we take the following approach: First we compute the individual feature value $f \in F$ for each user $u \in U$ on each data set S . Next we plot a distribution $dist(f, U, S)$ for all features in F and all three of our topics. Results of this analysis are shown in Figure 4.9, and arranged as follows: each row represents a computed feature, identified by the title on the left side. Each column represents a data collection, identified by the seed hashtag in the header row. This arrangement of distributions is useful since allows us to quickly compare across data collections and across features. All values are shown in percentages with the exception of the first row, which is a time-based value (seconds).

Let us first discuss the behavior of individual features, with a view to locating thresholds that may yield information about competence of users as information providers about the topic. The recollection feature shows distribution of users as a deviation from the mean time that the topic was discussed on Twitter, meaning that the leftmost group are early adopters, those at the peak are discussing the event as it is happening, or close to it in time, while the users to the right are talking about it after-the-fact. The users on the right of the peaks have the important benefit of hindsight. Note that for the Sochi data set, the gaussian curve is cut off because the data runs up to the time of writing of

this article. Table 4.3 shows the crawl times for each plot. Both Sochi and BostonStrong data sets show clusters of early adopters on the negative slope –an interesting subset for further analysis.

For the recognition/coverage feature all three collections show clusters of accounts with relatively high coverage. Manual inspection of these showed that they were official, government, media or other dedicated accounts to monitor the event during the crawling time, and are therefore a potentially useful information source. The decision feature shows the most interesting result across the three collections. Clearly there is a large amount of sentiment and opinion expressed about the Boston and BostonStrong collections, and the dedicated account clusters are clearly visible on the right. Looking at the sentiment polarity shows a more detailed account of the public feeling at the time. During the event time, the sentiment was all negative relating to the bombing incident, but when we look at the polarity score for the aftermath movement BostonStrong, we see clear signs of positive sentiment relating to the topic. These are likely tributes and other encouraging, hopeful messages stemming from the tragic event. For the olympics data, there is a more even distribution, which is intuitive given the winners and losers at the games.

Last, the awareness metric examined the immersion of a user in a topic, but weighted the score based on the number of tweets in T . These plots (bottom row of Figure 4.9 show a few accounts that are far more dedicated than the others. These accounts are again, likely to be dedicated to covering the topic for one reason or another.

In summary, the best values for thresholding these graphs to best identify the transitions from Figure 4.8 are likely to be in the areas that segment small clusters from the remainder of the users. The following section outlines an experiment to evaluate the competence of users that exist within the extremities of each of the feature distribution plots from Figure 4.9.

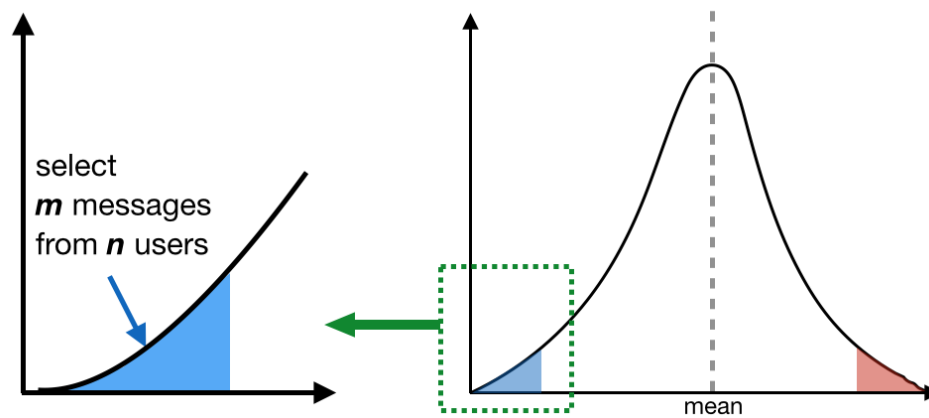


Figure 4.10: Procedure for sampling user profiles from each of the 6 feature distribution graphs for evaluation in the crowd sourced experiment.

4.4.3 Evaluation

Thus far have described a mapping process between an abstract behavioral model from the field of educational psychology, and a measurable set of features in the Twitter network. We have performed an analysis of the behavior of each individual feature. The next step in our general framework is to evaluate data samples from the distributions in an effort to find useful thresholds for building a prediction model. Figure 4.10 illustrates the process on a sample distribution. m messages were sampled from n users from the extremities of each distribution plot. In this experiment, we chose $m = 2$ and $n = 3$ for each of the 6 features on each of the Sochi data collection and gauged perceived levels of competence, newsworthiness and topic-relevance in a crowd-sourced study.

Feature-based Competence Assessment

A study was run using Amazon’s Mechanical Turk crowdsourcing tool. In total, 150 participants completed the study. Participants were 62% Male, 38% Female, ranged in age from 18 to 58 and took an average of 12 minutes to complete the study. Most participants reported that they had strong reading ability and had at least a Bachelor

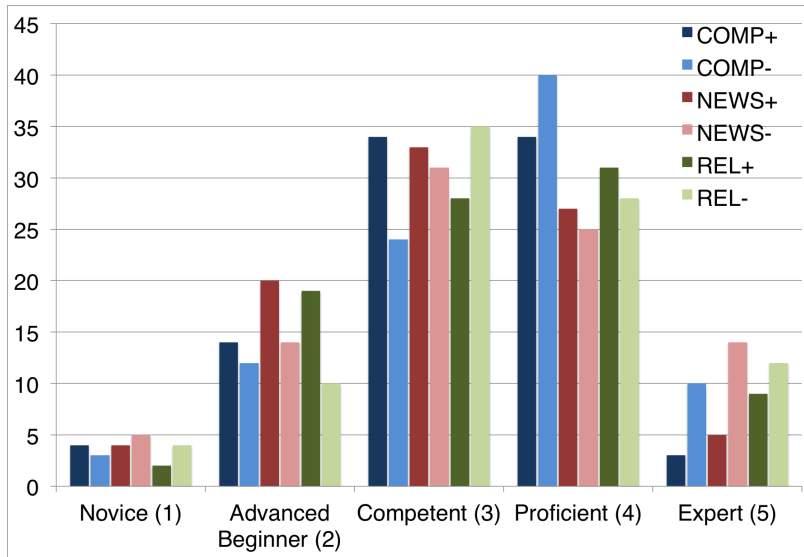


Figure 4.11: Distribution of ratings in AMT study for Competence, Newsworthiness and Relevance on the Sochi Winter Olympics data collection.

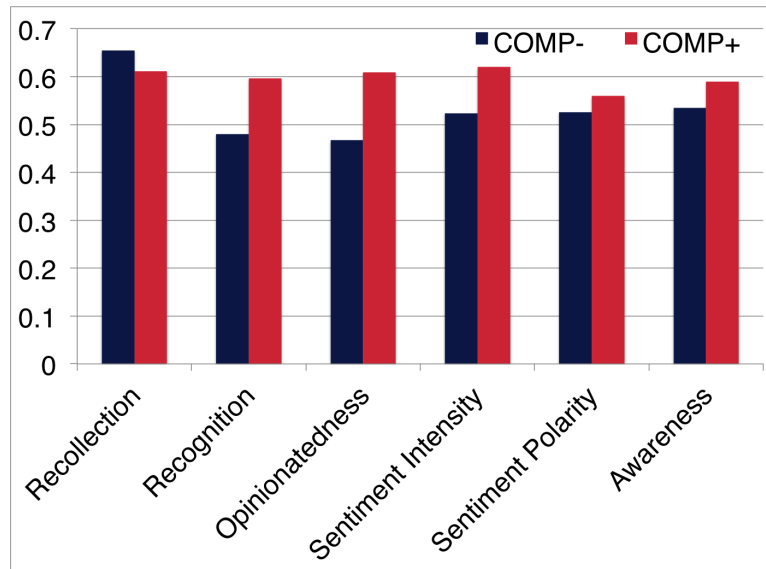


Figure 4.12: Comparison of ratings for each feature grouped by the users sampled from *COMP+* and *COMP-* areas of the feature distribution curves. This graph was computed on the Sochi data collection.

level college education. A small payment of 50 cents was provided for completed studies. Sampled messages were presented to AMT evaluators in a simple web form. Participants were asked to read groups of three messages (coming from an individual user), and evaluate that user's competence as an information provider in the target topic. Competence ratings were provided on the 5-point Dreyfus Scale from Novice to Expert. In addition to competence, newsworthiness and topic-relevance was also assessed. Table 4.5 lists all of the metrics that were recorded in the study. Here we focus only on the competence annotations (*COMP+* and *COMP-*). Figure 4.11 shows the mean competence score (y-axis) on the Sochi data set for each feature in our mapped model (x-axis). The x-axis is grouped by *COMP+* and *COMP-*, reflecting the users and messages sampled from the right and left sides of each feature distribution curve in Figure 4.9 and also illustrated in Fig 4.10.

Figure 4.9 shows some interesting results for each feature. The only instance where *COMP+* is lower than *COMP-* is on the recollection feature. In other words, the users selected from the left side of this feature distribution, i.e. the early adopters of the topic, received higher competence scores than those who began tweeting about the topic later in its evolution. This is a good indication that recollection is a useful feature for measuring competence in Twitter. The second group in Figure 4.9 (recognition) shows us that those users who covered a greater portion of the topic were considered to be more credible. The largest difference between competence ratings is for the opinionatedness feature. Here we can see that users in *COMP+* (right side of distribution curve, and highly opinionated) were rated as more competent than those in the *COMP-* group (left side of distribution, less opinionated), with a relative increase of 35.5%. The smallest difference was shown for the sentiment polarity group (12% relative increase for *COMP+* group), meaning that polarity of sentiment was less correlated with the competence annotations than intensity of sentiment, coverage of a topic or opinionatedness.

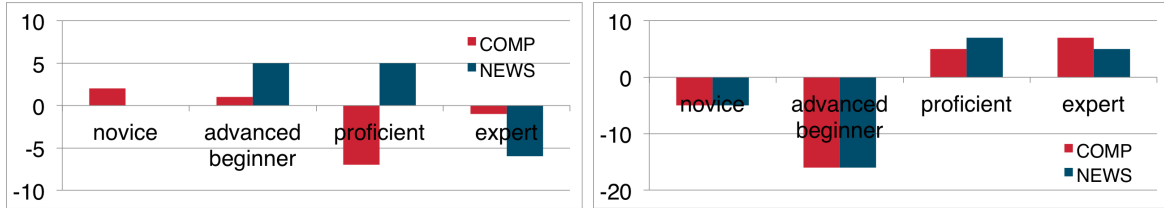


Figure 4.13: Differences between AMT competence ratings for the Recall and Opinionatedness features. Differences shown for $COMP+$, $COMP-$, and $NEWS+$, $NEWS-$. The x-axis shows each rating bin from novice to expert.

| Series | Description |
|---------|---|
| $COMP+$ | Competence score for tweets on right side of feature distribution |
| $COMP-$ | Competence score for tweets on left side of feature distribution |
| $NEWS+$ | Newsworthiness score for tweets on right side of feature distribution |
| $NEWS-$ | Newsworthiness score for tweets on left side of feature distribution |
| $REL+$ | Relevance score for tweets on right side of feature distribution |
| $REL-$ | Relevance score for tweets on left side of feature distribution |

Table 4.5: Description of recorded results from AMT study.

Figure 4.11 shows the general distribution of the ratings from the study, for each of the metrics in Table 4.5. This trend was evident across all data sets and features evaluated in the study, with mean ratings between 3 and 4 on the 5 point rating scale. Figure 4.13 shows a different perspective on the AMT data. Here, we focus on the trend in the difference between $COMP+$ and $COMP-$ across the rating bins from novice to expert. The upper chart shows the differences for the recall feature. This tells us that there are far more early adopters of the topic in the proficient and expert bins than in the the novice and beginner bins. Interestingly, this was a significant trend for the competence annotations, but not for the newsworthiness annotations. The lower chart in Figure 4.13 shows the opposite trend for the opinionatedness feature: more highly opinionated users exist in the proficient and expert bins than the beginner and novice bins. These trends show that opinion and adoption-time (time of first tweet about the topic) are strong indicators of competence, but less so of newsworthiness.

4.4.4 Summary and Discussion

In the field of information system (IS), provenance of information, or information source, has long been studied by researchers as an important proxy of information quality. This study has presented a step-by-step generalizable framework for linking existing models of human behavior from the social and cognitive sciences with real world measurable features from the Twitter social network. The main assumption of our approach is that every user on the Social Web can be understood as an information provider. A user can either publish an original content or forward a piece of content produced by another user to other information consumers in the network, and such activity can be interpreted as the dissemination of information. In this vein, we attempt to gauge the degree of competence of the user by mapping an established conceptual model of competence to Today's Social Web. Specifically the research proposed 5 integration steps and provided a worked example using the Dreyfus model of skill acquisition as a representative model. Features were mapped to a computational model over the Twitter network and behavior of each feature was analyzed over three large data collections. A study of 150 participants evaluated the competence levels of users sourced from both poles of the feature distributions. Results and manual analysis indicate that there is potential in the distribution plots to identify useful (competent) information sources related to a particular topic. A feature-by-feature comparison outlined a range of interesting effects between competence ratings for users selected from the poles of the feature distribution plots for the Sochi data collection. As a follow up study the authors propose to compare against a range of other models from the behavioral sciences, and to combine the resultant features into a predictive model and run accuracy-based evaluations over multiple ground-truth metrics. In conclusion, while there are many assumptions in the mapping stages of the approach, the authors believe that the methodology can help both algorithm designers for the So-

cial Web and researchers in the behavioral sciences to better understand complex data interactions in Twitter.

4.5 Modeling News Content in Microblogs

In recent years the greater part of news dissemination has shifted from traditional news media to individual users on microblogs such as Twitter and Reddit. Therefore, there has been increasing research effort on how to automatically detect newsworthy and otherwise useful information on these platforms.

In this study, we present two novel algorithmic approaches—content-similarity computation and graph analysis—to automatically capture main differences in newsworthy content between microblogs and traditional news media.

For the content-similarity algorithm, we discuss why it is difficult to capture such unique information using traditional text-based search mechanisms. We performed an experiment to evaluate the content-similarity algorithm using a corpus of 35 million topic-specific Twitter messages and 6,112 New York Times articles on a variety of topics. This is followed by an online user study (N=200) to evaluate how users assess the content recommended by the algorithm. The results show significant differences in user perception of newsworthiness and uniqueness of the content returned by our algorithm.

Secondly, we investigate a method for identifying unique content in microblogs by harnessing network structure of the information propagation graphs. In this approach, we study how these two types of information differ from each other in terms of topic and dissemination behavior in the network. The results show that the majority of sub-graphs in the traditional group have long retweet chains and exhibit a giant component surrounded by a number of small components, unique contents typically propagate from a dominating node with only a few multi-hop retweet chains observed. Furthermore,

results from LDA and BPR algorithms indicate that strong and dense topic associations between users are frequently observed in the graphs of the traditional group, but not in the unique group.

4.5.1 Introduction

Over the last decade, microblogs have evolved from an online communication channel for personal use to a central hub for information exchange between users. On microblogging platforms, users produce or share information with friends or strangers. Recent studies revealed that the greater part of today's internet users rely on information on microblogs [165] (e.g. Twitter and Reddit) as a primary source of a wide range of information, particularly news. Accordingly, this new paradigm highlights the importance of automated tools that detect reliable and newsworthy information on microblogs.

Going beyond typical information consumers, professional journalists also admit to relying heavily on social media streams for their news stories [183, 12]. During the last decade, microblogs have been studied by researchers in communication and journalism as an essential news gathering tool and several guidelines are proposed⁴. Many users favor to browse microblogs such as Reddit and Twitter on a daily basis since these platforms provide personalized news content based on their previous browsing patterns. Recent research also highlights that traditional news outlets still play an important role in the provision of reliable, well curated news content [165].

However, news outlets are typically biased in some way or other, and do not always act as the best information filters in all cases. A recent study by [184] highlights the polarizing political bias that exists across most of the top US traditional news outlets. Despite the possibility for bias, we believe that curated news from a variety of sources can be leveraged to help identify and classify newsworthy messages in social media streams. In particular,

⁴http://asne.org/Files/pdf/10_Best_Practices_for_Social_Media.pdf

we propose a novel method for identifying *niche* user-provided topics from social media that is a) not reported in traditional curated news, and b) is newsworthy information. Figure 4.14 shows an overview of our first approach. Each data point represents a Twitter post, located on the x-axis by similarity to a target set of news articles, and on the y-axis by general newsworthiness of the message content. The distribution shows a linear trend indicating the correlation of newsworthiness and similarity to curated content, as we would expect to see. In this case however, we are interested in the highlighted “niche content” section in the top left of the graph, which contains those unique messages that are *not similar* to mainstream media, but do have newsworthy content based on other metrics. This content could be found through a series of text based search queries, but defining relevant keywords is difficult, and may potentially only uncover a given slice of the true overlap between the data sources.

To explore this concept, we study a variety of topics from 37 million Twitter posts and 6,112 New York Times articles and attempt to answer the following research questions:

1. **RQ1** How can we best detect newsworthy information in social media that is not covered by traditional media?
2. **RQ2** How do information consumers perceive the detected information?
3. **RQ3** How do the niche information get propagated differently from traditional news in the network?

In this study, we propose two distinct approaches to capture unique news content on microblogs. First approach is based on a variety of content-similarity metrics. Simply put, we compute different content-based similarity metrics on microblog posts and a corpus of traditional news articles. Using these similarity metrics with our newsworthiness scores computed on individual tweets, we can locate the niche (unique and newsworthy)

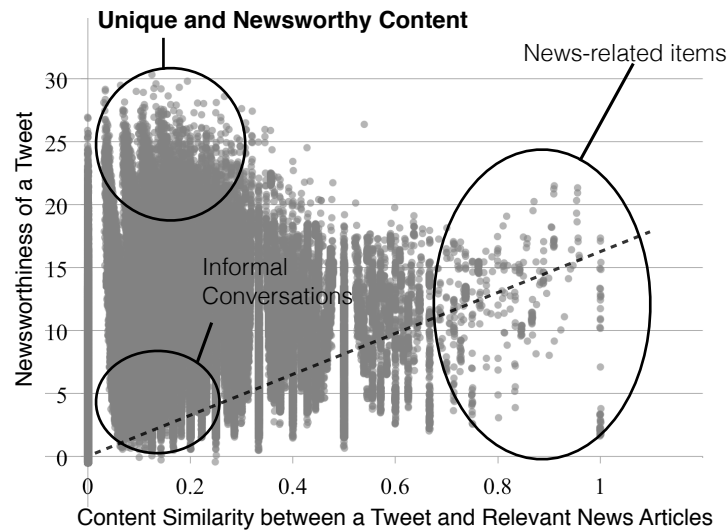


Figure 4.14: Overview of approach to filtering unique and newsworthy content. Y-axis tweet newsworthiness is computed from NLTK and from Human Evaluation. X-axis is tweet similarity to mainstream news.

contents and analyze them to find important features that can be utilized for developing automated detection algorithm. Specifically we describe two experiments: first, an automated evaluation is performed to test a variety of mechanisms that predict overlap between a microblog post and a corpus of news articles. These include manipulations on n-grams, part-of-speech tags, stop words and stemming techniques. A co-occurrence score is produced for each message, which is in turn compared to a set of manually annotated newsworthiness scores, combined with a content-based newsworthiness score. The different strategies are ranked by the resulting distance and the best approach is used for experiment 2. Manual annotations of newsworthiness were collected using a crowd-sourced study described in [49]. The second experiment samples data in various ways from the highlighted areas of Figure 4.14 for a range of topics and presents an A/B style questionnaire about newsworthiness, similarity to traditional media content, and personal focus to 200 participants in an online study.

Results of experiment 1 show that a simple n-gram approach with word-stemming but without stop word removal produced the most accurate approximation of the manual

annotations. Results from experiment 2 show that there is a significant difference in reported “similarity to mainstream news content” for messages sampled from the top left area of Figure 4.14 compared with a random sample from the right side, indicating that the method is capable of automatically identifying newsworthy content that is not covered by mainstream media.

To address **RQ3**, the second approach (network analysis on microblog news contents) has been demonstrated in Section 4.5.3. In this approach, we apply a variety of commonly used network measures of structural and functional connectivity to microblog information to unveil unique characteristics that represent both niche and generic news contents on microblogs. Particularly, two experiments are performed on the collection of 2.4M Twitter dataset to find the differences between the two groups (niche and traditional groups) in network topology (**Exp 1**) and topical association across users (**Exp 2**).

Results of **Exp 1** show that the majority of subgraphs in the traditional group have long retweet chains with a giant component surrounded by a number of small components. On the other hand, unique contents typically propagate from a dominating node with only a few multi-hop retweet chains observed. Furthermore, results from **Exp 2** indicate that strong and dense topic associations between users are frequently observed in the graphs of the traditional group, but not in the unique group.

The differences between the unique and traditional news groups that we found in this study will benefit future studies for intelligent and scalable algorithms to automatically classify or predict unique or interesting news in microblogs. We will discuss our future work and possible applications for which our model can be applied in Section 4.5.4 and 4.5.5.

4.5.2 Content Similarity based Approach

This section describes our approach to filtering unique and newsworthy content from microblog streams based on comparison with contents from mainstream media. According to the study in [142], Shoemaker claims that newsworthiness is not the only attribute which represents news. However, in this study, we adopt newsworthiness as the central indicator of news contents in general. Basically, we assume here that curated news articles are newsworthy. Our first approach exploits news articles as a reference to identify Twitter postings about a target topic that are newsworthy but are not the focus of curated mainstream news. We begin by exploring a set of mechanisms for computing similarity between a microblog post and a topic-specific corpus of news articles.

Data Collection

To examine real-world microblog messages and news contents, we choose “Twitter” and “New York Times” as representative examples for microblogging platforms and traditional media outlets. Both provide well documented application program interfaces (APIs)⁵ through which we can retrieve microblog messages or news articles as well as a rich set of metadata (e.g. keywords, embedded multimedia items, urls). With the two APIs we collected about 35 million (35,553,515) microblog messages from Twitter and 6,112 news articles from New York Times and other sources such as Reuters and Associated Press (AP). An overview of this data collection is shown in Table 4.6. Before the crawling stage, we selected major news events such as natural disasters, world cup and various political issues over the course of 4 years (2012 - 2015) to examine how both media differs from each other and see if there is topic-specific bias across different events.

⁵New York Times Article Search API: <http://developer.nytimes.com/docs>
Twitter API <http://dev.twitter.com>

Table 4.6: Overview of the data sets collected from New York Times and Twitter.

| topic | <i>world cup</i> | <i>ISIS</i> | <i>earthquake</i> | <i>hurricane sandy</i> |
|----------|------------------|-------------|-------------------|------------------------|
| tweets | 22,299,767 | 8,480,388 | 921,481 | 3,851,879 |
| articles | 4,097 | 422 | 329 | 1,264 |
| from | 6/24/14 | 1/20/15 | 1/20/15 | 10/29/2012 |
| to | 7/17/14 | 3/29/15 | 3/31/15 | 12/31/2012 |
| days | 24 | 69 | 71 | 64 |

We collected topic-specific data sets⁶ using related keywords to retrieve microblog messages and news articles from Twitter and New York Times databases. In particular, for Twitter data, we used the Streaming API to monitor transient bursts in the message stream while we collected regular data about the events.

Similarity Computation

A key challenge in this approach is to discover meaningful mappings between a short microblog post and a larger corpus of news articles. Since traditional text-matching mechanisms such as TF-IDF or topic modeling do not work well with short messages, a variety of simpler mechanisms were evaluated. Table 4.9 shows an overview of the mechanisms tested and their performance with respect to manually labeled “ground truth” assessments of newsworthiness. An initial pre-processing was applied to all messages to remove superfluous content such as slang and gibberish terms.

Word n -grams Next, a set of word n -grams as described in [185] were computed, varying n from 1 to 3. Part-of-Speech (POS) tagging was applied to identify potentially useful noun, verb, pronoun and adjective terms. A standard stop-word list was identified and systematically removed as shown in Table 4.9. A Twitter-specific stop-word list was compiled from a manual analysis of posts. This list contained platform-specific terms such as “twitter”, “rt”, “retweet”, “following” etc., based on a term frequency analysis. In total, 24 combinations of lightweight NLP techniques were applied to 4 topic-specific

⁶Dataset available upon email request

collections of twitter posts and NYT news articles. These are detailed in Table 4.9. Each method computed a content-based similarity score between a *single* microblog post and a larger collection of news articles.

For each topic studied, we obtained thousands of n-grams from the NYT article collection and use it as a corpus of news n-grams ($n = 1, 2, 3$). Next, we applied n -gram extraction on the entire tweet collection and computed the number of co-occurrences of n -grams from each post with those in the news n-gram corpus. To account for length deviation, this score (*Score*) was normalized by the total number of n-grams in each tweet.

Newsworthiness In this study, we apply a two dimensional approach to newsworthiness: (1) news term frequency in each tweet ($News_{Term}$) and (2) newsworthiness score labeled by real-world microblog users ($News_{User}$) in [0-5] Likert scale.

For $News_{Term}$, we compute number of tokens that contain news terms using the Reuters news word corpus in NLTK⁷ and divide this number by total number of tokens in each message.

$News_{User}$ is the human-annotated newsworthiness score, and is also normalized by the maximum score. Normalization is performed on both metrics in order to eliminate bias of different message sizes in tweets and take the average of the two metrics for Equation 4.10. Table 4.7 shows the selected set of similarity metrics that we employ in this study.

Strategy Selection

We define a simple inverse distance metric in order to evaluate our content-based similarity measure (*Score*) and select the best performer among 24 candidates. This metric is then applied to the composite sets of multiple metrics to select the best feature

⁷NLTK Reuters Corpus has 1.3M words, 10k news documents categorized <http://www.nltk.org>

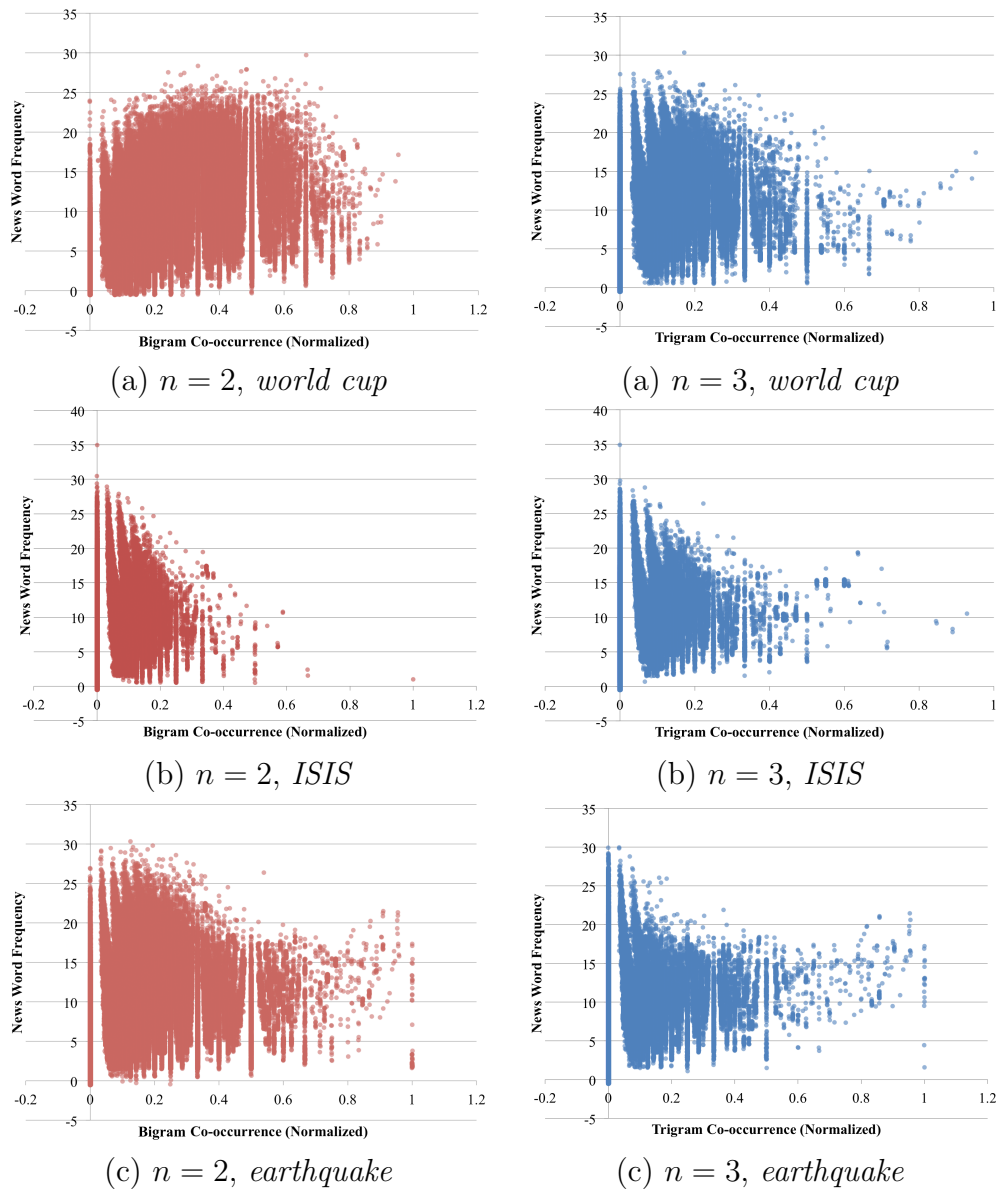


Figure 4.15: News word frequency on tweets and n -gram ($n = 2, 3$) co-occurrence with mainstream news articles (NYT) on different topics.

Table 4.7: Metrics analyzed in the study.

| Metrics | Nomenclature | Description |
|----------------------|---------------|--|
| n -gram Similarity | $Score$ | Number of n -grams that co-occur between news article corpus and a tweet |
| News Word Frequency | $News_{Term}$ | News word frequency with NLTK Reuters corpus |
| Newsworthiness Score | $News_{User}$ | Human annotated newsworthiness score [0-5] on a tweet |

based on the linear relationship between the similarity score and newsworthiness of a message. We discuss the procedure in detail in this section. Afterwards, we explain our evaluation method and procedure in Section 4.5.2.

Definition 4 Each event-specific data collection T contains N messages where $T = \{m_1, m_2 \dots m_N\}$, and we represent individual message as m where $m \in T$. Inverse distance of a message between newsworthiness and content similarity to news corpus is represented as $InvDist$.

$$InvDist(m_i, c_N) = \frac{1}{|News(m_i, c_R) - Score(m_i, c_N)| + 1} \quad (4.10)$$

Where $News(m_i, c_R)$ is:

$$News(m_i, c_R) = \frac{News_{Term}(m_i, c_R) + News_{User}(m_i)}{2} \quad (4.11)$$

Please note that c_N and c_R are a corpus of news articles on a topic and the Reuters news vocabulary corpus in NLTK, respectively.

Since we compare one strategy against others in the selection procedure, we use the average of inverse distance for a strategy over all messages, computed using Equation

4.10.

We apply a fractional function to the inverse distance metric in Equation 4.10. Intuitively, this approach maximizes gain in highly correlated messages and, likewise, penalize un-correlated messages between newsworthiness $News(m)$ and content similarity $Score(m)$. As briefly mentioned earlier in this section, we believe that both $News_{Term}$ and $News_{User}$ represent different aspects of newsworthiness. Unlike the n -gram co-occurrence ($Score$), which reflects the word-based association on a specific-event, $News_{Term}$, which is corpus-based news word frequency, represents topic-independent association between a microblog message and the Reuters news word corpus. To validate our inverse distance metric, we performed Pearson and Spearman correlation tests with the best feature selected and they are shown in Table 4.8. The best feature selection is summarized in Algorithm 1.

Algorithm 1: n -gram strategy evaluation (Best feature selection)

Result: Best performing strategy

initialization;

for all n -gram strategies **do**

for all message m where $m \in T$ **do**

$nGram \leftarrow \text{computeNGramScore}(m, \text{strategy}, \text{corpusNYT});$

$newsTerm \leftarrow \text{computeNewsTerm}(m, \text{corpusReuters});$

$news \leftarrow \text{mean}(newsUser, newsTerm);$

$similarity \leftarrow \text{computeSimilarity}(news, nGram);$

end

$\overline{similarity} \leftarrow 1/n \sum_{i=1}^N;$

end

$best \leftarrow \underset{\text{strategy}}{\text{argmax}} \overline{similarity};$

return $best$

As shown in Table 4.9, *unigram with stemmer only* feature has the highest correlation. Therefore, we select this feature for our user experiment and evaluation.

Table 4.8: Correlation coefficients between newsworthiness $News(m)$ (arithmetic mean of news word frequency and user annotated newsworthiness score) and n -gram co-occurrence score $Score(m)$ (all metrics normalized $[0,1]$)

| | Correlation Coeff. | 2-Tailed Test Significance |
|----------|--------------------|----------------------------|
| Pearson | 0.47063 | $< 1e - 10$ |
| Spearman | 0.41414 | $< 1e - 10$ |

Experimental Setup

In this study, we aim to identify unique newsworthy contents on microblogs that differs from those in mainstream news media like New York Times. So far we have explored different features based on content similarity metrics and text processing techniques. To validate our approach discussed in the previous section, we conduct an experiment including a crowd-sourced user study.

Random Sampling For the experiment, we randomly sample 10,000 tweets from each collection. This sampling task allows us to avoid possible scalability issue from the high volume of our data sets and fit the experiments and user study. We sampled tweets that are primarily written while events were taking place or shortly thereafter. For the NYT articles, however, we aggregate them together first before we compute similarity features.

Niche Content Extraction Our hypothesis is that, in general, newsworthy contents on microblogs do not completely overlap with mainstream news contents. In this study, the term “niche content” was coined for microblog exclusive (unique) newsworthy information. As the coined term implies, we assume that this type of information has a unique value and, thus, we believe that it is worth to investigate. The aim of this study is to find the unique characteristics of the niche content on microblogs and exploit our findings to provide a guideline to design more effective newsworthy information filtering algorithm in many applications.

| Avg # Terms in News | Avg # Terms in Tweets | # Co-occurrence | Stopword Removal | Stemming | Noun Only (POS-tag) | <i>n</i> -gram | Inverse Distance |
|------------------------|--------------------------|-----------------|---------------------|----------|------------------------|----------------|---------------------|
| 3,863 | 17.952 | 10.509 | N | N | N | 1 | 0.774 |
| 12,085 | 16.965 | 1.713 | N | N | N | 2 | 0.814 |
| 15,246 | 16.011 | 0.162 | N | N | N | 3 | 0.689 |
| 1,719 | 7.401 | 2.678 | N | N | Y | 1 | 0.777 |
| 4,596 | 6.532 | 0.144 | N | N | Y | 2 | 0.75 |
| 5,792 | 5.762 | 0.014 | N | N | Y | 3 | 0.714 |
| 1,596 | 17.952 | 3.868 | N | Y | N | 1 | 0.96 |
| 4,592 | 16.965 | 0.165 | N | Y | N | 2 | 0.758 |
| 5,790 | 16.011 | 0.006 | N | Y | N | 3 | 0.740 |
| 1,564 | 7.401 | 2.654 | N | Y | Y | 1 | 0.736 |
| 4,509 | 6.532 | 0.145 | N | Y | Y | 2 | 0.8 |
| 5,678 | 5.762 | 0.014 | N | Y | Y | 3 | 0.769 |
| 1,557 | 11.161 | 1.744 | Y | N | N | 1 | 0.857 |
| 4,495 | 10.171 | 0.068 | Y | N | N | 2 | 0.714 |
| 5,664 | 9.251 | 0.006 | Y | N | N | 3 | 0.666 |
| 1,557 | 6.217 | 1.473 | Y | N | Y | 1 | 0.857 |
| 4,495 | 5.345 | 0.057 | Y | N | Y | 2 | 0.8 |
| 5,664 | 4.611 | 0.007 | Y | N | Y | 3 | 0.666 |
| 1,557 | 11.161 | 2.949 | Y | Y | N | 1 | 0.857 |
| 4,495 | 10.171 | 0.163 | Y | Y | N | 2 | 0.833 |
| 5,664 | 9.251 | 0.013 | Y | Y | N | 3 | 0.8 |
| 1,557 | 6.217 | 1.99 | Y | Y | Y | 1 | 0.857 |
| 4,495 | 5.345 | 0.136 | Y | Y | Y | 2 | 0.8 |
| 5,664 | 4.611 | 0.015 | Y | Y | Y | 3 | 0.666 |

Table 4.9: [*n*-gram table] Overview of different NLP mechanisms applied to computing co-occurrence between a microblog message and a news corpus (topic:*occupysandy*). Each row in this table represents a different combination of text-matching mechanisms that were evaluated in our study.

We apply both statistical and heuristic approaches, including manual inspection on the contents with semantic relatedness in mind, to the experiment. Specifically, we manually inspect frequently used unigrams (see Table 4.11) after removing noisy information via stop word removal. Next, we classify these frequent terms into three different groups. Exploratory analysis such as frequency and burst analysis was also performed to scrutinize the data collections and compare contents from different categories with the features. We then sample microblog messages from two different groups: contents with high/low similarity with regard to mainstream news media contents. To perform this second-phase sampling task, we choose 20 and 80 percentile in n -gram feature distribution as the thresholds. We will provide some insights into the distinction that we interpreted from the experiment and discuss limitations later in Section 4.5.2.

User Study Following our content extraction and comparative analysis, we conduct a crowd-sourced user study to validate our hypothesis. In the user study, the participants were shown two groups of 10 tweet messages. Each group of tweets were randomly sampled from the messages with high similarity and low similarity to main stream news media contents in $News_{n-gram}$ metric, respectively. The participants were then asked to answer 6 different questions regarding (1) similarity to traditional news articles, (2) newsworthiness and (3) how personal the shown content is. They were also asked to answer to general questions such as demographic information (gender, age, education level, etc.) and their microblog usage.

Evaluation

We now discuss evaluation of the research questions posed earlier. Using the best performing co-occurrence method from the 24 mechanisms for computing similarity between a short Twitter message and a larger collection of news, showing in 4.9, we conducted

| Topic | # of Terms in News | | | Avg # of n -grams in a Tweet | | | Avg % of Co-occurrences | | |
|--------------------|--------------------|--------|---------|--------------------------------|--------|---------|-------------------------|--------|---------|
| | unigram | bigram | trigram | unigram | bigram | trigram | unigram | bigram | trigram |
| <i>world cup</i> | 9,274 | 75,036 | 122,573 | 18 | 17 | 16 | 77.7% | 25.6% | 6.3% |
| <i>ISIS</i> | 2,573 | 9,764 | 12,724 | 19 | 18 | 17 | 63.1% | 14.9% | 2.4% |
| <i>earthquake</i> | 2,303 | 7,114 | 8,772 | 18 | 17 | 16 | 64.3% | 15.9% | 4.1% |
| <i>occupysandy</i> | 3,078 | 11,865 | 15,190 | 18 | 17 | 16 | 60.5% | 10.3% | 1.0% |

Table 4.10: Statistics overview across different data sets (stemming only)

a user experiment to assess perceived differences between messages sampled from the niche areas shown in Figure 4.14 and a general sampling of messages in the topic. The experiment consisted of two conditions: 1) message sampling along the 20th and 80th percentiles of the x -axis from Figure 4.14 (I.e.: the co-occurrence score between a tweet and the NYT article corpus), and 2) messages sampled from the top left corner of Figure 4.14. I.e.: co-occurrence score combined with a content-based newsworthiness score for the message. This area represents messages that are inherently newsworthy but do not frequently occur in the mainstream corpus. In both conditions, the samples were shown alongside randomly sampled messages about the topic and user perception was evaluated. Information consumers can perceive newsworthiness differently over time, so we first examine a sample of temporal distributions of topics across the two domains (NYT and Twitter).

Frequency Analysis Figure 4.16 shows a frequency analysis of Twitter postings and NYT articles related to the 2014 world cup. Multiple peaks on both line plots show sudden bursts of discussions (on microblogs) or reports (from news outlets) on the corresponding topic (*world cup*). In this representative example, both streams follow a similar trend, but the bursts are more pronounced on Twitter than in traditional news. This trend in bursts is representative of several analyzed topics, so, while Twitter appears to be more reactive to events in terms of bursts, both streams show peaks of interest for critical events (semi-final and final in this case), indicating that newsworthiness of events is similar on both sources.

| | Article | | Common | | Tweet | |
|-------------|--------------|-----|------------|------|-----------------|------|
| | word | # | word | # | word | # |
| worldcup | 2014 | 412 | worldcup | 4801 | fifaworldcup | 1011 |
| | thursday | 231 | world | 2492 | bra | 763 |
| | skiing | 86 | cup | 2363 | arg | 706 |
| | longman | 76 | soccer | 1161 | ned | 551 |
| | table | 65 | brazip | 1077 | joinin | 418 |
| | association | 64 | germany | 873 | mesutozil1088 | 296 |
| | 1994 | 61 | ger | 656 | worldcup2014 | 294 |
| | golf | 60 | final | 598 | gerarg | 273 |
| | governing | 60 | team | 580 | fra | 214 |
| | christopher | 59 | argentina | 509 | crc | 211 |
| ISIS | 8217 | 33 | isis | 4872 | amp | 665 |
| | adeel | 16 | iraq | 445 | via | 497 |
| | 2015 | 13 | syria | 370 | dress | 294 |
| | fahim | 12 | obama | 340 | cnn | 170 |
| | schmitt | 11 | islamic | 339 | isil | 162 |
| | 1973 | 10 | video | 295 | share | 134 |
| | fackler | 8 | state | 281 | foxnews | 126 |
| | corrections | 6 | us | 274 | bokoharam | 119 |
| | badr | 6 | alive | 259 | usa | 113 |
| | abdurasul | 5 | jordan | 225 | daesh | 107 |
| earthquake | sniper | 31 | earthquake | 5165 | utc | 484 |
| | 2011 | 22 | magnitude | 835 | amp | 333 |
| | kyle | 19 | japan | 515 | breaking | 309 |
| | defense | 15 | tsunami | 451 | feel | 274 |
| | former | 14 | california | 348 | via | 261 |
| | marine | 12 | usgs | 345 | newearthquake | 254 |
| | tea | 10 | new | 333 | mar | 192 |
| | routh | 9 | ago | 295 | alert | 191 |
| | navy | 8 | strikes | 256 | sismo | 186 |
| | nations | 8 | quake | 245 | map | 161 |
| occupysandy | blackouts | 49 | sandy | 641 | occupysandy | 5867 |
| | andrew | 32 | help | 410 | sandyaid | 598 |
| | presidential | 30 | new | 343 | ows | 425 |
| | conn | 29 | need | 298 | sandyvolunteer | 340 |
| | newtown | 28 | hurricane | 248 | please | 329 |
| | barack | 26 | relief | 207 | occupywallstnyc | 310 |
| | education | 25 | nyc | 205 | 520clintonos | 269 |
| | connecticut | 24 | volunteers | 194 | today | 264 |
| | gasoline | 21 | occupy | 193 | info | 216 |
| | senate | 21 | rockaway | 182 | thanks | 210 |

Table 4.11: Top 10 frequent words extracted from tweets on each topic.

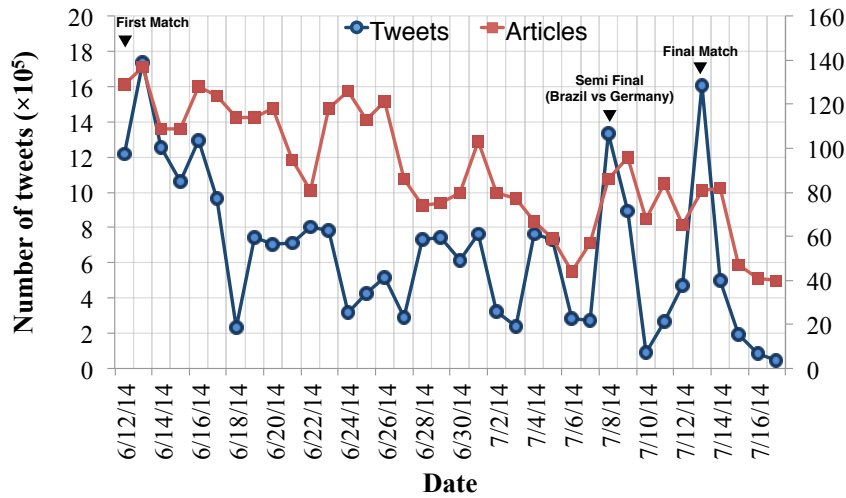


Figure 4.16: Temporal distribution of the microblog messages (tweets) and news articles on the topic *worldcup*. The time period shown in this graph corresponds to the 2014 world cup held in Brazil.

Study Participants and Procedure Participants for the user experiment were recruited through Amazon’s Mechanical Turk (MTurk). A total of 200 participants took the study which lasted an average of 8 minutes. 48% of participants were male and 52% were female. All participants were active microblog users. Age ranged between 18 and 60, with the majority between 25 and 50 (78%). 69% of participants reported having a 4-year college degree or higher. Participants were all located within the United States and had completed a minimum of 50 previous successful tasks on the MTurk platform.

Participants were shown a Qualtrics survey⁸ that asked basic demographic questions. Next, they were shown two groups of 10 microblog posts, side by side with random ordering. Two conditions were evaluated. Condition 1 showed groups of messages randomly sampled from within the 20th and 80th percentiles along the x -axis of Figure 4.14. To recap, this axis represented the co-occurrence score of the best performing mechanism from Table 4.9. Condition 2 users were shown ten messages that were sampled from the top left portion highlighted in Figure 4.14 (the ‘unique’ and ‘newsworthy’ messages),

⁸www.qualtrics.com

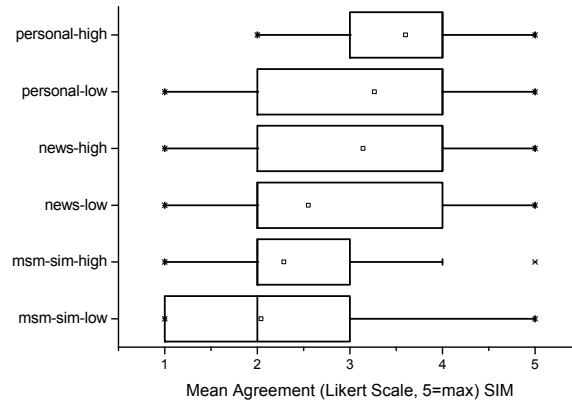


Figure 4.17: Mean agreement of the responses from condition 1 – SIM

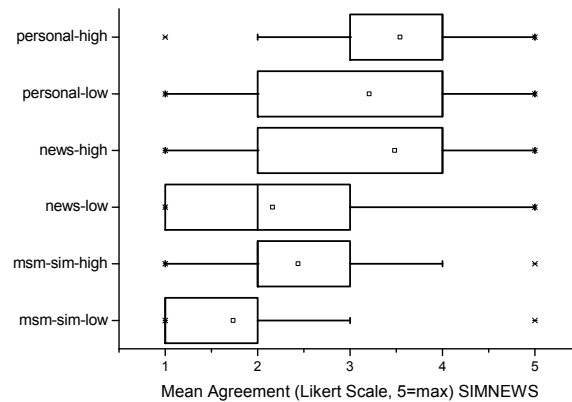


Figure 4.18: Condition 2: Mean agreement of the responses from the user study – SIMNEWS and ten randomly sampled from within the topic. This selection used both the x -axis similarity and the content-based newsworthiness score described earlier. In each case, participants were asked to rate their agreement with three statements for each group shown (total of 6 ratings):

1. *The messages in group x are similar to what I would find in mainstream news such as the New York Times.*
2. *The messages in group x are newsworthy*
3. *The messages in group x are personal*

Results Results of the experiment are shown as box plots in Figures 4.17 and 4.18. Our first task was to assess the effect of the co-occurrence metric chosen from the 24 options in Table 4.9. Two random groups of 10 tweets were sampled from the poles of this distribution (shown as the x -axis in Figure 4.14) and displayed side-by-side to participants. Participants were asked to rate their agreement with the questions listed above on a Likert scale of 1-5, with 5 indicating full agreement with the statement. Responses to the above questions are shown in Figure 4.17. Participants reported that the similarity to mainstream media was higher for messages with high co-occurrence, but, we did not observe a statistical significance for this result. Figure 4.18 however, does show a significant difference at $p < 0.05$ between the sampled messages. So, by augmenting the co-occurrence score with a content-based newsworthiness score, shown in Equation 2, we achieved a significant shift in perception of uniqueness of content. Interestingly, the perception of newsworthiness for these messages was reasonably high and did not change significantly along the x -axis (similarity to NYT), meaning that the approach did find messages that people felt were unique to the microblog domain and were also newsworthy.

Results of a term-based analysis are shown in Table 4.11 which displays three sample topics (“worldcup”, “ISIS” and “Earthquake”). The table shows the top $n=10$ terms from each data set as they overlap with the source data. The left column (Article) shows terms that are mostly unique to news articles. The center column shows combined terms, while the rightmost column shows terms that are popular on Twitter but not overlapping with the mainstream news. From manual inspection, the combined terms in the middle column in Table 4.11 appear to be a good descriptor of the topic. For example, the “ISIS” topic contains “ISIS”; “IRAQ”; “SYRIA”; “OBAMA”; “ISLAMIC” as the top 5 terms. Terms unique to mainstream media appear to be focused more on official structures and laws, while terms unique to the microblog tend to be more personal and emotional. Interestingly, the term “BOKOHARAM” is listed in the microblog column.

This is a good example of a global news phenomenon that is covered extensively in most countries, but is relatively under-reported in the United States. Now we will discuss our results in the context of the research questions presented earlier.

RQ1: How can we best detect newsworthy information in social media that is not covered by traditional media? We have examined 24 mechanisms for computing the similarity between a short microblog post and a corpus of news articles. Our findings show that a simple approach using simple unigram term matching and a porter stemming algorithm provides a better approximation of manually labeled examples than other methods tested, including POS tagging, stop-word removal and matching on bi-grams and tri-grams. Our initial expectations were that bi-gram and tri-gram overlap would produce better matches to the manual labels. Our experimental data showed that single term overlap was a better metric. We assume that since microblog posts have a limited number of terms, overlap in bi and tri-grams was sparse, as highlighted by the statistics in Table 4.9. For example, unigram co-occurrence for the topic “ISIS” shows 78% overlap with the news article database, while bi-gram overlap is 26% and trigram overlap is just 6.3%. For future work we plan to apply a combination of n -gram overlaps to create better mappings between microblog posts and news articles. *RQ2:* How do information consumers perceive the detected information? Our online evaluation of 200 paid participants shows us that sampling messages from the distributions created by the co-occurrence computation produces a significant increase in perception of the uniqueness of messages, while not affecting perception of newsworthiness. We believe that this is a promising result for the automated detection of niche and newsworthy content in social media streams.

4.5.3 Network based Approach

Following the previous approach, we propose another approach to capturing unique news content on microblogs using structural and functional metrics of network. In this section, we demonstrate our strategies to find differences in network structure and topic association between niche and traditional groups of tweets.

The main idea that penetrates our two proposed approaches is that there is a unique portion of newsworthy content in microblogs that are not covered by traditional media. The underlying assumption in the second approach is that such unique content travels from a node to its neighbors in a different fashion from those covered by traditional news outlets. Let us assume that a node u_i produces a “newsworthy” content m_i in the network and m_i becomes exposed to u_i ’s neighbors in a given time Δ_t . Unlike one-to-many propagations for contents directly provided by traditional media (e.g. tweets posted by @BBC), we expect arbitrary one-to-one or one-to-few type of propagations in the unique content group.

In this approach, we apply 1) network and 2) topic association analyses to our microblog datasets. First, we convert the crawled tweets and their associated users into two different graph data structures (network and topic spaces) based on the typical vertex/edge graph structure ($G = (V, E)$). Before analyzing the two spaces, for the network space, we reconstruct a retweet chain graph using our datasets. In this graph structure, every node, or a vertex, i represents a user u_i , and an edge $(i, j) \in E$ ($E \subset V \times V$) that connects nodes i and j becomes a retweet. We can say that $i \sim j$ if $(i, j) \in E$. For the topic space, we apply topic modeling to microblog messages using Latent Dirichlet Allocation (LDA) and extract associated topics from the messages. Using the topics extracted from the tweets, we construct a bipartite graph G_{LDA} . In this graph, we have a set of users $U = \{u_1, u_2, \dots, u_m\}$ and another set of topics $T = \{t_1, t_2, \dots, t_m\}$ that are asso-

ciated with the users $\in U$. Afterwards, we generate the final graph G_{BPR} using Bipartite Projection via Random Walks algorithm proposed by Yildirim and Coscia [186].

Hypotheses

In this study, inspired by our motivations, we aim to answer the last research question (**RQ3**) we have in Section 4.5.1.

- **RQ3:** How do the niche information get propagated differently from traditional news in the network?

As a recap, in this study, we assume that the unique and newsworthy contents on microblogs do not completely overlap with mainstream news contents. Accordingly, the following hypotheses are derived to further shape the experimental setup for our network-based approach.

Hypothesis 1 *A difference in network structure can be observed between the spread of niche (unique) and traditional media content.*

Hypothesis 2 *A difference in network dynamics can be observed between the two groups.*

Data Collection and Preprocessing

To utilize real data from the microblogging platform Twitter, microblog posts, or “tweets”, were crawled for specific keywords. In this study, we have crawled the total of 2,353,334 tweets using Twitter REST API on three different topics: *#Calais* (86,627), *#prayforparis* (1,431,467), *#paris* (835,240). After examining all datasets, we decided to focus on the *#paris* dataset which covers most news threads and relevant discussions on related subtopics. The datasets were collected during the terrorism in Paris (Nov. 8 ~ Nov. 15.) This crawling process is shown in Figure 4.19.

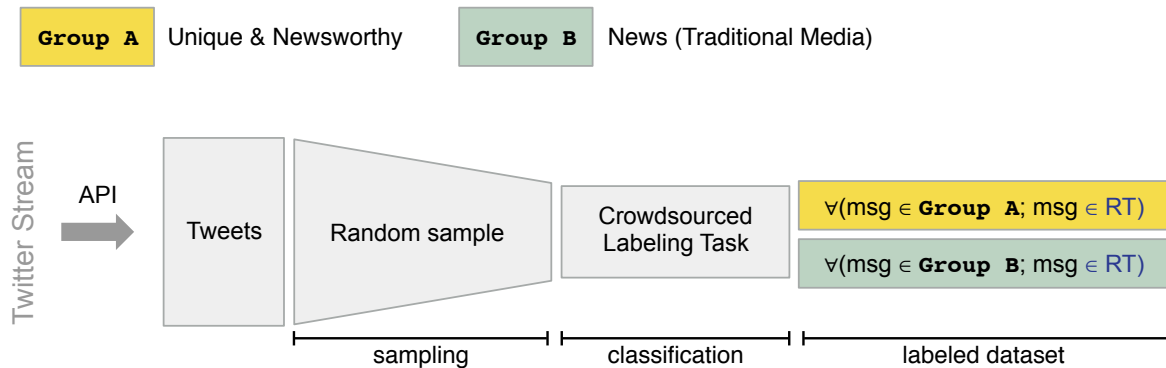


Figure 4.19: A diagram that describes crawling and labeling data sets.

Using the crawled datasets, we reconstructed retweet chain graphs in which the nodes represent users and the edges between them represent a retweet. In the data pre-processing task, the content (message text) of each tweet and corresponding metadata such as retweet count, number of friends/followers, user id and screen name, language, self-reported location are extracted using a document-oriented database⁹ and parsing scripts.

Labeling Tweets

Before the comparative analysis on the two groups of contents (Group A and B), we need to classify the messages into one of the groups. Since both newsworthiness and uniqueness of content are subjective metrics, we conducted a labeling task on a crowdsourcing platform¹⁰. Each individual message of the 300 sampled retweets from our data collection is shown to three different participants. During the task, each user was asked to rate *newsworthiness* and *uniqueness* of the given tweet in [1-10] Likert scale and answer the foundation of their judgement on newsworthiness among usefulness, timeliness, novelty (rarity) and interestingness (see Table 4.12.) We asked multiple participants to

⁹A NoSQL database (Mongo DB) was used.

¹⁰Crowdfunder (<http://crowdfunder.com>) was used for the labeling task.

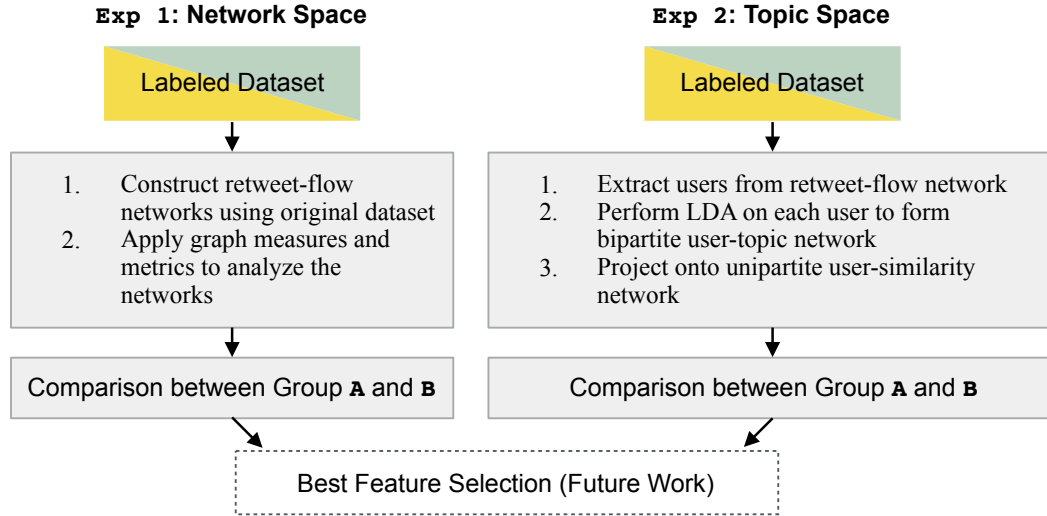


Figure 4.20: A diagram that demonstrates how we process data and evaluate the model proposed in the study.

Table 4.12: Distribution of the foundation of newsworthiness assessment in the labeling task

| News type | # Responses | News type | # Responses |
|-----------------|-------------|-------------------|-------------|
| Usefulness | 314 | Timeliness | 210 |
| Interestingness | 233 | Novelty or Rarity | 143 |

label on each message to avoid personal bias towards/against specific topic or information source. Thus, we only use the tweets that have high agreement on both newsworthiness and uniqueness of the content across three participants.

Network Analysis (Exp 1)

In this study, we are interested in investigating how “newsworthy and unique” content differs from other generic news contents. In particular, we want to analyze who generally produces this unique content and how this content is structured, i.e. propagated, in the network. Borrowing the perspectives from graph mining and social network analysis, we assume that each node corresponds to a message (or a user who posts/re-posts that message) and each edge to a propagation of a message from a node to its neighboring

Table 4.13: The metrics used for analyzing the network structures of the groups A and B

| Metric | Symbol | Description |
|-----------------------------|---------------------|---|
| Node/Edge Count | N/M | Number of nodes and edges of a graph G |
| Average degree | $\langle k \rangle$ | The mean of number of edges connected to all nodes of the graph G |
| Closeness Centrality | Cen_C | Inverse average distance to every other vertex |
| Betweenness Centrality | Cen_B | Fraction of shortest paths that pass through the vertex |
| Eigenvector Centrality | Cen_E | Importance of a node in a graph approximated by the centrality of its neighbors |
| Mean Clustering Coefficient | C | The mean clustering coefficient of the graph G |

node. The list of network metrics we use are shown in Table 4.13.

Besides the metrics we use to indicate network structures, in this study, we examine how vertices are associated with their neighbors by looking at the structure of the graphs through graph visualizations. We will discuss our findings in Section 4.5.3.

Topic Association (Exp 2)

The second experiment seeks to explore topological differences in topic-similarity networks of users that are responsible for spreading unique versus non-unique posts. The Bipartite Projection via Random Walks (BPR) method from [186] is utilized to create a user-user content similarity network for this purpose.

From the original retweet network, each user is extracted along with their 100 most recent tweets, which are aggregated into a single document. Latent Dirichlet Allocation (LDA) [187] is then performed and a document-topic matrix is produced. From this, a two-modal bipartite graph is constructed. For the *#paris* retweet network, LDA was performed with 25 topics ($K = 25$). If a user u_i 's last 100 tweets contain topic t_j , an

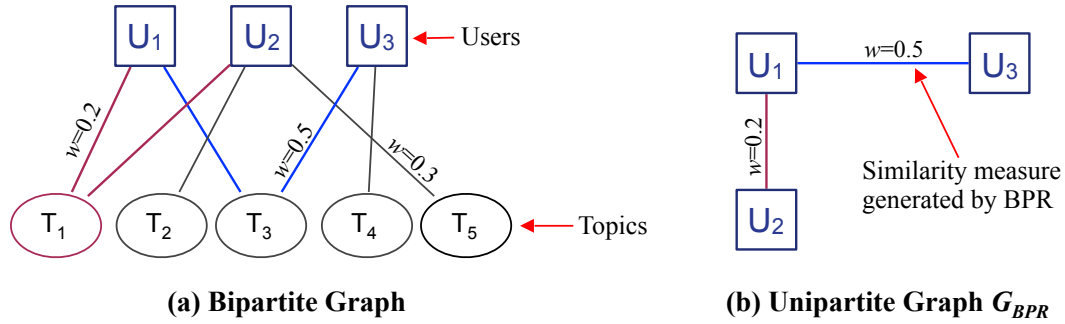


Figure 4.21: Bipartite graph construction using Bipartite Projection via Random Walks (BPR). Note that although there is an inherent weight assigned to edges in the (a) by the document-topic matrix, (b) is constructed using a simple binary adjacency matrix.

edge is drawn between i and j . Figure 4.22 shows that this network is connected and edges exist only between $(u_i \sim t_j)$ pairs; requirements for utilization of the BPR method can be found in [186].

Thresholding To construct a unipartite graph G_{BPR} , described in Figure 4.21, we set the threshold τ , not establishing every edges when two users share at least one topic regardless of the weights between the users. This strategy is considered for the ease of understanding the topology of the graph and scalability of computation. For a given bipartite graph \mathcal{G} , let $\theta_{\mathcal{G}} \in [0, 1]$ denote the threshold of weight between the user u_i and the topic t_j such that

$$(u_i, t_j) \begin{cases} \text{exists} & \text{if } \text{weight}(u_i, t_j) \geq \theta_{\mathcal{G}} \\ \text{not exists} & \text{if } \text{weight}(u_i, t_j) < \theta_{\mathcal{G}} \end{cases} \quad (4.12)$$

The BPR [186] projection method, shown in Figure 4.21, is performed on this user-topic content similarity network, and thus predicts edges in the user-user content similarity network. This technique accounts for the overall structure of the bipartite graph, which helps ensure that topic hubs do not saturate its unipartite projection with unlikely

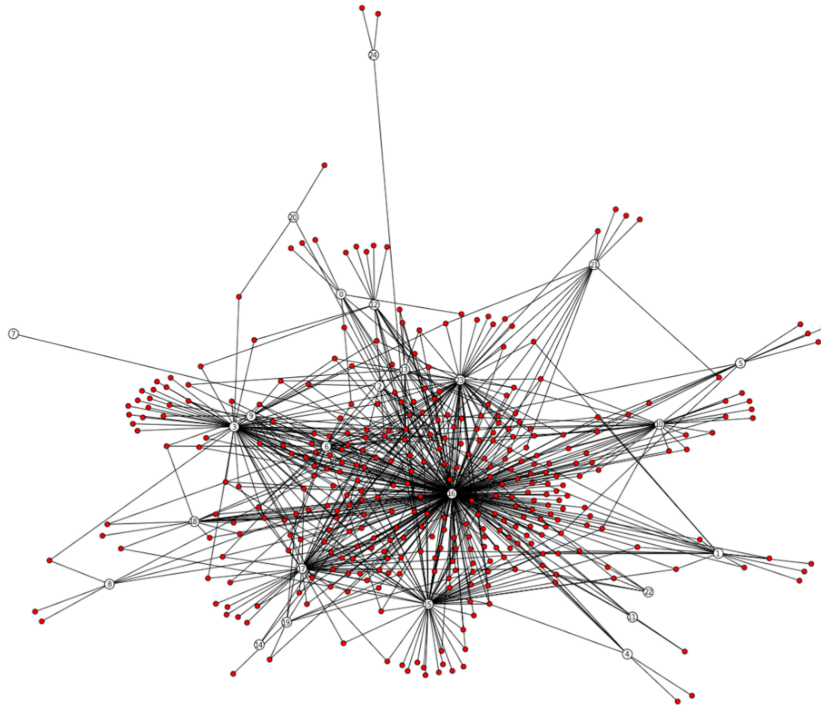


Figure 4.22: *User-topic Content Similarity Network for #paris*. Smaller nodes represent users, and white nodes represent topics.

links.

Figure 4.20 shows the overall process of data processing and evaluation of our approach.

Results and Discussions

In this section, we will discuss the findings from our two experiments (Exp 1 and Exp 2).

Network Analysis (Exp 1) Since our primary interest is how information is produced and propagated along the connections in microblogs, we study how they differ between Group A and Group B by re-constructing retweet chains from the dataset into undirected graphs and compute the graph metrics in Table 4.13 on these graphs. These metrics can

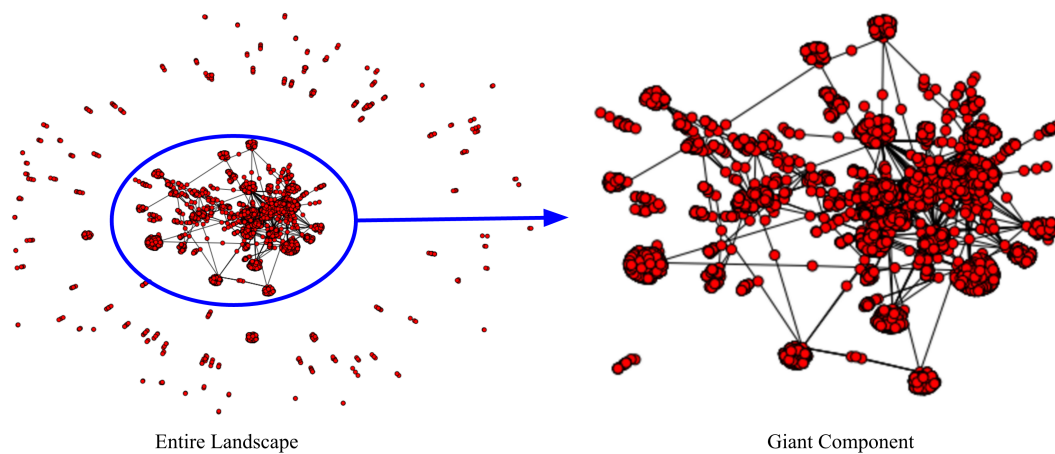


Figure 4.23: The landscape of the retweet chain graph reconstructed from the dataset “#paris.” One giant component and a number of small components were observed.

help us gain some insight into the structure, behavior, and dynamics of the given network. Specifically, for example, we can answer to such questions: 1) what are the dominating nodes in the propagation chain/network; 2) how densely do the nodes connected to each other; 3) can we partition this network into N different components; 4) does a giant component exist in this graph. To evaluate structural characteristic of the graphs in each group, we visualized the landscape of the entire data collection, and this is shown in Figure 4.23.

Figure 4.23 shows the network on the topic of *#paris* with a giant component surrounded by many isolated nodes and small components. In this graph, the giant component is loosely connected with many subcomponents via single or a few edges. Intuitively, we can divide the giant component into multiple clusters (or subcomponents) through these low-connectivity edges with high betweenness centrality. Intuitively, this type of structure can be sparsified into a simplified graph structure using sparsifier graph H (d -regular Ramanujan graph). According to Benczur-karger approximation model [188], we can sample low-connectivity edges (with high probability), eliminating high-connectivity edges within densely connected components.

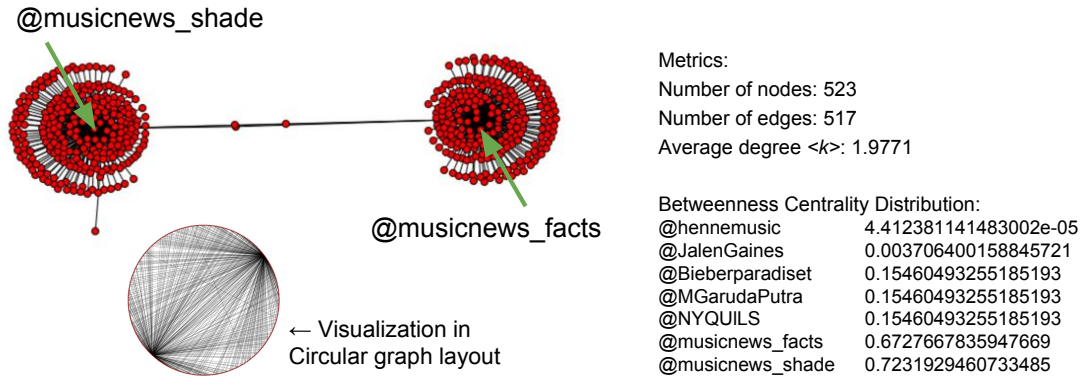


Figure 4.24: An example of dumbbell type graph found in *#paris* dataset.

For the comparison of **Group A** and **Group B**, we sampled 2 most representative sub-graphs for each group from the dataset. Structural characteristics of each set were then analyzed through visual and computational assessments.

Group A *Unique and Newsworthy Contents*: Our labeling task performed on the crowdsourcing platform revealed that the participants favored unique 3rd-party news providers or quotes from celebrity accounts (e.g. *@musicnews*, *@BrianHonan*) and labeled them as niche contents. For example, the tweet “*RT @BrianHonan: With the news breaking from Paris it’s wise to remember this. <https://t.co/bKZP5Vh46n>” was rated as highly newsworthy and unique (in other words, less likely to be seen in or covered by traditional news outlets.) Interestingly, many tweets that contain both personal opinion with sentiment and a short news headline (sometimes with a url that directs users to an external source of information) within a tweet received high newsworthy and uniqueness score.*

Group B *Traditional News*: Most tweets that fall into this category are, expectedly, news headlines or blurbs provided by major news providers or other institutional accounts. Most of the graphs in **Group B** has long retweet chain that either spans across the comparatively big component or connects two neighboring components. In some cases (an example is shown in Figure 4.24) one or two nodes exist(s) that bridges two

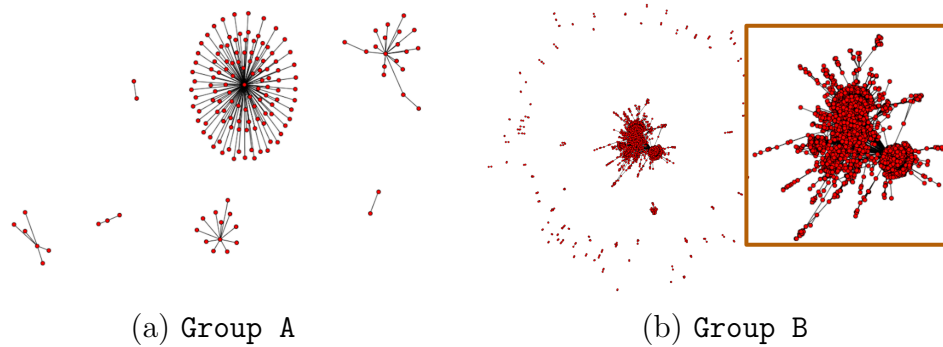


Figure 4.25: Graph visualization of the unique news group **Group A** and traditional news group **Group B**

| Group | # Nodes | # Edges | $\langle k \rangle$ | C |
|------------------------------|---------|---------|---------------------|------------|
| @globalnews (Grp A) | 159 | 151 | 1.8994 | ~ 0.0 |
| @CNN (Grp B) | 5,245 | 6,563 | 2.5026 | 0.045 |

Table 4.14: Basic metrics computed for the representative subgraphs (retweet chain graphs) of group A and B. ($\langle k \rangle$: mean degree, C : mean clustering coefficient)

small components in similar size, constructing a dumbbell-shaped graph. An example might be where the New York Times tweets about an event to its many followers, one of which is CNN News, who then retweets to its many followers. Another example of this effect that occurred in the crawled data about the Paris terrorism event involved a popular Dutch journalist who re-tweeted false information about the lights in the Eiffel Tower being turned off as a mark of respect for the victims. This created a dumbbell shaped graph between the Dutch and French communities, that also happened to contain misinformation, since the lights were actually turned off as a matter of routine.

Topic Association (Exp 2) The user-user content similarity network generated by the BPR is the network of interest. Figure 4.27 shows this network for the *#paris* example. For each projection, many possible networks can be formed based on the threshold of similarity τ between users needed to form an edge between them. Continuing the *#paris*

example, the power law is reflected in Figure 4.28, which plots the number of edges in the user-user network versus the similarity threshold used to form that specific network. This relationship seems to fit a power-law distribution, which would suggest that the BPR method has successfully captured scale-free decay in the number of similarities as the similarity threshold increases. Without any threshold, the giant component does not in fact grow to the entire network; the network remains unconnected. Notably, the unconnected nodes in the user-user network have an average degree of only 1.03 in the user-topic network, which explains why BPR did not predict any edges for these users.

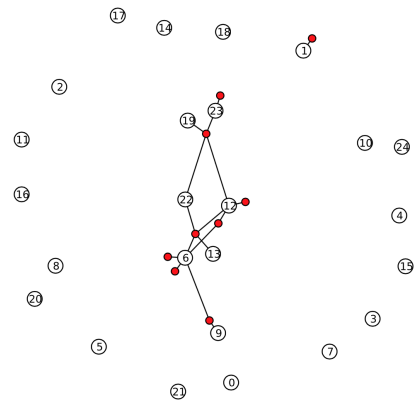
Additionally, some user-user content similarity networks that were generated for #paris are suspected of exhibiting a power-law degree distribution themselves; an example of which is shown in Figure 4.29. To corroborate this claim we will investigate further into the degree distributions of these networks as a future work.

To fully utilize the power of these user-user similarity networks in comparing unique versus non-unique content spread, the same process was carried out on a set of sampled retweet networks of the groups A and B.

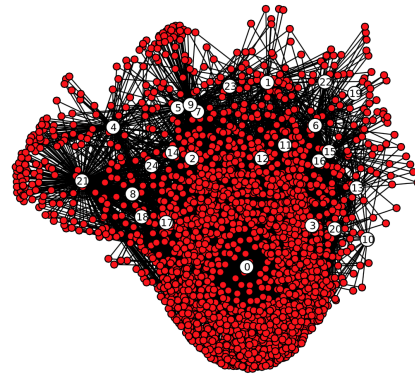
It is suspected that user-user content similarity will differ between users that spread non-unique (**Group B**) posts versus users that spread unique (**Group A**) posts, as these group's corresponding retweet-chain network structures are different. Also, comparing outlying users (users that become unconnected in user-user similarity networks) to those in the giant component of the opposite group could help provide insight to any overlap in users that spread content from both groups.

4.5.4 Future Work

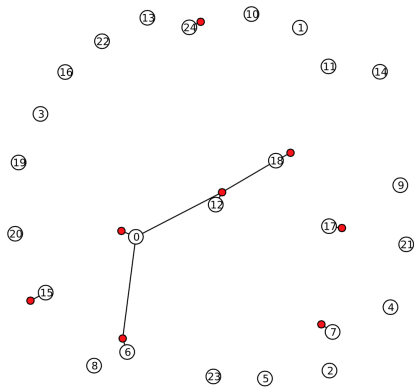
In this study, a few challenges have been discussed in order to achieve our final goal: developing a reliable and automated detection algorithm for unique news content on



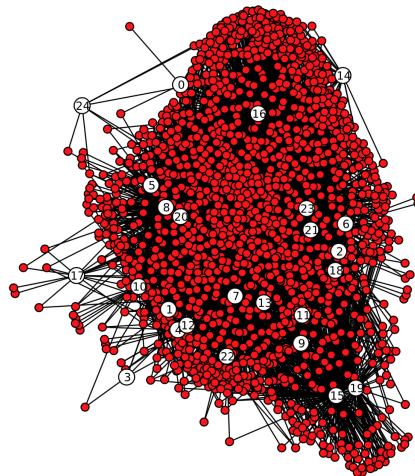
(a) Group A (@BrianHonan)



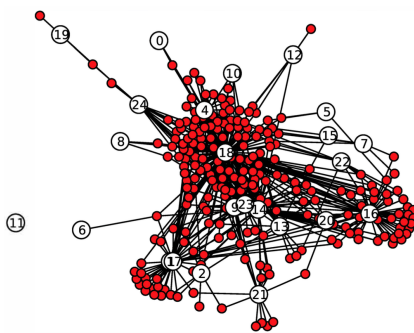
(b) Group B (@CNN)



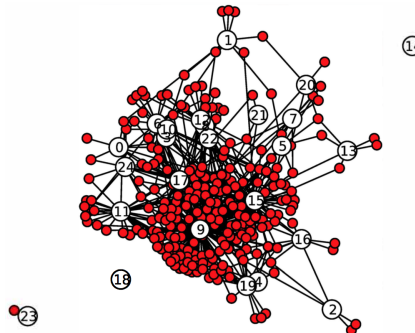
(c) Group A (@margotwallstrom)



(d) Group B (@FoxNews)



(e) Group A (@musicnews_facts)



(f) Group B (@RasmusTantholdt)

Figure 4.26: Bipartite graphs of user-topic association network. Please note that LDA topic nodes are labeled with index numbers (from 1 to K; K=25). Please note that (e) and (f) are the examples of crossover accounts.

| <i>Group</i> | <i>N</i> | <i>M</i> | <i>#CComp</i> | $\langle k \rangle$ | <i>C</i> | <i>Cen_B</i> | <i>Cen_C</i> | <i>Cen_E</i> |
|-------------------|----------|----------|---------------|---------------------|----------|------------------------|------------------------|------------------------|
| A-BrianHonan | 9 | 8 | 4 | 0.889 | 0.367 | 0.7 | 1.0 | 0.545 |
| A-musicnews_facts | 228 | 6895 | 2 | 30.241 | 0.573 | 0.033 | 0.766 | 0.189 |
| A-margotwallstrom | 8 | 1 | 7 | 0.125 | 0.0 | 0.0 | 1.0 | 0.707 |
| B-CNN | 1743 | 375 | 1647 | 0.215 | 0.029 | 0.554 | 0.521 | 0.287 |
| B-NBCNews | 226 | 7934 | 7 | 35.106 | 0.551 | 0.025 | 0.777 | 0.181 |
| B-FoxNews | 1565 | 226 | 1494 | 0.144 | 0.029 | 0.382 | 0.769 | 0.322 |

Table 4.15: Network metrics computed for the bipartite (topic-user) graphs of group A and B. Please note that all centrality metrics are computed on the max centrality nodes in the main connected component (*#CComp*: number of connected components. For other symbols, see Table 4.13).

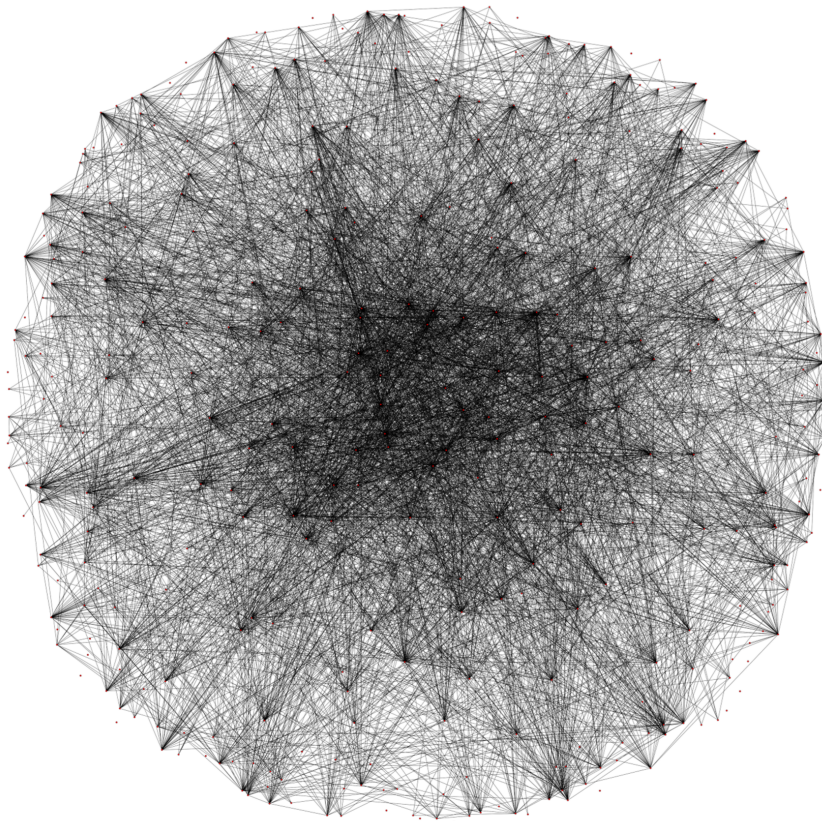


Figure 4.27: *User-user Content Similarity Network for #paris*. This specific network was constructed using a similarity-threshold of 0.00001 (4295 edges).

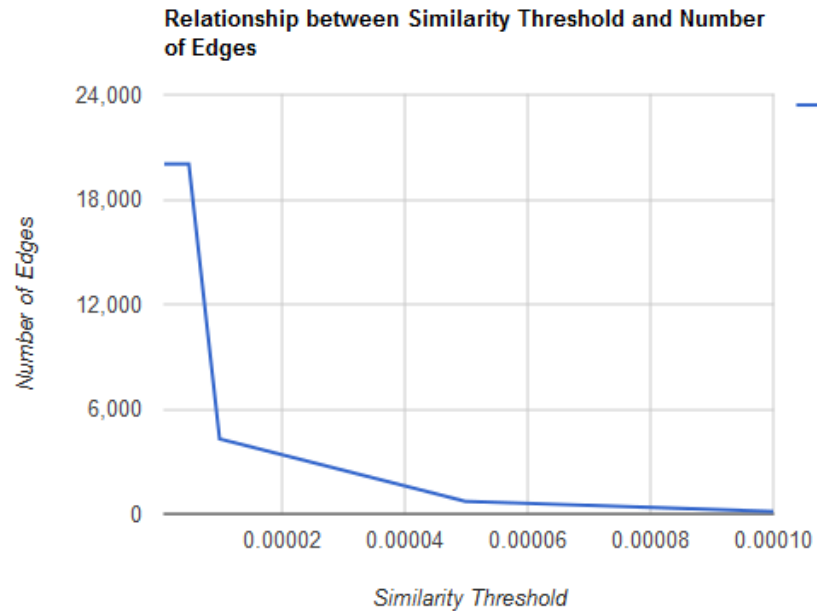
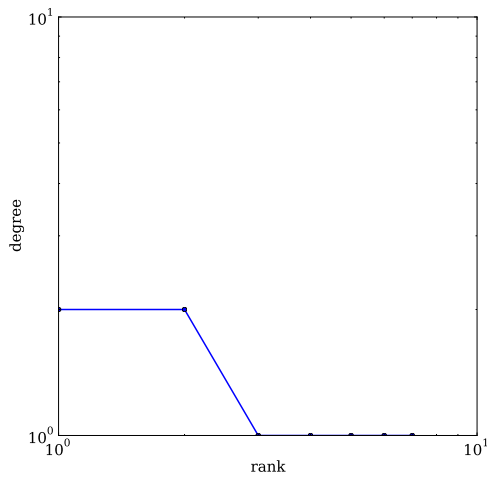
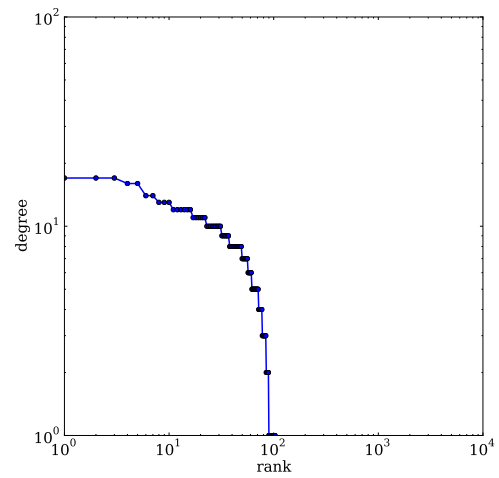


Figure 4.28: Plot of similarity threshold versus number of edges generated in user-user content similarity network using the threshold τ . Calculated power-law constants using $\tau \times 1000$: $\alpha = -1.113$, $B = 20.358$.

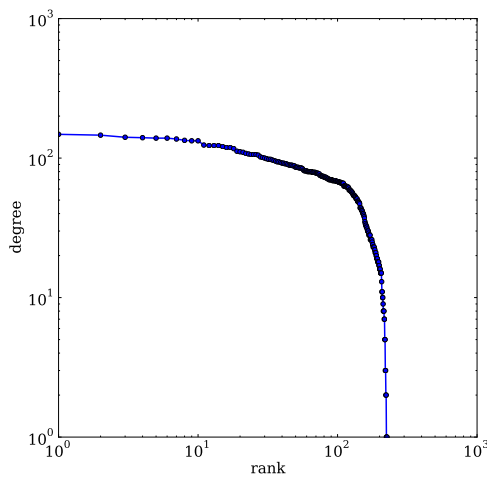
microblogs. For future work, we will apply the salient features that we found in this study to different machine learning algorithms and find an effective way to automatically locate niche microblog contents. Moreover, a temporal analysis will be performed on retweet-chain graphs in order to reveal differences in network dynamics between the groups. Any temporal patterns, found by the analysis, may allow online learning algorithms to predict niche content across time. Specifically, by investigating multiple snapshots of each network, we can measure the temporal differences and compute related metrics over the course of development of each network. In this type of analysis, tensor and different decomposition methods such as high order SVD, PARAFAC/CANDECOMP (CP) decompositions can be applied to find out multidimensional characteristics of the given network. When we incorporate the best features into an automated algorithm, however, the algorithm might need to be optimized requiring occasional user feedback



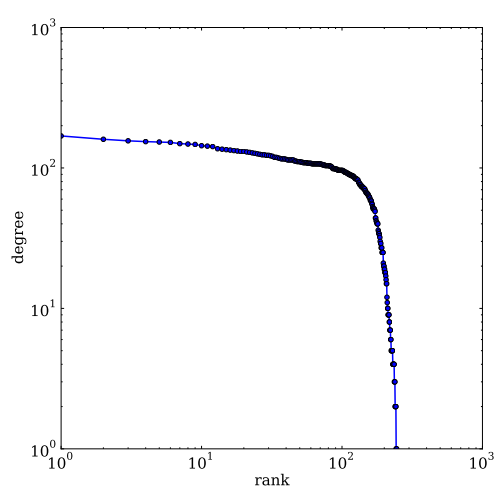
(a) Group A (@BrianHonan)



(b) Group B (@CNN)



(c) Group A (@musicnews_facts)



(d) Group B (@RasmusTanholdt)

Figure 4.29: Distributions of topic associations of users in group A and B. X-axis shows users in rank order (log scale) and Y-axis shows number of topic associations, also on a log scale.

due to the ambiguity and subjectivity of newsworthiness.

4.5.5 Conclusion

This study evaluated novel approaches for automatic detection of unique and newsworthy content in microblogs, using a comparative analysis between a corpus of curated news articles from traditional media and collections of “uncurated” microblog posts. Our initial approach examined differences in content similarity between the two. 24 combinations of simple NLP techniques were evaluated to optimize a similarity score between a short Twitter post and a corpus of news articles about a target topic. Next, a user study was described that gathered human annotations of newsworthiness for use as ground truth to evaluate our filtering method. Results showed general agreement between predicted scores from our approach and the human annotations.

We extend our news detection method to include information about the underlying network and dynamics of the information flow within it. LDA and BPR algorithms were used to explore structural and functional network metrics for the purpose of predicting newsworthiness and uniqueness of content. Primarily, we have studied the structure of various subgraphs underlying multiple topic-specific collections of microblog posts. Moreover, we have proposed a method to explore the topical association between different nodes in a graph, i.e. the vertices that tend to belong to either unique or traditional news groups. The results of our empirical analysis show that structural differences are observed between the unique and traditional news groups in microblogs. For example, the majority of subgraphs in the traditional group have long retweet chains and exhibit a giant component surrounded by a number of small components, unique contents typically propagate from a dominating node with only a few multi-hop retweet chains observed. Furthermore, results from LDA and BPR algorithms indicate that strong and dense topic

associations between users are frequently observed in the graphs of the traditional group, but not in the unique group.

4.6 Modeling User Influence for Social Marketing

In this section, we demonstrate a formative study on modeling user influence in microblogs in the context of social marketing. In this study, we design an audience manager for social marketers by extracting a content-based landscape of influential individuals in the social network. This study aims to identify influential users based on the content unlike demographic and profile-based approaches.

As the importance of Social Influence Marketing (SIM) increase, people try to focus on “how to identify influential users or entities in the Social Web” for effective and efficient social marketing. In social network, only a small portion of users has an important role with respect to information production or flows. Social marketers like to target these high influential profiles and harness their significant and immediate impact on the network. To this end, we propose a novel influence model based on the three distinctive roles of influential users on the microblog space. The model computes user influence based on each user’s topical relevance to the marketer’s interest and the potential of information propagation on the network. Our goal is to present the described solution in an interactive demo application.

4.6.1 Approach

Our approach is based on modeling microblog user behavior. More specifically, we base our assumption from our observation on how influential users behave differ from others in terms of their usage of information, type of information they produce or interact with, dynamics (activity) of the users. The approach mainly focuses on the unique roles of influential users in the network.

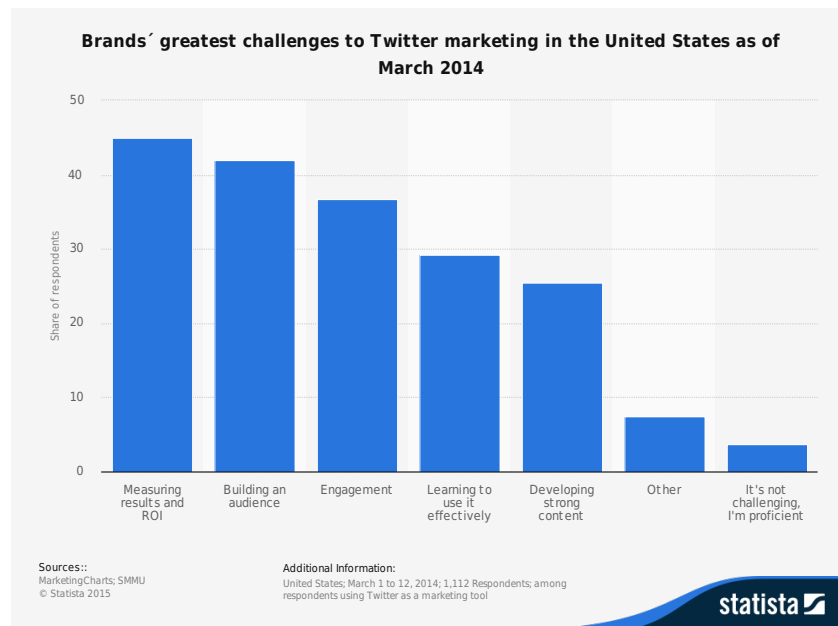


Figure 4.30: Statistics that illustrate main challenges to Twitter marketing in the U.S. as of March 2014. (Excerpt from statista.com)

Role-based User Identification

From our observation on user behaviors and information flow in Twitter we found that there are three typical roles that represent the influentials. Three-dimensional, non-exclusive metrics are used to model and quantify the influence of individual users.

To the best of our knowledge, this is the first approach that provides an influence model that maps each user in the social network to a multi-dimensional space. The model computes multi-faceted influence score of a user. Using this metric, the system returns a ranked list of influential users based on the given information. In this study, we use a corporate account (e.g. @AdobeSocial) for the initial value on which the system sets off its computation. Furthermore, since the proposed algorithm is computationally tractable, it can be integrated in a real-time monitoring system and also hugely benefit from distributed computing technologies.

Since social media, particularly microblog services, have been one of the major chan-

nels for both industrial and self marketing places in public, influential user identification problem has been sought by many researchers (Please see related work in Section 2.5.3). However, many works in the literature employ simulation-based approaches such as expectation maximization in propagation. Specifically, our approach can be differentiated from prior works in the following aspects.

Multi-faceted influence measurement For our user influence model, we propose the User Behavior Disposition score that quantifies multi-faceted user behavior in the social network. This composite score can be adapted to and harnessed for different marketing scenarios. For example, if a marketer wants to promote a newly launched service and attempts to identify early adoptors of the service in the network, she/he can selectively target those highly ranked users in the information provider group. This is possible since the User Behavior Disposition score can provide the ranking of users with regard to a specific type of influence (e.g. list of prolific users). On the other hand, for a brand marketing, information qualifier group would be a better choice.

Scalability Since the algorithm is designed to avoid expensive computation, it is scalable enough to handle real-time events using MapReduce and available distributed computing resources. We implemented the algorithm with Apache Spark to achieve near real-time performance for the prototype of the system.

Beneficial for both marketers and individual users Currently available systems that recommend influential social network users and other works from the prior arts take the end users' perspective only. For example, Klout¹¹ provides their influence score in the form of social recommendation service, recognizing individual users as their main user

¹¹<https://klout.com/home>

pool. However, our approach can be utilized for both enterprise marketers and individual users interested in self-marketing.

4.6.2 Model

In the recent years, there have been growing interest in social networks regarding the impact of the platforms from various aspects such as credibility, influence, information propagation and event detection to name a few. For example, Cha et al. [128] revealed several interesting observations from their study on measuring user influence in Twitter.

- Popular users who have a high indegree are not necessarily influential in terms of spawning retweets or mentions.
- Most influential users can hold significant influence over a variety of topics.
- Influence is not gained spontaneously or accidentally, but through concerted effort such as focusing tweets to a single topic.

To recap, in this study, we define a three dimensional representation of user influence in social networks. To model influence of users, we first dwell on the topology of information flow by deriving it from our observation on user behaviors and dynamics of information. Second, we assorted users in the network into three most typical groups: information provider; information disseminator; and information qualifier. This classification was made based on their role in information handling.

As can be seen in Figure 4.31, each role model is defined with respect to the type of impact that the user can trigger on the network. We define the influential roles in this project as follows:

- *Information Provider (P)* : A node in the network that mostly provides fresh information to its followers. The contribution that this type of node provide to the

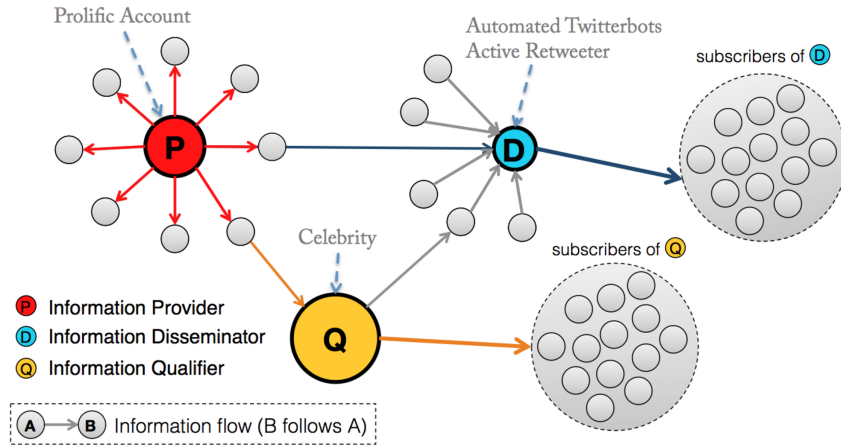


Figure 4.31: A conceptual diagram of multi-dimensional aspect of user influence in the social network.

community can be measured with the amount of information on a specific topic that the community is interested in. Most of the prolific accounts with community membership, for example, fall into this category.

- *Information Disseminator (D)* : A node that acts as an information conduit. This type of node rather curates significant amounts of information from different nodes (information providers or other disseminators) than create its own content (retweet to non-retweet ratio $num(retweet)/num(tweet)$ – is a good indicator). In general, this type of node delivers curated information to its followers by retweeting them. This user type can be characterized by (a) following many accounts that produce contents on similar topics and (b) having many followers (subscribers, $num(followers)$). Bot accounts, for example, fall into this category.
- *Information Qualifier (Q)* : A node that is verified by the service (Twitter). Celebrities or accounts that represent institutions/organizations/communities can be considered as this type of node. Due to the fact that they are verified by the general public out of social network, they already have strong bond with their followers/fans. In addition to the aforementioned loyalty from their audience, they also

possess high credibility which comes from their social reputation. It is notable that whether they speak about something in either positive or negative way is extremely important when it comes to a social marketing context.

We first winnow out the three influential role types among the users by computing the *User Behavior Disposition score (D)*.

$$D = \text{popularity} + \text{prolificity} + \text{throughput} \quad (4.13)$$

Popularity in the network features: $\text{num}(\text{follower})$, $\text{num}(\text{follower})/\text{num}(\text{following})$, listed_count .

Popularity of user u_i can be computed using following function f_p ,

$$f_p = w_p \times (\log_{10}(n_{fo} + 1) + n_l) \quad (4.14)$$

where w_p , n_{fo} , n_{fr} and n_l are popularity weight coefficient, number of followers, number of friends, and listed count, respectively.

And,

$$w_p = \frac{n_{fo} + 1}{n_{fr} + 1} \quad (4.15)$$

he popularity weight coefficient reveals how much balanced of the users behavior in the network. This metric is frequently used in order to find outliers in the network such as bots or fake accounts. We will utilize this metric to differentiate information curators or automated bots from other types of users.

Prolificity features: $n(\text{tweet})$, $n(\text{retweet})$, listed_count , $\text{listed_count}/n(\text{follower})$

Prolificity of user u_i can be computed using following function f_{pr} ,

$$f_{pr} = w_{pr} \times ((n_m - n_{RT}) + n_l \times \frac{n_l}{\log_2(2 + n_{fo})}) \quad (4.16)$$

where w_{pr} , n_m and n_{RT} are prolificity weight coefficient, number of tweets, and number of retweets, respectively.

And,

$$w_{pr} = \frac{n_m - n_{RT} + 1}{n_m + 1} \quad (4.17)$$

Pass-through rate features: $n(\text{following})$, $n(\text{follower})$, $n(\text{retweet})$, $n(\text{tweet})$, account_age

Pass-through rate of user u_i can be computed using following function f_{ptr} ,

$$f_{ptr} = w_{ptr} \times \frac{n_{RT}}{n_m} \quad (4.18)$$

where w_{ptr} and n_d are pass-through weight coefficient and account age in days.

And,

$$w_{ptr} = \frac{n_{fo}}{n_d} \quad (4.19)$$

Since we have these three metrics, we can represent the User Behavior Disposition score (D) as a point in a 3 dimensional space in a Cartesian coordinate system. Thus, D of a user u_i can be re-written as follows.

If we assume that there are n users to evaluate, we will have a 3 dim matrix ($n \times n \times n$). If we normalize each dimension (f_p , f_{pr} , f_{ptr}) we can vectorize D of u_i :

$$D_i = x\mathbf{i} + y\mathbf{j} + z\mathbf{k} = f_p\mathbf{i} + f_{pr}\mathbf{j} + f_{ptr}\mathbf{k} \quad (4.20)$$

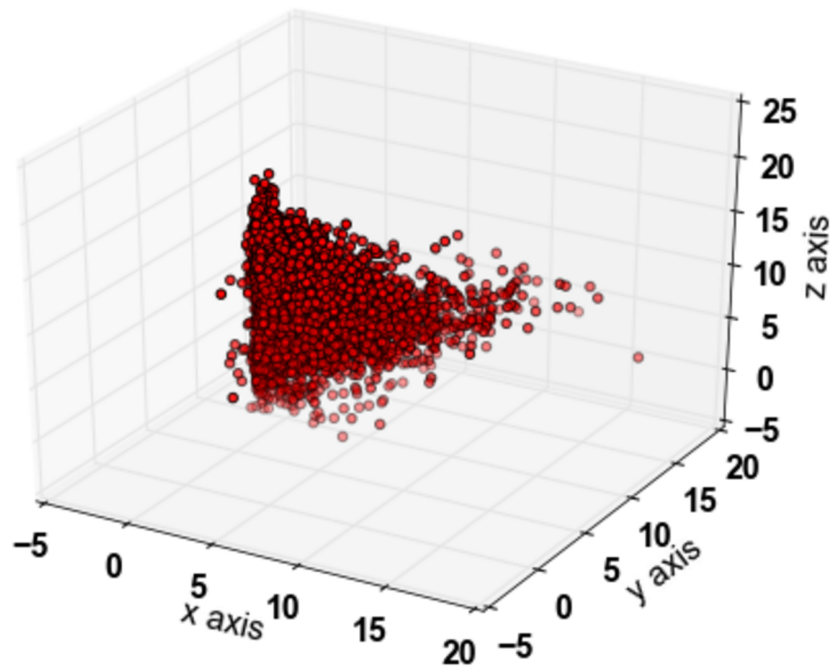


Figure 4.32: 3 dimensional plot of the vectorized User Behavior Disposition scores.

4.6.3 Data Collection

For our initial analysis of information flow and user behavior in social networks, we collected multiple datasets from Twitter. Each dataset has been crawled using one of the account names that we arbitrarily selected which represents a product, service or brand. Table 4.16 illustrates our data collection used for the data analysis and modeling. Among them, two profiles—@AdobeSocial and @Photoshop were selected for further investigation since they showed the most active profile updates at the time of our evaluation.

Random Sampling of Tweet Stream In addition to the topic-specific data collections, we crawled 233,037 number of messages along with author information without any keyword using Twitter Streaming API (September 5, 2014, 11:00am - 13:15pm). This dataset was crawled for randomly sampling a set of tweets and users in order to find reasonable thresholds for influential user detection. This approach can help us avoid

| Account | #Followers | #Followings | #Tweets |
|----------------|------------|-------------|---------|
| Adobe | 321K | 1.6K | 20.4K |
| AdobeCare | 34.6K | 8K | 52K |
| AdobeMktgCloud | 149K | 684 | 1K |
| AdobePR | 18.3K | 60 | 1K |
| AdobeSocial | 27.6K | 542 | 5K |
| creativecloud | 247K | 2K | 10.6K |
| Gap | 423K | 1515 | 19.5K |
| GapInc | 4700 | 375 | 1837 |
| Photoshop | 728K | 627 | 2.6K |

Table 4.16: Statistics overview of the collected datasets used for data analysis.

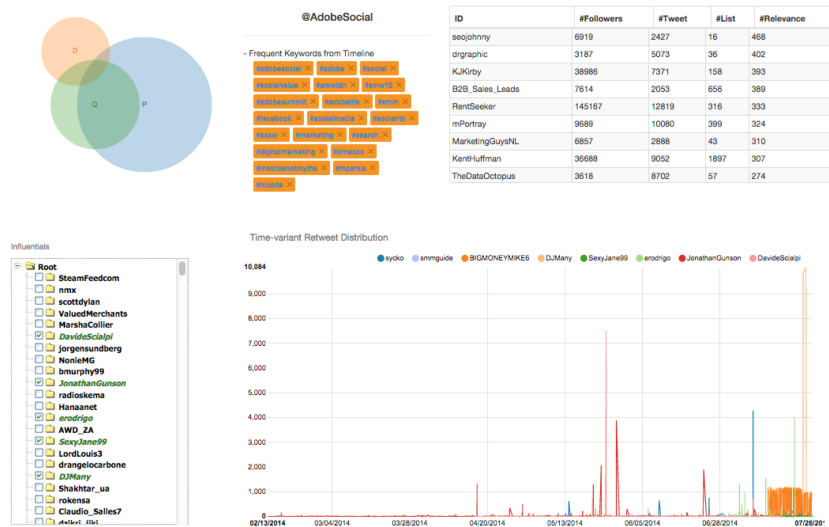


Figure 4.33: Screenshot of the implemented recommender system for influential users

possible bias of given topic/keyword.

4.6.4 Conclusion

In this study, we proposed our novel computational model that identifies influential user profiles in the social network. The proposed model is based on multi-faceted user behavior disposition score. Specifically, we carefully designed three most typical user behaviors with respect to information production, dissemination and verification in the

network through statistical distribution of user metadata and other footprints of the users sampled from our datasets. These different user types are measured by quantifying a user’s popularity, prolificity as well as the pass-through rate of the information that travels across the given account. Our evaluation through a lab-based qualitative user study showed user satisfaction and improved usability based on the participants’ self-reported responses. For our future work, we plan to conduct a quantitative, more systematic, evaluation on both the proposed algorithm and the web-based recommender system that were developed during this study.

4.7 Conclusion

In this chapter, we have discussed modeling information reliability in general, including different aspects that need to be considered and several attributes of information reliability we covered in our recent studies. In addition to that, in Section 4.1, we also provided simple guidelines for modeling reliability metrics based on the findings from the studies introduced in this chapter. We believe that these guidelines can shed some light on information modeling studies on social media. In particular, studies that build a predictive model using one or more computational algorithms (e.g. machine learning) may benefit from the implications of our recent modeling works.

Moreover, the last two studies (Section 4.5 and Section 4.6) in this chapter model rather subjective and unique metrics of information reliability: *newsworthiness* and *user influence*. Unlike traditional reliability metrics such as information credibility and user competence (expertise), which have been widely studied for many decades, modeling such metrics require stricter definitions and more rigorous evaluation. Additionally, in real-world practices, these rather subjective attributes often entail delimitation of topic or context of interest in the quality assessment. Nevertheless, over the last few years,

the importance of these information attributes, such as newsworthiness, influence and interestingness, have been exponentially increased in both the academia and the industry since they affect the type and amount of information disseminated on the Social Web. In the future studies, we will combine the proposed models with context-aware systems that harness rich user interactions.

Chapter 5

Validation

In the previous chapters, we have discussed important factors that affect 1) how humans perceive information reliability on the Social Web and 2) how to model information reliability using such factors found in our studies. In this chapter, we discuss how to construct reliable and robust ground truths using a range of available features.

Increased popularity of microblogs in recent years brings about a need for better mechanisms to extract reliable or otherwise useful information from noisy and large data. While there are a great number of studies that introduce methods to find reliable (e.g. credible or newsworthy) data, including our methods and approaches, there is no accepted reliability benchmark. As a result, it is hard to compare different studies/methods and generalize from their findings. In this chapter, we argue for a methodology for making such studies more useful to the research community with a focus on *information credibility*. First, the underlying ground truth values of credibility must be reliable. The specific constructs used to define credibility must be carefully identified. Second, the underlying network context must be quantified and documented. To illustrate these two points, we conduct a unique credibility study of two different data sets on the same topic, but with different network characteristics. We also conduct two different user surveys, and

construct two additional indicators of credibility based on retweet behavior. Through a detailed statistical study, we first show that survey based methods can be extremely noisy and results may vary greatly from survey to survey. However, by combining such methods with retweet behavior, we can incorporate two signals that are noisy but uncorrelated, resulting in ground truth measures that can be predicted with high accuracy and are stable across different data sets and survey methods. Newsworthiness of tweets can be a useful frame for specific applications, but it is not necessary for achieving reliable credibility ground truth measurements. We also show that the underlying model for predicting credibility can differ depending on the underlying network context, which needs to be clearly identified and reported in credibility studies to improve their impact.

5.1 Introduction

What are the desired properties of a credibility study? First of all, the exact definition of credibility must be made very clear by defining the underlying construct of credibility and the classes of credible and not credible messages. Methods to measure and obtain this ground truth must be justified. These methods must be robust, giving predictable results over repeated experiments. Perhaps, one of the possible purposes of information reliability studies, including credibility studies, is to find models that can predict the quality metric of interest with high accuracy and study the important features in such models. These models must also significantly improve on the baseline of random prediction especially in imbalanced prediction tasks. We will refer to a ground truth value as stable if it satisfies all these requirements. Our hypothesis in this study is that credibility models trained using stable ground truth measures are portable to multiple data sets and studies. To increase portability further, credibility studies must also identify the underlying network context that shape how credibility is communicated.

Without a stable definition of credibility, it is hard to judge to which degree a credibility model presents a novel scientific contribution.

Credibility is generally defined as the believability of information [29, 30]; please see Section 2.5.1. People judge credibility based on many different constructs such as accuracy, objectivity, timeliness and reliability, and rely on different cues like source credibility, social prominence and domain knowledge [189]. To which degree a credibility cue is used depends strongly on the decision making context [190]. Since credibility judgements are subjective, researchers must pay careful attention to the way “ground truth” credibility data is collected. The methods for obtaining ground truth may vary considerably.

User surveys for judging credibility are usually unbiased and uninformed.

Large-scale online user surveys such as those on Amazon’s Mechanical Turk or CrowdFlower offer a direct way to measure credibility. As the raters of credibility tend not to know the message senders and do not have knowledge about the topic of the message, their ratings predominantly rely on whether the message text looks believable. The results of such studies can be extremely noisy at *a single tweet level* simply because the given amount of information is too small to reliably assess credibility. Casual observations may not be as easy to classify as declarative statements. There is variation in how surveys are conducted, but the general expectation is that the survey results are unbiased except for the bias introduced by the cues presented to the raters such as the message sender’s social network or the number of retweets for the message, and the way credibility is framed in the survey. Definitions given to the user or the other questions in the survey may be used to frame which specific credibility construct should be considered when judging the credibility of the message. The surveys are also typically uninformed about the topic. As they are performed post-hoc, they do not capture how credibility would have been judged at the time of the message based on the information that was

available at that time.

In-network proxies for credibility are informed, but noisy and biased. In network behavior at the time of the message is a good proxy for credibility. For example, retweeting is often understood and used as an endorsement for the quality and interestingness [191], and credibility [20, 30, 110] of the message. Given a retweet may mean many different things, it is a noisy factor. It can also be affected by other factors such as the trust for the sender if the sender is known personally, her reputation, information cascades and corroboration in the network. Note that this type of bias may actually improve the quality of credibility judgments obtained from observed behavior. In addition, the behavior reflects how credible the message was at the time it was sent, judged by people who have a stake in a given topic. It also takes into account the level of uncertainty in the network. Some messages may also be rumors that are later found to be false, but the social media network may actively stop rumors as well [102]. Overall, behavioral proxies for credibility can be noisy, but they are also informed by the knowledge of the senders and the topic.

Studies of credibility based on analysis of factors are incomplete if they do not introduce meaningful controls. It is hard to compare and use different studies due to lack of control variables in these studies. The network characteristics and network behavior differ greatly depending on the topic, who participates in the discussion and the level of uncertainty that exists at the time. There is little work that investigates what these controls might be. As a result of all these difficulties in determining ground truth and measuring it within the proper context, there are no widely accepted or used benchmarks for credibility.

In this chapter, we provide a guideline for how to construct a stable ground truth value by carefully considering pros and cons of different ways to obtain it. To this end, we conduct a unique study. We collect two data sets on the same topic, but from different

perspectives. The first one is on *Hurricane Sandy*, collected during the storm. There is great uncertainty about what is happening and in fact there are even reported cases of misinformation being distributed [192]. The second data set is for the relief effort after the storm, from a period of lower uncertainty. Further, the network in the second data set is much more connected as it is initiated by people who have an existing social network. It is likely that people who know each other talk differently than those who talk to a general audience.

In this study, we construct different base ground truth values. We consider two different surveys in which participants are shown different information about the same tweets. This allows us to test to which degree credibility judgments across different surveys are comparable and how the survey method influences the results. We also consider two different ways to quantify retweets, overall and at the time of the message, capturing the importance of the message at two different time granularities. We show that overall survey ratings for individual tweets are hard to predict and can vary greatly from survey to survey. Prediction of retweets may vary from data set to data set. Overall, user surveys and retweet behavior are noisy indicators of credibility, but they are uncorrelated and provide different type of information.

Then, we show that it is possible to construct multiple sophisticated ground truth values by combining individual base measures. We demonstrate with examples how different definitions of credibility yield significantly different but valid sets of messages. We conduct a comprehensive study on the predictive accuracy of these ground truth values across both data sets. We show that it is possible to predict the credibility of individual tweets with accuracy values of 0.93-0.95 for different ground truth definitions, highest shown in the literature. Furthermore, the prediction improves significantly over the baseline. This finding is true for both datasets, regardless of how the survey is conducted, which allows us to conclude that our ground truth values are stable. We show

that newsworthiness of tweets can be a useful frame for specific applications, but it is not needed for constructing ground truth values that can be predicted with high accuracy. Furthermore, our combined ground truth values are not only easier to predict, but also capture the best aspects of credible information: judged *credible* by survey participants and found *interesting/relevant* within the network. We believe our method provides a step towards a more standardized approach to studying information credibility.

Our findings provide compelling evidence that reliable and meaningful credibility measurements can be constructed by combining uncorrelated and noisy measurements. We discuss how in the following sections.

5.2 Features

In this section, we describe the various features used in our study. While some of the features are novel, the rest have been proposed in prior work by us [193, 30] and others [29]. Our intention is to provide a set of features comparable with other studies of credibility. However, we remove features that are highly correlated with each other to increase the interpretability of the results. Note that our intention is not to provide a set of comprehensive features, but to give representative features that cover the frequently studied categories for user based and content based information. These include user's network, her behavior towards others, typical content of her messages, the properties of the message in question. Our features include most of the top rated features in prior work [29].

5.2.1 Content Based Features

Content based features evaluate the textual content alone, whether the text contains mentions, urls, specific type of words, sentiments expressed and so on. Note that when

judging credibility, the importance of the textual content cannot be disregarded [190]. For example, information that appears plausible is much more likely to be believed. Information that is familiar to the information consumer can be remembered quickly, and as a result may be judged more credible. These types of heuristics are often employed when judging credibility. In fact other cues such as the existence of cited sources (people or urls), whether it was retweeted or not, may be used to infer the authoritativeness of the message [189]. In fact, past work shows the importance of such features [29, 30]. However, the way authority is communicated differs based on who is speaking to whom. This well-studied notion in social sciences has not been properly studied in the credibility models. A person talking to public will use terms that are easily understood by everyone. A person talking to their social circle will use words and expression that are known within the circle. This distinction between tacit vs. implicit knowledge is shown to be very relevant in many social situations [194]. In our study involving two separate datasets, we aim to test if this is indeed true in our case. As we will show, one of the data sets comes from a much more connected network for the same topic. The list of content based features are given in Table 5.1. Details on these features can be found in [30].

5.2.2 User Based Features

User based, social features try to assess the credibility or expertise of a person by the size of their network. The number of friends gives one access to diverse information, while number of followers allows them to distribute information widely. In addition, number of followers is widely understood by researchers as an endorsement of the importance of a person in the network. There are many studies that elaborate on the importance of these features [193, 103]. Often, they signal reputation which serves as a proxy for competence or expertise [103]. The age information, i.e. number of years one has been on Twitter

Table 5.1: The set of content-based Twitter features analyzed in our evaluation.

| <i>feature name</i> | <i>description</i> |
|------------------------|---|
| <i>char/word</i> | # chars/# words |
| <i>question</i> | # question marks |
| <i>excl</i> | exclamation marks |
| <i>uppercase</i> | # uppercases in text |
| <i>pronoun</i> | # pronouns (count by corpus) |
| <i>smile</i> | # smile emoticons |
| <i>frown</i> | # frown emoticons |
| <i>url</i> | # urls |
| <i>retweet</i> | RT in tweet text, 0: not retweeted, 1: retweeted once, 2: multiple times |
| <i>sentiment_pos</i> | positive/negative word count based on lexicon sourced from NLTK ¹ |
| <i>sentiment_neg</i> | |
| <i>sentiment</i> | polarity (sentiment_pos - sentiment_neg) |
| <i>num_hashtag</i> | from entity metadata |
| <i>num_mention</i> | from entity metadata |
| <i>ellipsis</i> | counting ellipsis sign(. . .) |
| <i>news</i> | occurrence frequency of news sources |
| <i>lex_diversity</i> | proportion of unique words per tweet |
| <i>dialog_act_type</i> | category: statement, system, greet, emotion, ynquestion, whquestion, accept, bye, emphasis, continuer, reject, yanswer, nanswer, clarify, other |
| <i>news_words</i> | NLTK corpus of news article terms (sourced from Reuters), count of occurrence |
| <i>chat_words</i> | AOL messenger corpus, count of occurrence |

impacts the network of a person as older users are likely to be more central in the overall Twitter network.

In addition, it is possible to develop user features based on how the user behaves in the network and even more importantly how the user's followers behave towards the user. Often behavior from both the sender's and receiver's perspective reveals more detailed information than the simple structural information. Adalı et al. [193], show that behavior differs towards friends versus acquaintances. All our behavioral features are computed based on the statistical properties of behavior between pairs of individuals without considering message content. We compute them using only the topic based collection, thus their computation does not create an additional cost.

Propagation type behavior by followers of a user combined with high number of followers, assortativity (balance of the number of friends and followers computed by entropy) are signals for asymmetric relationships that are signals of reputation [193]. Conversation type behavior and reciprocity of messages are a signal of friendship. For each pair, we seek at least one directed message in each direction. We then aggregate features for each user across all the friends and followers to find the mean behavior for each user. Note that propagation in our features is not actual retweet behavior as we do not consider message content. It finds pairs of messages that statistically appear to be propagations [101]. Table 5.2 summarizes all the user features in this study.

5.2.3 Conversation Based Features

For conversation based features, we look for a sequence of directed messages that appear close enough in time compared to the rest to be considered a unit. For propagation-like behavior, we consider the timing of the messages from a user A to a user B , and use a linear time maximum matching algorithm developed in [195] between B 's incoming and

Table 5.2: The set of user-based Twitter features analyzed in our evaluation.

| <i>feature name</i> | <i>description</i> |
|-----------------------------|---|
| <i>u-friend/ u-follower</i> | # friends/followers (log) |
| <i>u-age</i> | # years on Twitter |
| <i>u-bal_soc</i> | ratio of follower to friends |
| <i>u-default_image</i> | user has default image or not (0/1) |
| <i>u-url/ u-mention</i> | mean # urls /mentions in tweets |
| <i>favorite-count</i> | # tweets favorited |
| <i>u-hashtag</i> | mean # hashtags in tweets |
| <i>u-length</i> | mean text length in tweets |
| <i>u-balance</i> | mean balance of number of followers |
| <i>u-conv-balance</i> | mean balance of conversations |
| <i>u-tweets/ u-favorite</i> | # of tweets / tweets favorited |
| <i>u-time</i> | mean time between tweets |
| <i>u-directed-ratio</i> | # directed tweets/#broadcast tweets |
| <i>u-retweet-ratio</i> | # retweets/#tweets |
| <i>u-prop-from</i> | # users the user propagates from |
| <i>u-prop-to</i> | # users that propagate the user |
| <i>u-convers-with</i> | # users that converse with the user |
| <i>u-propagated-tweets</i> | # tweets propagated by other users |
| <i>u-propagation-energy</i> | amount of propagation energy spent on this user by others |
| <i>u-worthiness</i> | proportion of user's tweets found worthy of propagation by others |
| <i>u-conv</i> | mean # conversations |
| <i>ss_length</i> | avg length of chain-like behavior |
| <i>ss_friends</i> | ss_length * avg number of friends (log) |
| <i>ss_followers</i> | ss_length * avg number of followers (log) |

outgoing messages satisfying a causality constraint with respect to time. A propagation-like behavior does not necessarily represent a retweet. If there are a lot of actions in which B appears to propagate from A , we can conclude that B receives a lot of messages from A and sends out a lot of messages. As a result, B is a good conduit. To further emphasize this concept, we compute chains of these behaviors and find the average length of such chains, which we will call social strength, `ss` for short. We also compute for each chain originating at node A , the average number of friends or followers along the chain multiplied by the length of the chain. We average these values and call it `ss_friends` (and similarly for followers). All `ss` features represent how well a node is as a conduit in the whole network. Details of behavioral features can be found in Adalı et al.’s work [193].

5.3 Collection and Annotation of Twitter Data

In this study, we introduce a unique comparative study of two different data sets on the same topic, Hurricane Sandy. The first dataset `FR` was collected during Hurricane Sandy using keywords “#sandy” and “#frankenstorm”, two keywords commonly used for the hurricane. The second data set `OS` was collected right after the Hurricane using keyword “#occupysandy”. Occupy Sandy is a coordinated relief effort to distribute resources and volunteers to help neighborhoods and people affected by Hurricane Sandy. It has been started by those who have participated in Occupy Wall Street demonstrations in 2012. Our choice of these two topics reflects two different perspectives about the same broader event. During the hurricane, there is a great deal of uncertainty. The topic is also of great interest to a large group of people, many of whom may not know each other. The relief effort involves a more localized group of people who are likely to know each other to some degree. In fact, we tested the connectivity hypothesis. We collected

samples of equal number of users from both datasets. We then computed the average number of connections to each other as friends in the sample. This value was 1.5 for FR and 6 for OS. Therefore, users in OS are much more connected to each other. As a result, both datasets offer us with a comparable study. They are on the same newsworthy topic. But, after controlling for topic, they represent two different contexts based on the level connectivity and uncertainty. We compare and contrast credibility measurements in these two different data sets.

We crawled both data sets using the Twitter Streaming API starting from Oct 29th, 2012 for two weeks. We applied keywords “#sandy” and “#frankenstorm” in order to generate the dataset FR during the storm and “#occupysandy” to generate dataset OS during the relief effort. The streaming API is not rate-limited, so it was possible to collect a large amount of tweets for our first data set FR. The second topic OS was far less popular (Table 5.3). From each dataset, we collected two basic samples of tweets, the first is a random set and the second is the set of tweets from users who had exchanged 2 or more messages with others in our collection. The survey tweets are a sample of 2,000 tweets each from each group with a total of 4,000 tweets. This allows us to have tweets from users with some social connectivity as well as random users.

In our samples, we excluded the users who are outliers, with more than 5K friends and 50K followers. These numbers were chosen as two standard deviations above the mean for typical Twitter users. We obtained these numbers by crawling the user info from the 2011 NIST Twitter dataset ² containing 16 million representative tweets from 2011.

²<http://trec.nist.gov/data/tweets/>

| | | | | | | |
|-----------|------------------------------------|---|---|-------------------------------|---|---|
| TWEET is: | <input type="radio"/> Can't Answer | <input type="radio"/> Strongly Non-credible | <input type="radio"/> Moderately Non-credible | <input type="radio"/> Neutral | <input type="radio"/> Moderately credible | <input type="radio"/> Strongly credible |
| | <input type="radio"/> Can't Answer | <input type="radio"/> Strongly Non-newsworthy | <input type="radio"/> Moderately Non-newsworthy | <input type="radio"/> Neutral | <input type="radio"/> Moderately newsworthy | <input type="radio"/> Strongly newsworthy |
| USER is: | <input type="radio"/> Can't Answer | <input type="radio"/> Strongly Non-credible | <input type="radio"/> Moderately Non-credible | <input type="radio"/> Neutral | <input type="radio"/> Moderately credible | <input type="radio"/> Strongly credible |

(a) 1st Survey

tweet 9: @rosefox: #OccupySandy kitchen volunteers: take this free food safety course online!
<http://t.co/UqMx6hdp> Protect yourself and the fo ...

Can't Answer

Not Credible Credible

(b) 2nd Survey

Figure 5.1: Screen shot from the two MTurk tweet assessment surveys.

5.3.1 Annotating Twitter Data

To analyze the tweets in terms of credibility we conducted two surveys. In survey 1 (Figure 5.1), we showed the users the message text, the source picture and retweet count, and sought three different types of annotations related to information credibility: the message is credible **E**, the message is newsworthy **N** and the user is credible **U**. Note that message credibility is extended with the additional source information, as a result, we will refer to this as **E**. In total 381 participants took part. Participants also had an option to select “can’t answer”. In all cases, assessments of 3 on the Likert scale and “can’t answer” responses were discarded.

The existence of images in the survey **E** may impact the evaluation of credibility as faces are often used to identify whether a source is trustworthy or not [196, 197, 198, 199]. In fact, facial evaluation is often much faster than the evaluation of text due to the dedicated processing of this signal in the brain. We expect that source credibility judgments can be impacted by this signal as there are only few other signals relating to the source. Furthermore, asking questions on newsworthiness of the tweet and the credibility of the user frame the message credibility judgment. We evaluate whether this

Table 5.3: Overview of the two topic-specific data collections mined from Twitter.

| <i>Set Name</i> | FR | OS |
|--|-----------|--------|
| <i>Seed Authors in Entire Collection</i> | 2,154,735 | 24,463 |
| <i>Seed Tweets in Entire Collection</i> | 3,801,395 | 60,671 |
| <i>Annotated Tweets in Survey E</i> | 8,728 | 6,503 |
| <i>Authors of Tweets in Survey E</i> | 7,974 | 3,239 |
| <i>Annotated Tweets in Survey T</i> | 3,471 | 3,639 |
| <i>Authors of Tweets in Survey T</i> | 2,654 | 1,657 |

frame had a noticeable impact in the next section.

To overcome possible issues related to the cues shown to the survey subjects, we conducted a second survey (Figure 5.1). This time subjects were presented with a definition of credibility, given as: “The message states a true fact and/or is believable, regardless of whether it is a newsworthy item or a personal detail.”, and shown only the textual content. In this way, we collect ground truth on both factual and personal (including opinions) content in text. We sought only a single ground truth T, whether the text is credible or not. In total 206 participants took part and at least 3 annotations were obtained for each tweet and the majority score was taken. If majority of the raters agreed on whether the message was credible or not, we used the corresponding label. Otherwise, this message was excluded. In this case, no information other than the text is available to judge credibility, but credibility is defined for the users as a construct more general than newsworthiness.

Both surveys were run using MTurk users. All assessments were given in 1-5 Likert scale. Participants were presented with instructions, followed by a pre-survey questionnaire and a set of simple filtering questions to test for bots and other noise such as rapid tab-click behavior. Each participant’s ability to rate was also tested using this set of pre-test questions. Those who did not answer the set reasonably were discarded, although this was unknown to them at the time of the study. Messages are then classified as

credible (1) or not credible (-1) based on their score.

As mentioned in the introduction, surveys are particularly useful for evaluating the text content of messages. But, there are a number of limitations. The survey subjects are unlikely to be familiar with the survey topic and are more likely to use heuristics to evaluate the credibility of the message. Furthermore, as it is very unlikely that the survey subjects are familiar with the senders of the information, source credibility information will not be based on prior information regarding the source. Hence, the use of the survey is limited in measuring the credibility of the message as a function of the expertise and reliability of the source.

To overcome this problem, we compute a secondary measure of credibility based on the fact that the message was retweeted in the network. This means that others in the network endorsed the message in some way. We compute two ground truth values, **RT** is the total number of retweets for a single tweet. All tweets in our dataset get the same **RT** (for retweet total) value as the original message that they are a retweet of. However, these retweets may have happened before or after our collection. We also computed a measure of the number of retweets of messages during the time of the collection by finding a set of tweets that are retweets of the same message either by text similarity or by their metadata. Again all the messages in a retweet group are given the same value. We call this second measure **RS** (for retweet sample). This second value represents how the message was propagating during the event we were monitoring.

We assign tweets an **RS** value of 1 if the message appears more than twice in our sample and a value 0 if the message appears only once in our sample, disregarding the rest. Similarly, we assign an **RT** value of 1 if the message has a retweet count greater than one and an **RT** value of 0 if the message has never been retweeted. A benefit of this method is that while the survey is a post-hoc analysis, propagation looks at how credible the message was at the time it was traveling in the network. Information that

| Abbr | Description | Credible | Not credible |
|-------|--|-------------------|------------------------------|
| U | user credible or not (survey 1) | value 4,5 | value 1,2 |
| N | message newsworthy or not (survey 1) | value 4,5 | value 1,2 |
| E | message credible or not (survey 1) | value 4,5 | value 1,2 |
| T | message credible or not (survey 2) | value 4,5 | value 1,2 |
| RT | message retweeted or not | 2 or more times | 0 times |
| RS | message retweeted in the sample or not | more than 2 times | 1 times |
| NE | credible among newsworthy messages | $N \& E$ | $N \& \neg E$ |
| NT | credible among newsworthy messages | $N \& T$ | $N \& \neg T$ |
| RTE | credible among retweeted messages | $RT \& E$ | $RT \& \neg E$ |
| RTT | credible among retweeted messages | $RT \& T$ | $RT \& \neg T$ |
| RSE | credible among retweeted messages | $RS \& E$ | $RS \& \neg E$ |
| RST | credible among retweeted messages | $RS \& T$ | $RS \& \neg T$ |
| ERT | credible and retweeted | $E \& RT$ | $\neg E \& \neg RT$ |
| ERS | credible and retweeted | $E \& RS$ | $\neg E \& \neg RS$ |
| TRT | credible and retweeted | $T \& RT$ | $\neg T \& \neg RT$ |
| TRS | credible and retweeted | $T \& RS$ | $\neg T \& \neg RS$ |
| NERT | credible and retweeted among newsworthy messages | $N \& E \& RT$ | $N \& \neg E \& \neg RT$ |
| NERS | credible and retweeted among newsworthy messages | $N \& E \& RS$ | $N \& \neg E \& \neg RS$ |
| NTRT | credible and retweeted among newsworthy messages | $N \& T \& RT$ | $N \& \neg T \& \neg RT$ |
| NTRS | credible and retweeted among newsworthy messages | $N \& T \& RS$ | $N \& \neg T \& \neg RS$ |
| rERT | relaxed version of ERT | $E \& RT$ | $\neg E \vee \neg RT$ |
| rERS | relaxed version of ERS | $E \& RS$ | $\neg E \vee \neg RS$ |
| rTRT | relaxed version of TRT | $T \& RT$ | $\neg T \vee \neg RT$ |
| rTRS | relaxed version of TRS | $T \& RS$ | $\neg T \vee \neg RS$ |
| rNERT | relaxed version of NERT | $N \& E \& RT$ | $N \& (\neg E \vee \neg RT)$ |
| rNERS | relaxed version of NERS | $N \& E \& RS$ | $N \& (\neg E \vee \neg RS)$ |
| rNTRT | relaxed version of NTRT | $N \& T \& RT$ | $N \& (\neg T \vee \neg RT)$ |
| rNTRS | relaxed version of NTRS | $N \& T \& RS$ | $N \& (\neg T \vee \neg RS)$ |

Table 5.4: The list of different ground truth measures used

was uncertain at the creation time may be known by the time the survey is conducted. The propagation information also incorporates how credible the source of the message was. The longer a message has traveled in the network, the more credible we consider it to be. Also, messages that are part of a long chain are likely to originate from users with higher credibility and reliability.

These measures of credibility serve as the basis of ground truth. However, we note that it is possible to augment ground truth judgments by combining complementary approaches. For example, E and RT provide different type of information about credibility. We expect both to be noisy indicators, but we also expect the noise to be uncorrelated. As a result, the combination ground truth that looks at tweets that are judged as credible and were also retweeted, is likely to be less noisy overall. Furthermore, these tweets constitute a more meaningful measure of ground truth, as credible messages that others in the network found useful and/or interesting. We will test in the next section various ways to construct ground truth and how well they can be predicted. To our knowledge, this unique approach has not been studied in any of the related work on predicting ground truth.

The list of ground truth measures tested in this study are shown in Table 5.4. We consider multiple definitions of credible and not credible messages with different meaning and different levels of restrictiveness. We consider multiple criteria for framing credibility. For example, in NT, newsworthiness is the frame. The credibility is defined only for newsworthy messages. A message is considered credible if it is newsworthy **and** credible. A message is considered not credible if it is newsworthy **and** not credible. In RTE, retweets is the frame. We consider credibility only for those messages that are retweeted. The opposite class is defined by negating all the conditions for semantic clarity. For example, in NTRT, a newsworthy message that is not credible will be not credible with respect to T and not credible with respect to RT. We also test more relaxed versions of the

Table 5.5: Correlation of the various ground truth measures for the two datasets.

| | U | N | E | T | NE | NT | RT | RS |
|----|------|------|------|------|------|------|------|------|
| U | 1.00 | 0.48 | 0.55 | 0.06 | 0.47 | 0.42 | 0.06 | 0.06 |
| N | 0.48 | 1.00 | 0.55 | 0.06 | 0.83 | 0.81 | 0.05 | 0.05 |
| E | 0.55 | 0.55 | 1.00 | 0.04 | 0.64 | 0.49 | 0.06 | 0.36 |
| T | 0.06 | 0.06 | 0.04 | 1.00 | 0.07 | 0.29 | 0.04 | 0.03 |
| NE | 0.47 | 0.83 | 0.64 | 0.07 | 1.00 | 0.84 | 0.04 | 0.04 |
| NT | 0.42 | 0.81 | 0.49 | 0.29 | 0.84 | 1.00 | 0.09 | 0.09 |
| RT | 0.06 | 0.05 | 0.06 | 0.04 | 0.04 | 0.09 | 1.00 | 0.89 |
| RS | 0.06 | 0.05 | 0.04 | 0.03 | 0.04 | 0.09 | 0.89 | 1.00 |

(a) FR

| | U | N | E | T | NE | NT | RT | RS |
|----|------|------|------|------|------|------|------|------|
| U | 1.00 | 0.42 | 0.42 | 0.03 | 0.45 | 0.37 | 0.05 | 0.05 |
| N | 0.42 | 1.00 | 0.41 | 0.04 | 0.82 | 0.74 | 0.10 | 0.11 |
| E | 0.42 | 0.41 | 1.00 | 0.01 | 0.57 | 0.34 | 0.06 | 0.07 |
| T | 0.03 | 0.04 | 0.01 | 1.00 | 0.04 | 0.28 | 0.07 | 0.08 |
| NE | 0.45 | 0.82 | 0.57 | 0.04 | 1.00 | 0.76 | 0.12 | 0.13 |
| NT | 0.37 | 0.74 | 0.34 | 0.28 | 0.76 | 1.00 | 0.15 | 0.16 |
| RT | 0.05 | 0.10 | 0.06 | 0.07 | 0.12 | 0.15 | 1.00 | 0.91 |
| RS | 0.05 | 0.11 | 0.07 | 0.08 | 0.13 | 0.16 | 0.91 | 1.00 |

(b) OS

opposite set using versions $rTRT$, $rTRS$, $rNTRT$, $rNTRS$. The relaxed version of ground truth measures are motivated by whether the two sets of raters could agree if the message was credible. The positive class corresponds to the case where both sets or raters agreed that the message was credible while the negative class corresponds to the case where either of the two sets of raters did not find the message credible.

5.4 Evaluation

5.4.1 Ground Truth Selection

In this section, we study the results from the different ground truth collection methods. As described in Section 5.3, we have conducted two different types of user surveys.

In the first, users were shown tweet text as well as the image of the author and the retweet count for the tweet. They were also asked whether the tweet was newsworthy or not. In the second, only the tweet was shown. We even removed any RT in the beginning of the text to remove cues that the message was retweeted. As a result, participants were forced to read the messages and judge them on textual content alone. However, without any information regarding the source, the survey takers lacked a frequently used anchor for credibility. In the first survey, they had a chance to base their opinions on cues like author's user image, the RT in the text and actual retweet count.

We first look at the degree to which credibility judgements are correlated to each other in the two surveys (Table 5.5). High correlation would imply a stable way to obtain ground truth information. The first thing we notice is that E is not at all correlated with T. However, NE and NT are highly correlated (0.76-0.84). Assuming newsworthiness is a stable construct for surveys, we can conclude that credibility judgments are stable within newsworthy messages. But, without a specific construct, it is hard to get a stable survey response as subjects can use many different definitions. Note that in the second survey, we did not ask for newsworthiness directly, but used the ratings from survey 1. Our method is not directly comparable to survey in [29] that asks credibility for groups of tweets not individual messages and does not ask for newsworthiness of the message. However, the observation that asking for credibility of a single or multiple messages without a framing construct may provide noisy results is applicable in the general sense to any credibility study.

We also note that in the first survey, measures U, N, E are highly correlated with each other (0.4-0.55). This shows us that the source and information credibility are judged similarly. It is also likely that the existence of questions regarding the newsworthiness of a message had an impact on the framing of the judgments of credibility. For example, it is likely that newsworthy messages were more likely to be viewed as credible,

and vice versa. However neither **E** or **T** are highly correlated with retweet based measures. Hence, we cannot conclude that showing number of retweets in **E** had a significant impact in credibility judgments. This leads us to conclude that **RT** and **RS** constitute an uncorrelated hence complementary measure of credibility on top of the user defined credibility measures. We also note that **RT** measure is informed by the knowledge of the message topic and sender that is not available to the survey subjects, and hence provides an independent type of credibility judgment.

RT and **RS** are highly correlated with each other, which means that our sample (as in **RS**) is fairly representative of the actual retweet behavior. This is also due to the fact that we only consider whether a message was retweeted or not, and disregard the actual number of retweets. We do note however that **RT** and **RS** ultimately measure a different behavior, at the time of message for **RS**, versus in the long run for **RT**.



Figure 5.2: Accuracy of prediction using different ground truth values. The values next to each ground truth value represents the Kappa value, representing how much the classifier outperforms the random guess (ranges between -1 worst, to 1 best).

| Ground Truth | Baseline | Accuracy | Kappa | ROC Area | Ground Truth | Baseline | Accuracy | Kappa | ROC Area |
|--------------|----------|--------------|-------|----------|--------------|----------|--------------|-------|----------|
| N | 59.11 | 63.96 | 0.19 | 0.64 | N | 57.52 | 58.81 | 0.09 | 0.59 |
| E | 61.36 | 64.20 | 0.14 | 0.61 | E | 60.29 | 60.20 | 0.01 | 0.56 |
| T | 71.67 | 71.45 | 0.02 | 0.60 | T | 71.92 | 71.84 | 0 | 0.58 |
| RT | 64.59 | 86.97 | 0.72 | 0.92 | RT | 64.44 | 94.20 | 0.88 | 0.97 |
| RS | 94.60 | 94.84 | 0.24 | 0.83 | RS | 78.29 | 94.89 | 0.85 | 0.98 |
| NE | 87.03 | 87.03 | 0 | 0.51 | NE | 82.08 | 82.08 | 0 | 0.53 |
| NT | 75.57 | 75.82 | 0.02 | 0.63 | NT | 74.59 | 74.29 | 0 | 0.59 |
| RTE | 65.25 | 65.43 | 0.01 | 0.53 | RTE | 64.29 | 65.15 | 0.04 | 0.55 |
| RTT | 72.43 | 72.58 | 0.05 | 0.63 | RTT | 74.72 | 72.70 | -0.03 | 0.56 |
| RSE | 55.77 | 69.23 | 0.38 | 0.76 | RSE | 66.10 | 66.83 | 0.08 | 0.60 |
| RST | 67.16 | 76.12 | 0.43 | 0.78 | RST | 77.51 | 77.95 | 0.06 | 0.63 |
| ERT | 52.87 | 87.05 | 0.74 | 0.92 | ERT | 52.72 | 93.84 | 0.85 | 0.97 |
| ERS | 92.73 | 92.98 | 0.24 | 0.84 | ERS | 67.69 | 93.25 | 0.85 | 0.97 |
| TRT | 60.69 | 93.24 | 0.86 | 0.95 | TRT | 60.18 | 94.84 | 0.89 | 0.97 |
| TRS | 87.03 | 91.64 | 0.60 | 0.94 | TRS | 55.44 | 95.26 | 0.91 | 0.98 |
| NERT | 78.93 | 93.77 | 0.80 | 0.90 | NERT | 78.07 | 96.11 | 0 | 0.53 |
| NERS | 75.00 | 84.09 | 0.56 | 0.89 | NERS | 66.77 | 95.89 | 0.91 | 0.96 |
| NTRT | 67.40 | 95.89 | 0.90 | 0.94 | NTRT | 70.0 | 95.00 | 0.88 | 0.96 |
| NTRS | 84.85 | 93.18 | 0.73 | 0.77 | NTRS | 57.19 | 95.41 | 0.91 | 0.96 |

(a) FR

(b) OS

Table 5.6: Prediction for the various ground truth measures for the two datasets.

Predictability of Ground Truth

Table 5.6 shows the accuracy achieved by our model on the task of predicting different ground truth measures using 10-fold cross validation. For these tests, we chose the best features for each ground truth using all our features. The total number of features used in this section for each ground truth measure did not exceed 10. A best feature study is presented in the next section. We also show the baseline accuracy which is measured as the prediction accuracy a classifier would achieve if it ignored all predictors and always predicted the majority class. This also shows the class imbalance in the data. We also report the Kappa statistic and the ROC Area achieved by the model. The Kappa statistic represents how much the classifier outperforms a random guess (ranges between -1 for the worst, to 1 for the best). The ROC area represents the ability of the classifier to distinguish between the two classes (ranges between 0 for the worst and 1 for the best). Prediction rates were obtained using logistic regression which was chosen because it achieved the best results overall. We excluded all features that were computed based

on the retweet counts while training our models to predict ground truth measures that include RT or RS.

In general, FR is a more noisy data set in which prediction is harder. For the task of predicting RT and RS, our model achieved prediction accuracies of 0.87 and 0.95 in FR and 0.94 and 0.95 in OS. Survey based measures (e.g. N, E, T) seem to be very noisy in comparison and prediction using our features is not very effective (not significantly better than baseline).

We then consider the problem of predicting credibility of newsworthy tweets, for which we had high correlation between NE and NT, and hence concluded that these were stable constructs. These are also not much better than baseline. Finally, we look at using retweet behavior as an anchor, and try to predict whether retweeted messages are credible or not (RTE, RTT, RSE, RST). Despite the fact that it is easy to predict whether a message is retweeted or not, it is not as easy to predict the credibility of such messages. We surmise that newsworthiness and relevance as measured by retweet rate are not good frames for obtaining reliable assessments of credibility or to make good predictions.

Our features yield significantly better predictive performance (0.92-0.95) at predicting the combined ground truth values (TRS, TRT) over both the FR and OS datasets (tweets that are credible and retweeted versus not credible and not retweeted). Similarly, ERT, ERS also show similar improvements, but the improvement over baseline is less significant in FR. We are able to build better models to predict TRT than ERT, but the difference is very small.

By combining these ground truth measures of credibility and retweets, we are essentially finding agreement between two sets of raters, from the Twitterverse and from MTurk. We had already concluded that these two were complementary credibility measures. Therefore, we are effectively reducing noise by combining two independent judgments which enables us to make better predictions. This new ground truth reflects all

Table 5.7: Prediction for the relaxation of the ground truth measures NTRT and NTRS for the two datasets.

| Dataset | Ground Truth | Baseline | Accuracy | Kappa | ROC |
|---------|--------------|----------|----------|-------|------|
| FR | rTRT | 72.48 | 78.22 | 0.42 | 0.88 |
| FR | rTRS | 95.95 | 96.13 | 0.21 | 0.87 |
| FR | rNTRT | 68.94 | 79.42 | 0.53 | 0.88 |
| FR | rNTRS | 95.74 | 94.24 | -0.02 | 0.88 |
| OS | rTRT | 71.79 | 83.28 | 0.59 | 0.90 |
| OS | rTRS | 81.56 | 89.30 | 0.64 | 0.94 |
| OS | rNTRT | 65.36 | 82.16 | 0.62 | 0.88 |
| OS | rNTRS | 75.03 | 88.38 | 0.71 | 0.92 |

the desired properties of a credible message: it appears credible and it is propagated in the network. We see similarly improved performance when we attempt to predict credibility within the context of newsworthiness. The combined ground truth measures (NTRS, NTRT) represent the same measures of credibility on newsworthy tweets where both sets of raters agree and our model achieves similarly high prediction accuracy at this task (0.93-0.96). The classifier also performs significantly better than random, has high Kappa-statistic and has high ROC area values.

It is interesting to note that the models thus built on our features perform equally well at predicting credibility over all tweets and over newsworthy tweets only. We conclude that placing credibility in the context of newsworthiness is not necessary to make good predictions of credibility if we can find the right construct of credibility and a robust ground truth measure. Ultimately, it is important to define the correct construct for judging credibility if we wish to get reliable results. Furthermore, these results appear stable across both datasets.

Measures (rTRS, rTRT) and (rNTRT, rNTRS) represent a relaxation of these measures where the negative class comprises of messages where the two raters could not agree that the message was credible. Although this relaxation means we do not have as clear a separation between the classes, it does mean that we can make predictions and train

over a larger proportion of tweets. We present the performance results in Table 5.7. We find that, despite the more noisy description, our model achieves relatively good prediction accuracy (0.78-0.89 and 0.79-0.88 respectively) at this task as well over both topic datasets. However, these measures are not too different than baseline for the more noisy data set FR.

We present examples of top and bottom most credible tweets as predicted by our classifier in Table 5.8. We trained a supervised classifier using a 66% split of a sample of the data using all of our features, and obtained classification results of whether the tweet belonged to the positive or negative class. Top and bottom tweets were chosen based on the confidence of the classifier that the instance belonged in the class predicted. As we can see, newsworthy and credible tweets resemble news items more closely. On the other hand, credible tweets for rTRS also include messages meant for exchanging information. There is a noticeable difference between top and bottom tweets. Top tweets contain more credible sources and more important information. Bottom tweets on the other hand include more conversational tweets in both cases, however bottom tweets for NTRS also include links to other sources and declarative statements more frequently when compared with rTRS. We also note a statement in the bottom tweets for TRS happens to be incorrect: the New York City Marathon was in fact cancelled in contrast with the speculation in this tweet.

Our findings and prediction results indicate that combining human annotation and retweet based judgements on credibility yields meaningful and robust ground truth measures. Many such measures can be constructed. We also find that it is possible to make reasonably high quality predictions of credibility across different datasets using features that are relatively inexpensive to compute. We note that we achieve higher accuracy in predicting credibility of individual messages than reported in [29].

| Source | Tweet Text | Source | Tweet Text |
|---------|--|---------|--|
| FR/NTRS | <p>- RT @SportsCenter: Packers safety Charles Woodson said he's donating \$100,000 to the Red Cross for assisting families hurt by Hurricane S ...</p> <p>- East Coast power outages from Hurricane Sandy reach 8.1 million: (Reuters)</p> <p>- East Coast electric companies say o... http://t.co/wVER1VsO</p> <p>- RT @cnnbrk: At least 50 U.S. deaths now linked to #Sandy – among a total of 118 worldwide. http://t.co/W3BSwLBL</p> | FR/NTRS | <p>- RT @TimTebow: My thoughts & prayers go out to everyone effected by Hurricane Sandy. Please be safe & help each other through thi ...</p> <p>Garden City resident: Sandy was a rude awakening http://t.co/wtQl3JK9</p> <p>- @RZA ~ @RedCross Donation Info Video ~ @fema ~ #Sandy ~ http://t.co/hIbnBUz ~ to Donate \$10 Text REDCROSS to 90999 _breakingDawn_XXX</p> |
| OS/NTRS | <p>- RT @OccupySandy: Today @520ClintonOS received so many donations! So many, UPS has agreed to donate a fleet of delivery trucks #mutualaid ...</p> <p>- RT @OccupySandy: Dozens of new volunteers lined up to get to work at the St. Jacobi #OccupySandy hub. http://t.co/k66ZSbF9</p> <p>- RT @OccupyWallStNYC: Where do I find info to help #SandyVolunteer? http://t.co/HJEBBhdi Keepin' it Simple and REAL. #SandyAid #SandyHelp ...</p> | OS/NTRS | <p>- @OccupyWallSt @occupysandy ASIA ALMADEH woman 45 y lost her life by #death gas Crimes of the repressive regime in #Bahrain #HRW</p> <p>- @forwardretreat @520ClintonOS @OccupySandy Thanks for the suggestion. We are currently coordinating with @520ClintonOS</p> <p>- @ALAG_Aims check out the wedding registry. It has what they need. @OccupySandy</p> |
| FR/rTRS | <p>- SANDY: Bloomberg - NYC put stickers on homes & buildings in SI & other places in NYC with different level colors to notify if they can enter</p> <p>- Long Islanders Use Facebook, Google Docs to Find Loved Ones Post-Sandy: Whether looking for a sk... http://t.co/BbUiFuUU via @mashable</p> <p>- RT @piersmorgan: I've changed my mind about this - Mayor Bloomberg should postpone the NYC marathon. Priority must be the #Sandy rescue ...</p> | FR/rTRS | <p>- The adventures of @Hanssie & her 2 friends trying to make it home after #Sandy #Getmehome http://t.co/CajGamps</p> <p>- The mayor is clearly not going to cancel it, so it is up to the runners to do the right thing. #sandy #nyc #marathon #volunteer</p> <p>- It's cold here tonight so glad for heating. I can't think how cold it is for people affected by #Sandy with no power/heating thinkin of U</p> |
| OS/rTRS | <p>- RT @rhookinitiative: #RHISupports RT @shawncarrie: #ParkSlope needs volunteers to go to flooded areas. Meet at 8th & Garfield. #San ...</p> <p>- RT @AnthonyQuintano: RT @OccupySandy: Want to volunteer doing #SandyRelief in NYC? Start by filling our volunteer form: http://t.co/6CePyV2c</p> <p>- RT @OccupySandy: NEW: Want to do some #SandyAid in New Jersey? Check out our NJ info page: http://t.co/MLRAIm58 and map: http://t.co/35j ...</p> | OS/rTRS | <p>- @opXpress @AirOccupy @JemaNdunerkant Did they actually do this to anyone? #OccuChat #occupysandy</p> <p>- @OccupySandy thank you! Just making sure. :)</p> <p>- . @hey_haywood @OccupySandy The fun thing with Clothes Mountain is as soon as it's sorted and out the door, IT REAPPEARS from new donations!</p> |

(a) Examples of top tweets

(b) Examples of bottom tweets

Table 5.8: Examples of top and bottom tweets from a sample for various datasets and ground truth measures.

5.4.2 Best features in different network contexts

In this section, we study the most predictive features for different ground truth values and compare the two different datasets corresponding to the two different network contexts. As discussed in the introduction, **FR** contains messages from a time of high uncertainty when compared to **OS**. Similarly, users in **OS** have higher percentage of social ties.

We use a heuristic based forward subset selection regression (FSS) to find a linear combination of the features that best predict the annotations in a given segment. FSS first finds the best single feature that approximates the given ground truth annotation. Then, it adds the next feature that minimizes the leave-one-out cross validation (LOO-CV) error until no improvements can be made to the LOO-CV error. This process typically produces a very sparse set of features and prevents over-fitting. We report only on those features with significance at 1% in Figure 5.3.

There are many differences between the two data sets. There are also many differences between the ground truths as expected, illustrating that these are in fact different constructs. In almost all ground truth values, **FR** models employ a more diverse set of features than **OS**. Also, the features in **FR** tend to include reputational features like long chains (short chains in **OS** in contrast) and user properties that would imply that the user is a heavy Twitter user with the use of mentions, hashtags and picking of favorite tweets. There is also difference in the content based features from **FR**, including different punctuation and use of unique words. Overall, shorter messages are more credible in **FR** and longer messages in **OS**. These distinctions could be due to different reasons. Users in **FR** come from a more varied group, as anyone interested in the hurricane was likely participating in the discussion. In **OS**, a specific group of people organizing the effort was more active. These people knew each other and hence are likely to be more similar in

the way they talk. One conclusion could be that there are a diverse number of features associated with retweeted messages in **FR** as it comes from a diverse set of users. As **OS** is from a more tight group of individuals, the messages have more normative features. One can also attribute this to the difference in the nature of the discussion: in **FR** which is on trying to assess the damage, while in **OS** in trying to organize others and give information. In **OS**, having more broadcasts than directed messages is significant, but not in **FR**. It is likely that most users in **FR** have this property and the feature is not distinctive.

If we look at similarities across all the tests, citing news sources, not having mentions, use of ellipsis for explanation or emphasis are uniformly important for predicting credibility.

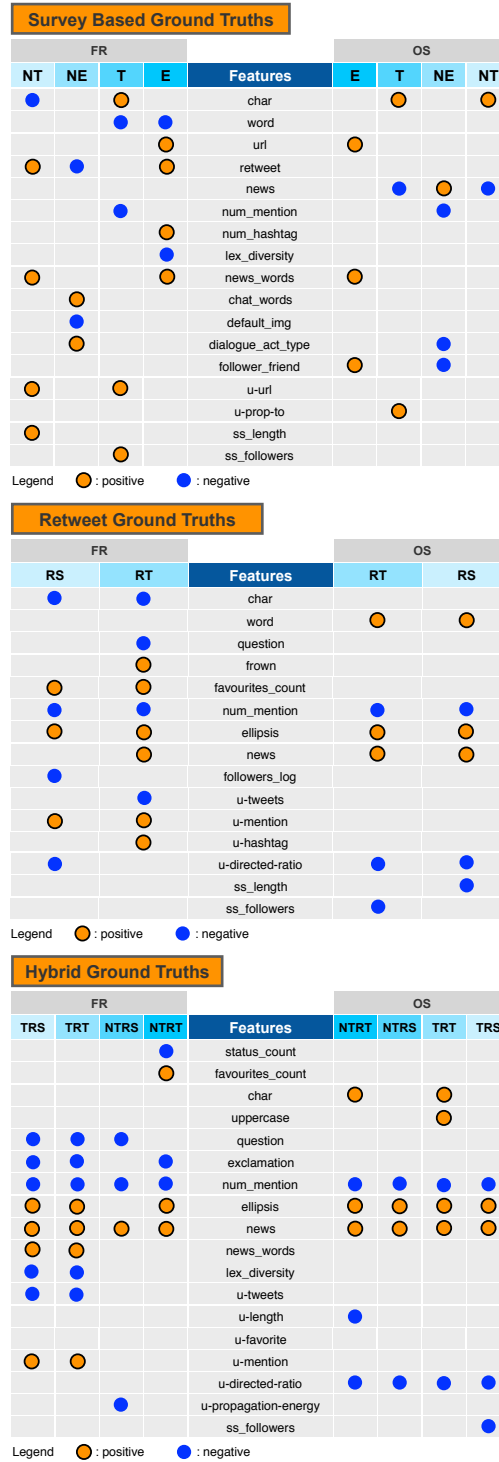
Overall, we can easily conclude that the best features for measuring credibility are highly dependent on the network context and the specific ground truth studied. This line of study also opens up many interesting new questions that can be asked by similar comparative study of different data sets and ground truth values. By seeing the differences between the best features, we can better understand when certain features are relevant and can make better informed choices on modeling credibility.

5.5 Guidelines for Studying Information Reliability

In this section, we summarize our findings and the implications of the study on information credibility into several action items that are applicable to all credibility studies in particular to feature based models in microblogs. Any study on information credibility must satisfy as many of the following suggestions as possible. We believe that this guideline can be extended to the studies on other subjective attributes of information reliability. Please see [50] for more detail.

- Define the exact notion of credibility used in the study and describe the credibility

Figure 5.3: Best features for the different ground truths in each data set.



definition as explicitly as possible.

- Use multiple credibility measures that are independent or have different biases as much as possible. Combine these measures in the ground truth construction.
 - ★ Credibility constructs that incorporate assessment of textual credibility as well as source expertise on a topic are expected to be richer and more valuable constructs.
 - ★ In network activity is a good proxy for measuring source expertise.
 - ★ User surveys are a good way to obtain textual credibility, but users may lack topical expertise.
 - ★ In network activity can also be a good proxy for measuring textual credibility, but may incorporate uncertainty arising from the lack of information at the time of message.
- Test credibility measures in multiple datasets to measure their correlation and prediction, to ensure that they are stable constructs.
- Define credible and not credible message classes carefully and unambiguously. Remove middle range of responses.
 - ★ One method to do this is by removing middle range of responses from surveys or other measurements.
 - ★ Another method to do this is by explicit construction in which the “not credible” class is the opposite of the “credible” class for every measure used in it. For example, if “credible” is defined as *A and B*, then “not credible” should be defined as *not A and not B*.

Note that this is not the logical opposite, it throws away classes that are ambiguous, i.e. *A and not B* and *not A and B*.

- If a specific frame such as newsworthiness is appropriate for the given study, then define these classes within the given frame.

Example: Given N is for newsworthy, credible with *N and A and B* and not credible with *N and not A and not B*.

- Design a survey to solicit user ratings that are appropriate for the given definition and frame of credibility.
 - ★ Users should be given the correct definition of credibility.
 - ★ Questions should be chosen carefully so that they do not influence how users perceive credibility.
- In feature based studies, do not forget that a single model is not likely to apply to all possible decision contexts. Identify the relevant contextual parameters for the study, measure and report them as much as possible.
 - ★ Possible contextual parameters include the strength and type of ties between those involved in a conversation. People will likely use different words when talking to a stranger than to a friend, and when talking to an audience versus a single person.
 - ★ Another contextual parameter is uncertainty and risk. In situations involving high risk (such as natural or other disasters) and in times of high uncertainty, people are likely to rely on the most trusted information sources and have limited resources for processing information.

We hope that future research will concentrate on standardized ground truth data sets developed by following these guidelines and be made available to the research community at large.

5.6 Summary and Discussion

This chapter described a novel method of constructing reliable and meaningful credibility ground truth values for microblogging sites like Twitter at the individual message level. We have shown that survey results can be noisy, affected by the specific framing of the questions and may differ greatly from survey to survey. Overall, it is hard to create prediction methods with high accuracy based on survey methods alone. Retweet behavior is easier to predict with network based features, but can differ from network to network. However, these two measures convey different and complementary information about credibility. We show that these two measures are uncorrelated in reality. Hence, by combining them, we are able to get ground truth values that are less noisy, can both be predicted with very high accuracy (0.93-0.95), and also capture the properties of the type of messages we would like to predict: credible text that has been endorsed as important by the network. We also show that while by framing credibility within the context of newsworthiness we do achieve high prediction accuracy, it does not necessarily result in a great increase in performance. The most important part is to choose a stable definition of credibility. In fact, we show multiple such definitions in this chapter and also illustrate some that do not work very well. We show that it is possible to measure credibility for newsworthy messages as well as for general messages with high accuracy. These findings are true in both datasets and the two different survey methods we study. We have also shown that the best features for credibility differ based on the underlying network context. As a result, feature studies must carefully consider and report on the

relevant contextual elements. Examples of such contextual elements are cultural norms for behavior, level of penetration of the social media site and the type of users. We show that the social connectivity of individuals is likely an important contextual factor. Our message based on our findings is clear: any credibility study must carefully define and measure ground truth, and quantify the relevant contextual factors.

We note that it is possible for messages to satisfy these extended definitions of ground truth and still contain misinformation. To improve further on such a metric, we can consider more sophisticated measures such as the embeddedness of different sources in the network. Investigating the effectiveness of such expensive measures is future work. We also intend to expand this study further by looking at how ground truth for other constructs such as expertise and interpersonal trust can be constructed using a combination of complementary methods. Additionally, we would like to expand our work towards understanding topic based expertise and credibility.

Chapter 6

Communicating Reliable Information

In the previous chapters, we have discussed the definitions and characteristics of reliable information in the Social Web and how we model underlying quality attributes using both qualitative and quantitative methods. To address the subjectivity of the attributes such as credibility or competence, how users perceive information differently across their personal background, tasks and social platforms in Chapter 3. In Chapter 5, we provided guidelines for reliability studies, focusing on the ground truth of information credibility.

In this chapter, we will further discuss the importance of visual interface and interaction between users and the information available to the users in the context of modern Social Web. Following the discussion on effective visual communication in general, we illustrate our contribution to communicating reliable social information through our recent works on intelligent user interfaces with real-world data.

6.1 Introduction

The goal of our study is to design and develop novel visualization of and interactive user interfaces for large and/or heterogeneous datasets for information search and discovery on the basis of reliability models and intelligent algorithms. Specifically, to this end, the following research questions have been tackled through our studies introduced in this chapter.

- How to develop scalable visualizations that allow users to easily comprehend the high dimensional structure of socially connected data?
- What is the most effective way to visualize social stream in real time?
- What are the best visualization and user interfaces that support decision-making and recommender systems?

In Chapter 3, we discussed our findings on the significance of visual perception of the presentation of information on the Social Web and the benefit of visual cognition in terms of its speed and effectiveness in information processing. Here, we begin with discussing how users make use of their visual sensory system during their information seeking and analysis tasks, and why harnessing this capability in communicating reliable information is important. Furthermore, we study the role of interactive user interfaces in such tasks through lab-based and crowdsourced user studies in Section 6.4.

We use the term *communicating* to refer to both 1) computer to human (visualization) and 2) human to computer (feedback via user interface) information flow, emphasizing the “interactivity” of user interface and visualization in information representation. The aim of our study is that we enable users rather “communicate” than simply identify reliable information through interactive and intelligent interfaces.

6.2 Visual Representation of Information

For information search, relevancy and accuracy have long been widely used as the primary measure of quality of search. Modern information filtering algorithms help users find most relevant information on the Internet. Recently, intelligent algorithms provide users with more reliable, and sometimes interesting, information by harnessing the state-of-the-art machine learning algorithms and collective intelligence. However, a number of challenges still remain in information search and knowledge discovery due to the massive volume and varying quality of available data on the Web. Given that systems or applications provide information of interest to the users, how the resulting information are represented is another problem to consider. For example, information can be represented in various forms such as plain text, text with hyperlinks, graphs or other visual elements. Effective representation can vary depending on the context and the type of information. Moreover, if the information is either large (e.g. a few gigabytes of text or thousands of images) or constructed as a set of heterogeneous data, it can be represented in a number of different ways.

In this section, we briefly discuss 1) why visual representation is important in communicating reliable information and 2) related research in the literature.

6.2.1 Visualization: Data to Information and to Knowledge

In Section 2.3, we introduced the knowledge pyramid [59, 60] which represents the Data - Information - Knowledge - Wisdom hierarchy (DIKW) [13]. Chen et al. [200] reviewed the relationship between the knowledge pyramid and visualization through applications in practice. Chen et al. provided an interesting insight into what each layer of the pyramid means in visualization process. With the assumption that “a visualization is a search process”, they interpret how the visualization process shapes information

from data and knowledge from information in the two separated spaces: Perceptual and Cognitive Space; and Computational Space; see Figure 6.1. For instance, they propose the definition of “information” in computational space in contrast to the Russell Ackoffs definition of information in perceptual and cognitive space [64] as follows:

- *information in perceptual and cognitive space (Ackoff)*: Data that are processed to be useful, providing answers to “who,” “what,” “where,” and “when” questions
- *information in computational space (Chen et al.)*: Data that represents the results of a computational process, such as statistical analysis, for assigning meanings to the data, or the transcripts of some meanings assigned by human beings

Additionally, Keim [201] asserts benefits of visual data exploration in his study as,

- Visual data exploration can easily deal with highly non-homogeneous and noisy data.
- Visual data exploration is intuitive and requires no understanding of complex mathematical or statistical algorithms or parameters.

Arguably, visual representation, including visualization, extends a user’s ability for communicating information through the humans’ visual system. By making use of this surprisingly efficient and well-designed brain module for processing information, users can improve their capability in understanding, interpreting and reasoning information. Therefore, visualization can address such challenges (information overload and varying quality) by properly mapping information into visual elements.

6.2.2 User Interfaces for Effective Communication

User interface plays another important role in communicating information along with visualization in practice. For example, in a visualization process, users can provide

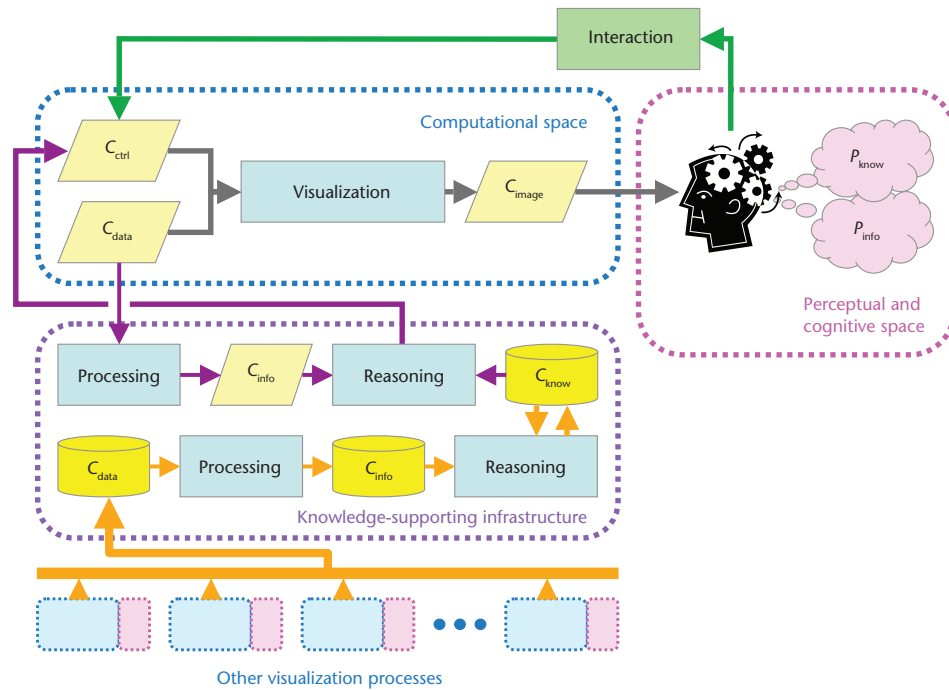


Figure 6.1: Chen et al.'s model of knowledge-assisted visualization with simulated cognitive processing. This system transforms the raw data arriving to the pipeline to knowledge through case-based reasoning (e.g. rule-based classifiers or machine learning algorithms). In this model, simulated cognition replace the role of expert users. Excerpted from [200] by Chen et al.

feedback to the system through interactive interfaces. As shown in Figure 6.1, a user can perform information search or analysis by interacting with the system through available interfaces. Furthermore, if the knowledge-based system provides a direct channel between the system and its user, the user's expert knowledge can be stored in the system as a feedback and, in turn, further refine the information represented to the user. This type of user interface is typically used for expert users such as data analysts.

6.2.3 Interactive Interfaces for Reliable Information

The deluge of networked data has been making reliable information retrieval more challenging and resource intensive. In [202], we studied the impact of visualization and interaction strategies for extracting quality information from data in complex social net-

works including microblogs and other participatory Web. The main motivation of this study is that finding optimal combinations of automated and human analyses of network data is a key challenge since data volume and reliability-determining factors vary greatly across domains, contexts and tasks. In this study, we applied two different approaches to interactive visual representations of data: an interactive node-link graph and a novel approach where content is separated into interactive lists based on data properties. To evaluate these two approaches with regard to information credibility, a reliability attribute, the TopicNets system [203] was compared with a novel system, named “Fluo” [204]. A scenario-based analysis was performed through each system on a set of big data filtered from the Twitter message service. The exposure of content, trade-offs between algorithmic power and interaction complexity, methods for content filtering, and strategies for recommending new content are assessed for each system. Fluo is found to improve on TopicNets ability to efficiently find relevant content primarily by providing a more structured content view, however, TopicNets is more customizable and boasts features which are critical for an expert analyst.

Approach To address this challenge, a scalable credibility analysis toolchain is presented that explores the limitations, potential synergies, and other theoretical boundaries between credibility analysis algorithms and credibility assessments made by human analysts. The toolchain starts with the transfer of data from a credibility analysis engine based on the Apollo system [205] and progresses to a second layer of algorithms that infuse additional modeling, such as the social and content-based credibility models described in [30]. Results are then represented in an interactive visual interface for human analysis. Both the visualization and interaction designs play a key role in an information analyst’s perception of Quality of Information in a system.

To explore the role of interaction design in depth, two distinct UI designs with dif-

ferent levels of visual and interaction complexity are analyzed in this paper. Both interaction and visualization approaches that are discussed in the context of a set of Twitter messages filtered by the the prototype toolchain. Before being visualized in either interface, messages are first annotated with information, such as credibility, generated by the Apollo system and subsequent algorithms[205].

Both systems are asked complex questions such as "What is the current difference in sentiment of tweets about #missile between the US and North Korea?" or "What are Twitter users in California saying about #Obama?" The first UI design uses the TopicNets interface from [203], which is a complex graph visualization of messages connected by topic associations. The second approach is a novel interface that organizes a graph view into several columns of ranked and truncated message lists, with a variety of filtering and sorting algorithms that are executed by interacting with data items in each column. In this paper the end goal is to assess interactive mechanisms for analysts that aid in comprehension of the data, the data model, and the underlying filtering algorithms. The longer term goal is to cognitively assess analysts' ability to provide informed feedback that improves underlying filtering models, the interface itself, and most importantly, the credibility-based filtering pipeline as a whole.

Analysis Table 6.2.3 shows a breakdown of the key elements that support visual inspection and interactive control in both interfaces. Table 6.2.3 provides a further breakdown of the advantages and disadvantages of each technique, based on a simple cognitive walk-through with expert users for the use case. Currently, a larger scale automated study is being set up to empirically evaluate both systems with data from large numbers of users. To summarize, TopicNets supports a far more diverse set of features, supporting multiple possible workflows to arrive at an answer to the task question, while Fluo has a more constrained set of functions, but in turn is far more efficient at answering spe-

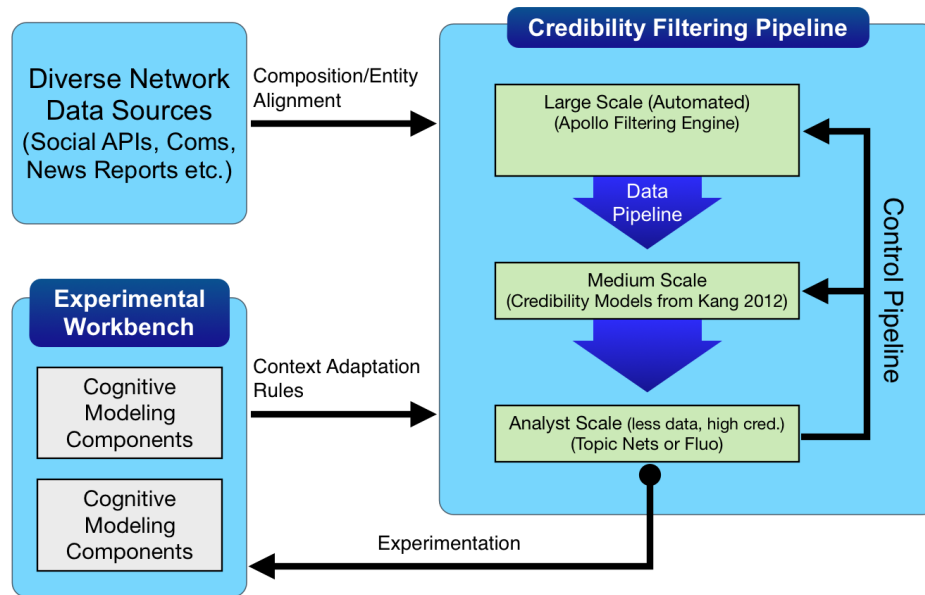


Figure 6.2: System architecture of the proposed credible information filtering pipeline and the associated experimental workbench.

cific questions and consequently is also easier for novice users to understand. TopicNets requires a longer learning-curve to be used to its full extent, making it less suitable for regular website users in an application such as Twitter, and more relevant for trained information analysts. One of the key differences between these systems is the primary modality for visualizing graph data: node-link graph with layout options (TopicNets) versus constrained column layout (Fluo). The graph view has a clear benefit for providing an overview of the entire corpus of data, which is not possible in Fluo's truncated list views. However, as edge complexity increases, the graph view becomes more cluttered and thus less effective for answering questions. Fluo overcomes this problem through ranking and truncating based on inherent data properties such as credibility, sentiment, location and other data scores provided interactively by the user during the analysis task.

As previously discussed, the data filtering toolchain's algorithms append credibility

| <i>System</i> | <i>Inspection Elements</i> | <i>Control Elements</i> |
|------------------|--|---|
| <i>TopicNets</i> | Interactive Node-Link Graph with Text and Tabular Elements | Node Dragging, Node Selection, Right Clicking, Control Panels |
| <i>Fluo</i> | Interactive List View with Text Elements | Node Selection, Slider List |

Table 6.1: Breakdown of Inspectability and Control Elements in both interfaces

and sentiment scores to each group of messages or “claim”. These values are used to guide the analyst towards potentially relevant information in both systems, however the data is utilized in different ways. Fluo presents a slider for credibility data, shown in the center column of Figure 6.3 (b). This slider affects the ranking of nodes to bias towards messages that have a certain credibility score. For example, by dragging the slider to the maximum value, the analyst is telling the system to boost the ranking of messages with higher credibility scores. This is done through a simple ranking and weighting mechanism over the results in the right column of Figure 6.3 (b). TopicNets (Figure 6.3 (a)) incorporates credibility data in a different way. Once again, a slider is presented, but this time to control a threshold value over connected entities in the node-link graph. For example, by placing the slider to the maximum value, only nodes with very high credibility will have edges drawn to topics in the document-topic graph. By default, nodes without any edges are not visualized. Both approaches have different benefits and limitations. The filtering approach in TopicNets is better at providing an overview of how the thresholding affects the entire information network and is much more effective when an analyst is searching for a particular node. Fluo’s scoring method, however, does not throw out nodes that were highly ranked by other mechanisms in the interface, increasing the quality and diversity of the results. This is especially useful when an analyst is looking for informative feedback on the algorithms that originally generated the credibility values or in the presence of noisy or erroneous metadata.

In this study the evaluation of an exploratory framework for scalable pipelining of different algorithms for filtering network data with respect to information credibility was illustrated. The framework ranges from highly scalable automated algorithms to smaller scale analyst-in-the-loop procedures that require data to be presented through interactive visualizations with controllability. As an initial experiment, data was collected through a scalable credibility filtering algorithm [205] and presented through two different user interfaces for analysis. A cognitive walkthrough of both systems is presented using the same overall task on the same data for both systems, which differ primarily in complexity of the interface and interaction capabilities. The key finding is that both systems are capable of recommending useful data by filtering based on credibility. The more feature-rich graph-based system (TopicNets) requires a greater familiarization period with the tradeoff that it can produce further perspectives on the underlying data, perhaps making it more suitable for trained information analysts than general users. Fluo does not provide many of the features of TopicNets, but improve on TopicNets ability to efficiently find relevant content by rendering a more structured content view. Besides, since content is discovered through a scoring process rather than filters, Fluo produces more diverse results, allowing it to correct for analyst or data error.

| Mechanism | Type | Advantages | Disadvantages |
|--|------------|---|--|
| Node-Link Graph (TopicNets) | Inspection | Good provenance. Easy to inspect paths, neighbor links etc. | Scales badly, gets cluttered quickly (abstraction / clustering can help). |
| List View (Fluo) | Inspection | Simple, can be reranked with provenance annotations. | Hard to display connectivity. |
| Interactive Interpolation (TopicNets/Fluo) | Inspection | Can handle lots of information. Creates a "game-like" feel to keep user interested. | Hidden functionality, usually has a learning curve, requires good annotation/help tools. |
| Tabular View (TopicNets) | Inspection | Easier to understand than a graph. | Hard to display complex connectivity/provenance. |
| Text-based (TopicNets/Fluo) | Inspection | Simple, lots of detail available. | Does not take full advantage of visual elements, does not scale well. |
| Node Dragging (TopicNets) | Control | Communicates impact of user input very well. | Not initially intuitive, difficult to re-rank vertically (crossed edges). |
| Node Selection (TopicNets/Fluo) | Control | Very useful for highlighting subset from a general overview. | Edges cause clutter quickly especially for large graphs. |
| Slider List View (Fluo) | Control | Clean look, most users familiar with slider input, can be reranked easily with provenance data shown. | Difficult to resize, less freedom. |
| Right-click (TopicNets) | Control | Useful for node-specific functionality. | Hidden functionality, has small learning curve. |
| Control Panels (TopicNets) | Control | Easier to understand than a graph, can be labeled more easily. | Can get cluttered quickly depending on the number and complexity of actions. |

Table 6.2: Advantages and disadvantages of Inspectability and Control Elements



Figure 6.3: The two visual interfaces illustrating data filtered through Apollo system for the query “Hurricane Sandy”.

6.3 Real-time Systems with Visual Interfaces

In this section, we illustrate our study on real-time visualization of social stream. In this study, we designed a novel real-time microblog stream visualization framework, named *TweetProbe*, and its contribution to identifying timely information on the Social Web. In this study, we present the design consideration of the TweetProbe and how the framework detects and highlights trending information with respect to the two important attributes of information reliability: recency (timeliness) and popularity (prominence). We evaluate the system and explain the underlying architecture focusing on two most trending topics (“#occupygezi”¹ and “#royalbaby”²) at the time of evaluation. As a motivating example, TweetProbe shows how 1) real-time social information stream can be effectively monitored and 2) interesting topics can be identified through animated visual elements. Through the proposed system, users can observe birth, growth, and death of ephemeral information (e.g. breaking news or short-lived but popular personal remarks) on microblogs through both sliding animation effect and logarithmic timeline.

6.3.1 Introduction

As user-centric social media such as Facebook and Twitter become more popular, user-generated contents serve as major information sources across various fields. For instance, recent marketing strategies give significant attention to social ‘big data’ and try to find meaningful patterns therein, in order to analyze consumer preferences or market dynamics. Moreover, information scientists have been conducting numerous research projects on social networks, applying state-of-the-art statistical models to extract topic-specific information, detect social events or extract sentiment on a specific topic.

¹201314 protests which took place in Gezi park in Istanbul, Turkey (Diren Gezi Park)

²The birth of English royal baby Prince George of Cambridge (George Alexander Louis), 20th of June, 2013

In this work, we present a real-time algorithmic visualization that shows trending topics, messages and their sentiments. *TweetProbe* (Tweet Stream Probe Framework)³, reveals live voices of microblog users and, by highlighting majority trends, we can easily sense current hot-button issues, social events and gossip. Our goal is to provide a novel efficient visualization technique for information consumers, scientists, and media arts audiences to help them easily understand and reflect real-time information from microblogging services. In this sense, the immediacy aspect and small time window used in this framework is the key component, since it enables users of this framework to detect real-time trends, local events, natural disasters and spikes of social signals at a microscopic level in a short time frame.

The objective of this research is a novel visualization design and its implementation based on real-time social media streams which provides

- identification of emerging (fastest growing) topics in real-time,
- identification of the most influential nodes in a long retweet chain,
- sentiment extraction from a topic of interest,
- event detection on a specific location,
- efficient algorithms which cope with massive amount of streaming data, and
- aesthetic visualization with intuitive visual components, suitable for media arts installations.

The main contribution of this study is the proposition and design of novel microblog visualizations which are carefully designed for real-time data streams from services such

³Sample video clip of TweetProbe is shown at <http://youtu.be/-MlPi1opnIk>

as Twitter. Our visualization framework is designed to detect instant updates in topic-specific discussions in the Twitter space and convey them to users through animated visualizations using time-window binning and sentiment extraction algorithms.

6.3.2 Design Considerations

The main goal of our visualization is to help users easily monitor trending messages, relevant topics and sentiment distribution of the given topic in real-time by supporting intuitive as well as thought-provoking visualization. Responding to user interest in staying on top of the information flow, numerous microblogging applications provide trendy topic ranking services to their users. However, it is still challenging to detect emerging topics on time, particularly if the topic is based on an emergent (new) event. For example, if a plane has crash landed a few minutes ago, it takes at least a half hour to become a trendy topic on microblogging sites and, thus, the original posts about the accident will not receive wide attention until they reach a sufficiently high number of retweet or favorites counts. By that time, network structures surrounding the author and retweeters (i.e. their followers) play a key role in this dissemination process. Since *retweet count* or *favorites count* are the key metrics for measuring popularity of messages in Twitter, these metrics can be used as important metadata in information analytics. To detect the most recent and emerging messages, we use a binning technique to find the messages predominantly shared by users in a given time window. The term “emerging topic” used in this study is considered as “the fastest-growing topic or message” in microblogging space. By considering the real-time message dissemination process, we decided to employ an animation-based design in our TweetProbe visualization framework. Our reasoning was that this design concept is most effective to convey real-time information flow in detail and reveal the overall dynamics of emerging topics in social network (from their

birth and growth to their decline and disappearance). The animated visualization reveals ranking transitions and the development of single topics across the entire network. A captured video of TweetProbe can be seen at <http://youtu.be/-MlPi1opnIk>. We will discuss the architecture of our framework in detail in Section 6.3.3. In this section, we discuss the primary principles of our design decisions.

Real-time Message Filtering

When a user applies filtering keywords to the system, they are sent as a parameter to the Twitter streaming server through the Twitter Streaming API. After this filtering phase, the system continuously receives tweet entities (a micro message and its metadata) in JSON⁴ data format. Each arrival of information through the streaming connection invokes a back-end data processing thread which in turn triggers item comparison, binning, ranking and sentiment extraction tasks. Between the comparison and ranking tasks, memory cache (bucket) is used to filter out irrelevant messages. This is a critical process in our system since it resolves scalability issues arising from the huge influx of data from the stream.

Time-window based Ranking

Trending messages in general in microblogs are ranked based on the total amount of sharing or occurrence in messages (retweet or hashtag in Twitter, respectively), which is the number of these events over a fixed period of time. However, monitoring a transient topic in real-time is still a challenging task since we need to collect a sufficient amount of messages during a reasonable time frame. In TweetProbe, we take a different approach to deal with the same problem. We assume that a burst of retweet action within a small time window can be considered a trendy topic in real-time. While there is a default rate

⁴JavaScript Object Notation (<http://json.org>)

set up (50 tweets per time window), the system enables each user to set a preferred rate as a threshold to detect trendy topics. Once a message's retweet count updates exceed the given threshold, the message is highlighted with a visual symbol for an emerging topic. The list of highly ranked messages or hashtags is being updated as new messages arrive in the system. The time window is initially set as 10 minutes, however it as well can be reconfigured by the user.

Color-coded Visualization

Each item in the timeline frame and the message frame are corresponding to each other in color. The color-coding scheme in our visualization is carefully designed to enhance readability of our system. It shows the scale in message ranking and aims to help users understand multiple facets of a single entity simultaneously.

Sentiment with Rain Drops

Aesthetic considerations are obviously crucial in the creation of artistic narratives and for provoking thought processes in audiences experiencing media arts installations, but they also have a big impact on visualization usability [206]. Interest in usability and influencing the audience's mood overlap when it comes to depicting results from sentiment analysis. Since sentiment scores express polarity in its scale (negative, neutral and positive), the sentiment of each message can be expressed as a color gradient, e.g. between red and blue. As the name implies, the 'stream of information' can be imagined as a flow in a continuous medium such as a current or stream. However, we can also think of each message as a discontinuous element in a flow of continuity. This abstract metaphor is the major motivation for our sentiment visualization which describes the message stream as a collection of rain drops. Detailed description of the raindrop visualization is available in Section 6.3.4.

Logarithmic Timeline

In microblogs like Twitter, we have a potentially huge span of timeline filled with countless message updates. However, as we mentioned in the previous sections, instant analysis of recent data stream is a crucial part of our approach. Therefore, we want to focus on the most recent messages. Perhaps, we can also look at presently active conversations among people regarding an old topic which has recently been brought up again due to some triggering event. This is our motivation for employing a logarithmic timeline in our visualization. This enables us to focus on recent messages with much higher resolution on the timeline and also show some old topics in approximate position on the same graph. Please note that we only show the original posting time of each retweet on this timeline.

A very prominent early logarithmic timeline visualization was designed by Sparks [207] about 80 years ago. Basically, logarithmic scale in timeline visualizations enables the depiction of historical events throughout time while focusing more on events closer to one end. Sparks explained this as follows: *As we travel forward in geological time the more complex is the evolution of life forms and the more are the changes to be recorded. Further, the most recent periods of evolution hold the most interest for us. We need therefore increasingly more space for our outline the nearer we approach modern times, and the logarithmic scale fulfills just this condition without any break in the continuity.* Both the old depiction from [207] and our timeline are shown in Figure 6.4

6.3.3 System Architecture

In this section, we present the overall system architecture of TweetProbe and discuss each visual component in detail. TweetProbe is comprised of two main components: A back-end data processing layer and a front-end visualization layer. Each layer is

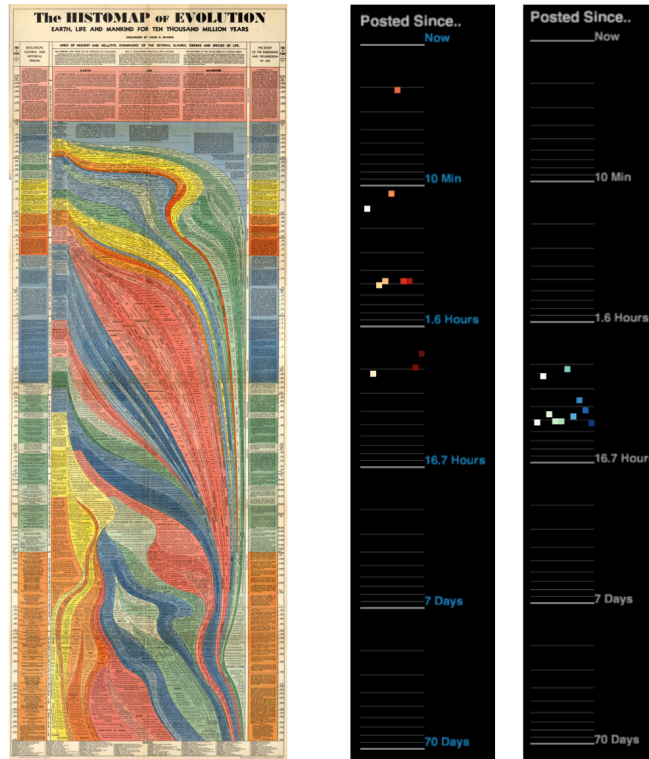


Figure 6.4: The Histomap of Evolution, the former logarithmic timeline visualization of geologic and human history, by John B. Sparks (1932) [207] (left) and the logarithmic timeline of TweetProbe (right)

interacting with the other by synchronizing two different threads, i.e., each layer has its own process thread. While the data processing layer responds to each message arrival, the visualization layer constantly gets updated at 40 frames per second, reflecting new message or backend analysis updates in its animation. The entire system is developed using Java and the front end makes use of Processing (Processing is an open source programming language and integrated development environment (IDE) built on the Java language.⁵). The overall system architecture can be seen in Figure 6.6

⁵<http://www.processing.org>

Twitter Stream Filtering

In our TweetProbe framework, the Twitter Streaming API is used to provide bulky tweet updates in real-time. The Twitter Streaming API brings a real-time stream of tweets into our system through a User Datagram Protocol (UDP) network connection. Users can either directly receive the entire message stream or extract topics of interest using a keyword filter.

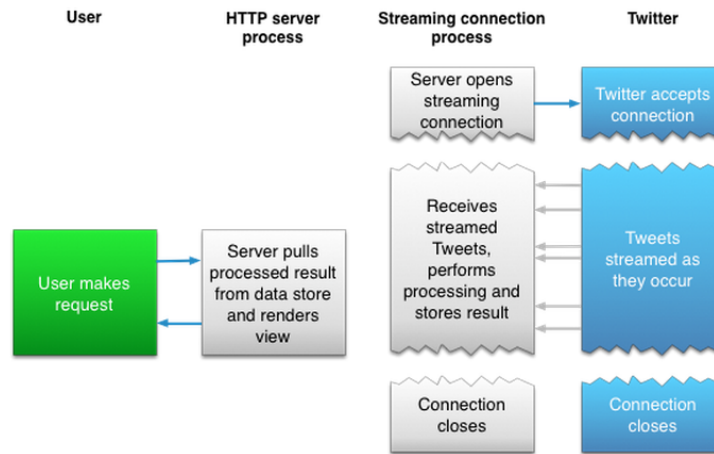


Figure 6.5: Twitter Streaming API Overview (Excerpt from dev.twitter.com)

Back-end Data Processing

TweetProbe has a back-end process that receives the incoming data stream, filters out irrelevant information when any keyword is applied to the system by a user, inserts new messages into the system's data structures or updates existing data entries by comparing each newly arrived message to the ones already stored. Both the tweet ID and message text are used for the comparison task and this information is enclosed with each message as a tuple in JSON format.

Data Structure The Twitter Streaming API delivers multiple tweets per second (up to 60 tweets per second) and each entity is encapsulated using JSON data structures. Each entity contains a tweet message and user (author of the message) information, along with a number of metadata that describe both the message and the user who updated that message. If the message is a retweet of another message, it also contains the original message and metadata. Simplified structure of an entity is shown in Figure 6.7.

Once we detect a new message that falls into our interest (e.g. retweet, popularity, recency, keyword-matching), the message is sent to the sentiment extractor module.

Sentiment Extractor One of the main components of our system is the sentiment extractor; see Figure 6.6. In TweetProbe, we applied both the corpus-based sentiment extraction algorithm from [30] and a simple emoticon extraction method to compute an overall sentiment score for each message. Once the score is computed, it is normalized into a linear scale from -10 to 10 which represents the degree of sentiment negativity or positivity respectively. This score is then visualized between two different colors (red and blue). When a message is closer to a neutral sentiment (sentiment score 0), it is expressed in white color by reducing the ‘saturation’ component of the color in HSV space. In a future version of the system, the sentiment processing can be updated to a more multi-faceted sentiment visualization such as [153].

Front-end Visualization Layer

The TweetProbe framework has four main visualization modules in which we present four different aspects of the microblog message stream: (1) sentiment and user distribution, (2) current most emerging messages, (3) most shared retweets and (4) most emerging hashtags (topics). Each module has its own screen and users can simply switch between each module by interacting with the system. Each visualization module is described in

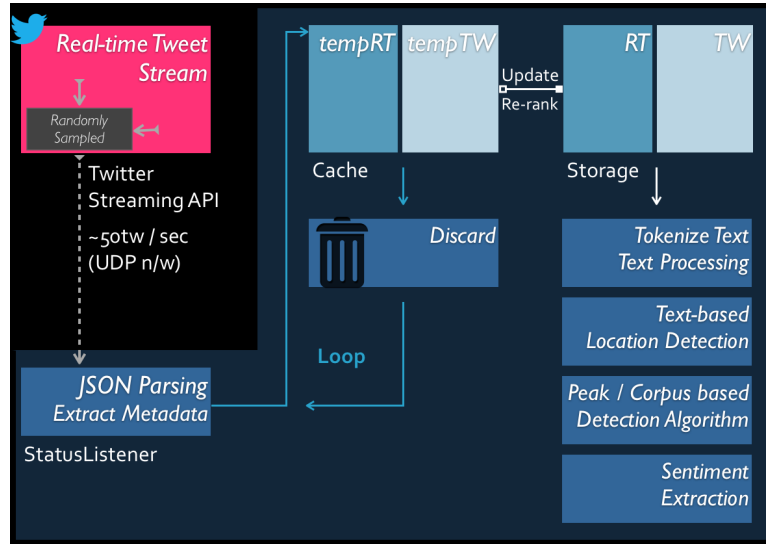


Figure 6.6: Tweet Stream Probe System Architecture

detail in the following section.

6.3.4 Visualization

In this section, we discuss our visualization designs by focusing on the individual visual components of *TweetProbe* system.

Sentiment Map (Raindrop Message Visualizer)

The sentiment map provides loosely-organized, but sentiment-oriented, raindrop visualization to users. We intended to keep this animation focusing more on aesthetic factors by compromising visual efficiency in order to deliver more natural feeling of information stream to end users. As can be seen from its sub-title: “Raindrop Message Visualizer”, each message is visually expressed as if a raindrop falling from the sky making a circular wave on water surface. Users can experience stream of messages coming through in multiple visual components. As depicted in Figure 6.11, each tweet arrival is expressed as a circle element. According to the legend upper left of the screen, each item is color coded

| Tweet Metadata | User Metadata |
|----------------------------|------------------------------|
| Update Time | Account Birthday |
| Update Source (App) | ID, Name, ScreenName |
| Reply to (username) | Location (text) |
| Geo Coordinates (optional) | Geo Coordinates (optional) |
| Place (text) | Account Description |
| Retweet Count | Follower/Friend Count |
| Favorite (Like) Count | Status/Favorite (Like) Count |
| Language | Language |
| URLs | Time Zone |
| ... | ... |

Figure 6.7: Tweet Metadata and User Metadata

upon its sentiment score. Since this visualization presents real-time data stream, we use fading animation effect to help users see exact time of arrival of each message. Moreover, each tweet is mapped on the grid (using logarithmic scale with base of 10) according to one's retweet count (*y-axis*) and the author's followers count (*x-axis*). This is because the both metrics show the potential of dissemination of an individual message in social network, particularly Twitter. For example, the biggest red-colored circle in Figure 6.11 shows that the message has been retweeted more than 100 times, having strong positive sentiment on the message. It also shows that the user posted this message has more than 100K followers. As with these clues, we can easily understand that this is an organizational account. Green-colored contour means it is a retweet of another article and text label next to each circle shows location of each user.

Figure 6.12 shows a series of screen shots from the first draft to the final version of the Real-time Sentiment Map. (The final version (Figure 6.12 (b)) includes sentiment extraction algorithm)

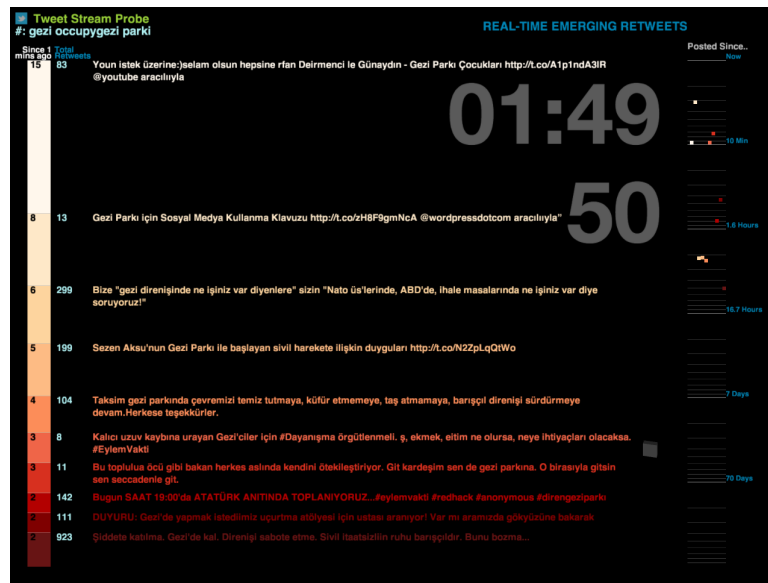


Figure 6.8: A screen shot of the real-time emerging retweet visualization. The animation on the left represents the number of new updates for each message since the application(TweetProbe) launched. Each segment and its corresponding message is color-coded with the square marker on the timeline graph, which is on the right side of the screen. The timer in the upper-right region of the screen shows the time elapsed since the program has been launched and the number below ('50') is the total number of new retweets since the launch of application.)

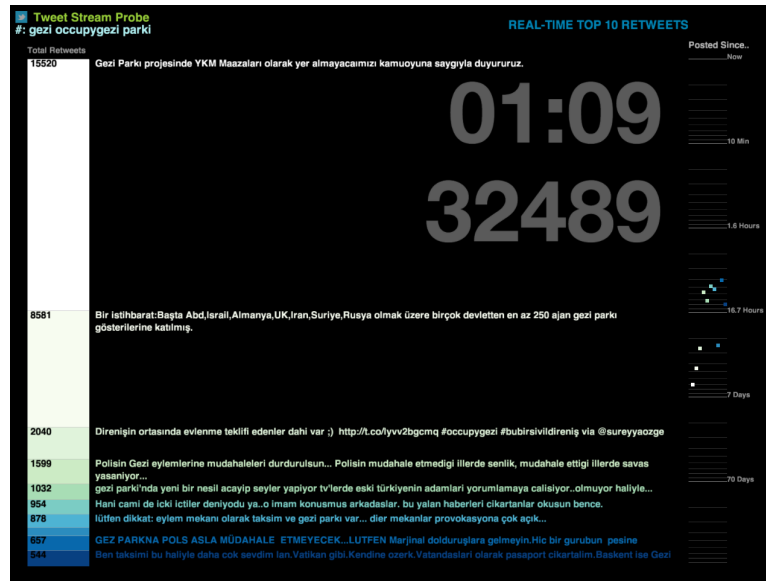


Figure 6.9: A screen shot of the real-time top 10 retweet visualizer (the Timeline graph on the right side of the screen shows each tweet origin’s time of creation. Since it is expressed in log-scale regarding relative time difference to current time (indicated as ‘Now’ at the top of the screen), it is easy to compare tweet times to one another or to current time.)



Figure 6.10: A screen shot of the real-time top 10 hashtag visualizer (The spinning box on the right-bottom side of the screen visualizes total count of the newly updated hashtags since the program has been launched.)

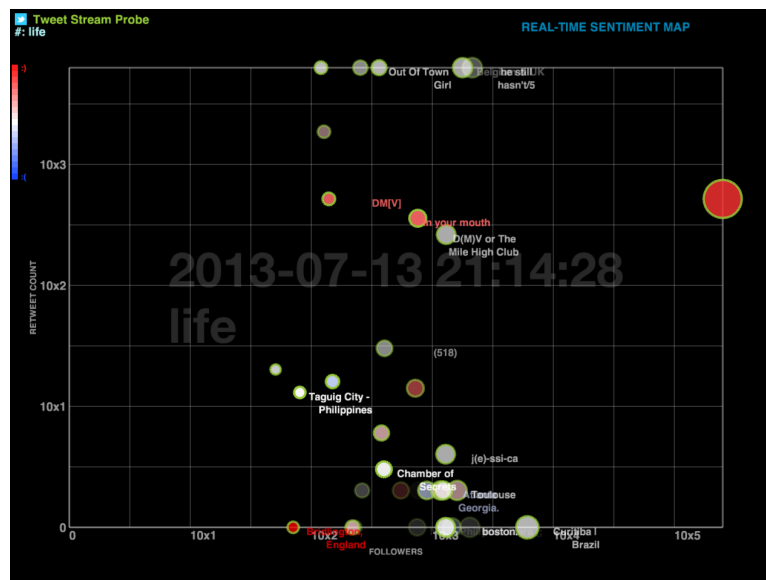
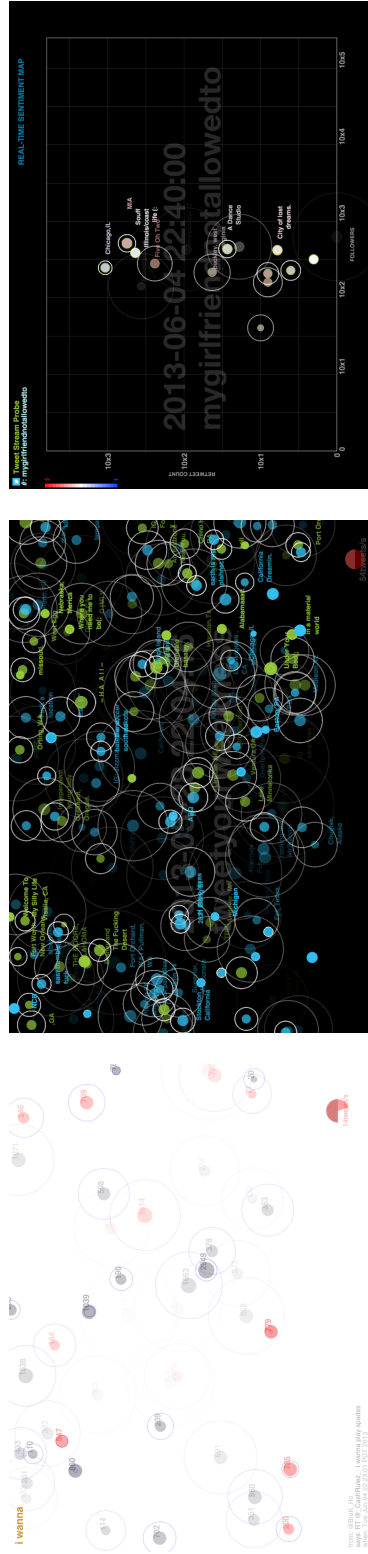


Figure 6.11: A screen capture of the sentiment map visualization (keyword “life” is used).



(a) Random scattering approach

(b) Spatial mapping enabled with grid

Figure 6.12: Raindrop visualization in different approaches. From the aesthetic perspective, we found that the random scattering approach (a) draws more attention (impression of raindrops) than the grid map (b) from the users.

Real-time Ranking Visualization

This visualization technique is designed to provide more organized information regarding retweets to its targeted users such as social network analysts. There are two main components: *sliding animation* and *log-scale timeline*. These components are applied to convey transient information (statistics and rankings) from Twitter stream on every 10 milliseconds. Users can also confirm that which message is currently competing against another through sliding animation with (1) message count (since the application launch time) and (2) total retweet count. The two counts are measured based on the original post of each retweet.

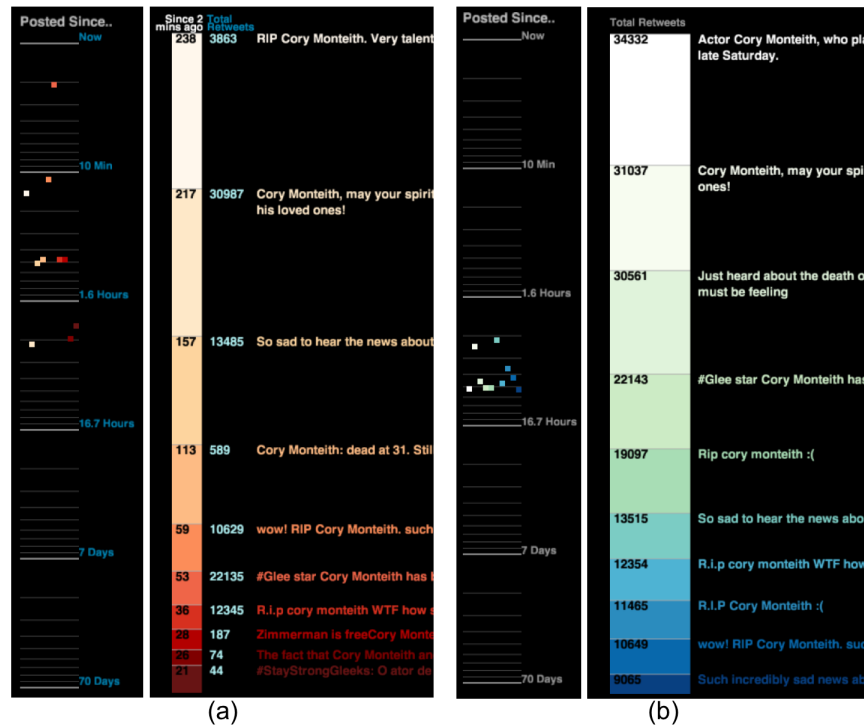


Figure 6.13: Real-time retweet ranking visualizations. (a) real-time emerging retweet ranking (right) and its update timeline in log-scale (left). (b) real-time retweet count ranking (right) and its update timeline (left).

Real-time Emerging Retweet Ranking The first view mode is named as “Real-time Emerging Retweet Viewer”. This visualization presents N (can be chosen by a user) most currently emerging retweets. As this view mode initiated, TweetProbe continuously counts new retweet arrival and updates a ranking of N (in this example, $N = 10$) retweets in real-time. To be specific, our back-end system counts origin of each incoming retweet and re-rank the list as a newcomer arrives on our system from the data stream. As can be seen in Figure 6.13 (a) and (b), live trending retweet does not always entail large number of retweet count, i.e., it can also be a new message just published a few minutes ago. Interestingly, many outdated messages that have large number of retweet count can be easily found if we compare visualization (a) to (b) in Figure 6.13. This phenomenon implies that timely information is the most important factor in Microblog space, especially when it comes to the latest social event, and it is one of the most significant and representative characteristics of social media. Another interesting point to note is the notable difference in the log-scale timeline between (a) and (b) in Figure 6.13. While emerging retweet ranking visualization shows various posting times of each message, most of messages in the ranking of retweet count often show a cluster that is shown in Figure 6.13 (b). We believe that this is caused by the time lapsed to accumulate enough number of retweet to be in this ranking regardless of how much it is trendy topic.

Real-time Retweet Count Ranking The second view mode shows top N retweets based on the number of retweet count of each message. In this visualization, we can see trending topics in macroscopic point of view regardless of the fact that they are currently active or not. This metric is also important because total retweet count reveals entire link of message dissemination since its birth. However, this visualization is more useful to detect past or present messages that already discussed with lots of users. This visualization can be seen in Figure 6.9.

Real-time Hashtag Ranking The last view mode provides top N hashtags, which can be considered as topics of messages, along with each hashtag’s time of birth. In this visualization, each hashtag’s text size is mapped with its ratio of hashtag count in the top-10 list so that users can see the quantitative contribution of each topic to the rank. This visualization is shown in Figure 6.10.

6.3.5 Deployment, Reception, and Discussion

As described in the previous sections, the TweetProbe system was conceptualized both as a tool for information workers, as well as a creative media arts installation that can alert audiences to up-to-the second information from the world around them. In this section, we discuss the deployment of our visualizations as part of a creative art work and discuss feedback that we have received during a showing to media arts professionals and the general public in May 2013 ⁶. This is followed by an application example for using our visualization tools on a real-world event for a specific time frame.

Continuum of Discontinuity

The main concept of our design is the expression of continuity from discrete data points existing in social media. We have been taking note of the fact that the online social space exhibits speedy and dynamic transitions of topics underlying the ongoing discussions. For example, even regarding the same story people focus on different facets of it as time moves forward, changing their stance on each topic. Thus, we aimed to visualize time-variant stories marked with the author’s sentiment through transient shapes and colors in abstract visual components. Since we need to deliver numerous information

⁶TweetProbe was part of the installation “Continuum of Discontinuity”, which was shown to hundreds of visitors of the public “Shadows in Space” event in May 2013 - an annual open exhibition at the University of California, Media Arts and Technology program. <http://show.mat.ucsb.edu/>

elements in real-time, visual components are designed to be as simple as possible. Tweet-Probe was set up as a media arts installation and exhibited on the 23rd of May 2013, as part of the UCSB Media Arts and Technology End of Year show "Shadows in Space." Different visualization techniques including random scattering raindrops were projected on the wall taking in turn. Audiences were allowed to select their own keywords of interest and enjoy the resulting visualization animations. Throughout this exhibition, hundreds of people visited our installation work and left valuable comments about their perception on it. Most of the people communicated positive impressions on both the sentiment map and ranking visualization. It became clear from the feedback that the raindrop visualization seemed more intriguing and engaging in an abstract sense and worked better at pulling people in for a closer look while the ranking visualization provided more information and was easier to comprehend without any guidance. Moreover, audience feedback confirmed that their interest was indeed aroused by the raindrop metaphor, particularly on the random scattering approach as depicted in Figure 6.12-(a) with color-coded sentiment scores. A few spectators also commented that it is interesting to see the live competition in ranking among different messages on the real-time emerging retweet visualization.

A Scenario-based Observation

For further discussion, the potential practical benefits of using our visualizations are observed in a real-world scenario. We ran our visualization framework with the keyword *#royalbaby* which was one of the trendy topics as a few weeks ago (20th of June, 2013). We have captured the most emerging tweets in real-time and the most retweeted message based on retweet counts. The observations reported here were performed on July 22, 2013 between 5 and 10 pm. As can be seen in Figure 6.14, each message represents the most emerging tweet in a 10-minute time window since 8:30pm (in BST - British Summer

Time). In this figure, the topic transition for the same event can be easily observed throughout each time window. The second message at 8:40pm announces the birth of the royal baby and the fourth message delivers an image through an embedded URL from an official information source. The next peak of previous message is found on the 5th message. While this emerging retweet visualization reveals live news on a social event, the retweet count ranking visualization showed the same message as a top-ranked tweet, which is several hours behind of the latest news.

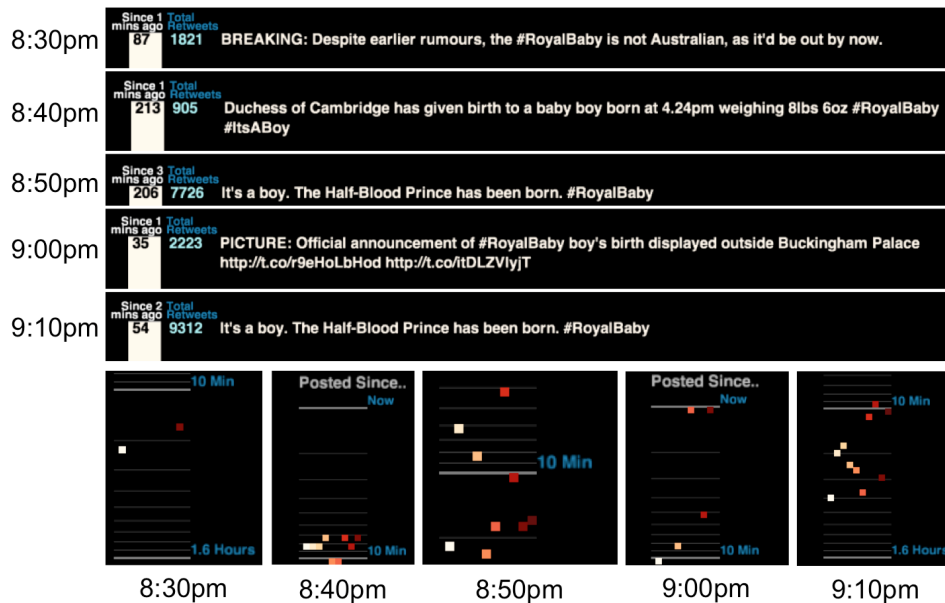


Figure 6.14: Top emerging tweets and their update times captured from the real-time emerging retweet ranking visualization. (Keyword *royalbaby* was used.)

6.3.6 Summary

The Tweet Stream Probe visualization framework is designed to sense real-time topic-specific trending information on Twitter. In this visualization framework, we implemented both a back-end data processing layer and front-end information visualization layer using the Java and Processing programming languages. The first data processing

layer filters out unnecessary information from the connected tweet stream, updates trending tweets, extracts underlying metadata and sorts tweets, retweets and hashtags. All of these tasks are performed multiple times each second. We believe that this system can serve social media analysts well for finding useful information or interesting patterns. At the same time, the real-time depiction of social media information can be the basis for engaging and intriguing public art installations, as it reflect the current state of the world from a specific medium's perspective. Our work presented some steps in this direction.

For the future work of TweetProbe, we will add additional features such as network analysis, community detection algorithm and a richer user interface. Additionally, we plan to extend our visualization on the Web in order to reach larger audience.

6.4 Inspectability and Personalization in Social Content Discovery

The primary focus of our study is how to model information reliability and identify highly reliable information on the Social Web. While we developed robust models with high prediction accuracy, if users can not communicate such high-quality information with ease, our good models and algorithms may be a good-for-nothing. In particular, when the information window that connects both users and information as a communication channel is not available or sufficient enough to users with given amount of information, this problem becomes critical.

In this study, we focus on the informational and user experience benefits of exploring topics within microblog communities, in a transparent, controllable and personalized manner. To this end, we introduce HopTopics – a novel interactive tool for exploring content that is popular just beyond a user’s typical information window in a microblog. We present results of a user study (N=122) to evaluate HopTopics with varying complexity against a typical microblog feed in both personalized and non personalized conditions. Results show that HopTopics system, leveraging content from both the direct and extended network of a user, gives users a better sense of control and transparency. Moreover, participants had a better mental model for the degree of novel content discovered when presented with personalized data. Lastly, we provide design improvements for better user experience in social content exploration based on the results.

6.4.1 Introduction

With large amounts of noisy, user generated content in social media, we have no choice but to rely on automated filters to compute relevant and personalized information that

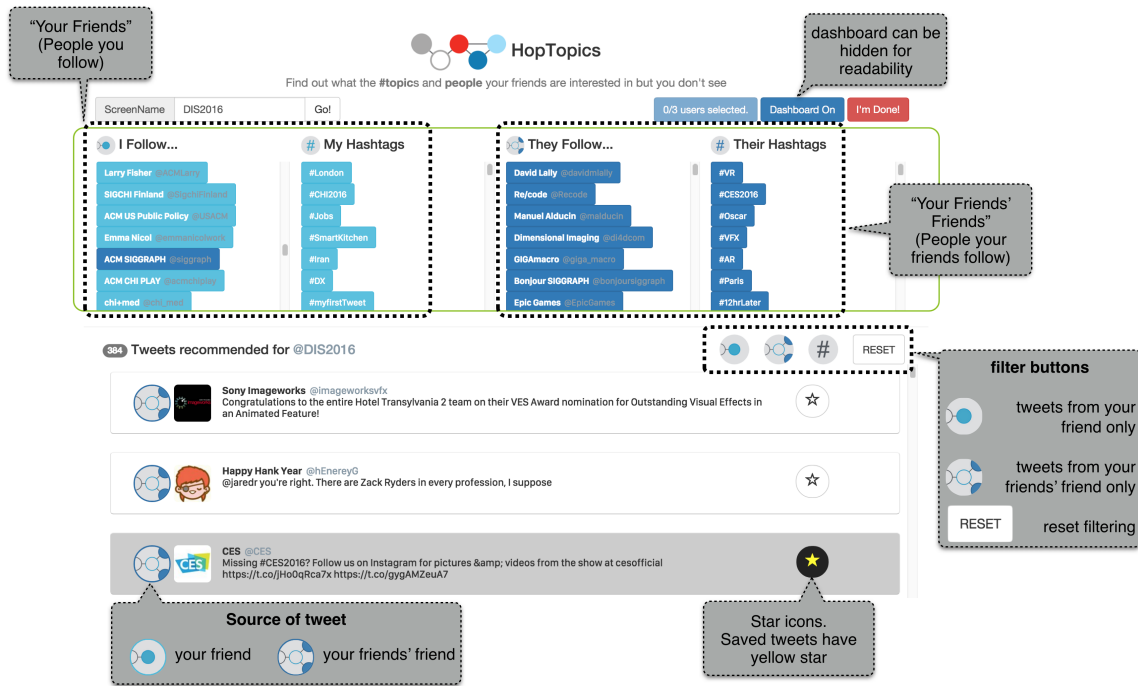


Figure 6.15: Screenshot of the system (condition D), with labels indicating how to interact with the system and what the various components are. At the top there is a dashboard controlling the source of the tweets using community structure (1-hop and 2-hop followers) and topics (hashtags). The resulting tweets are shown below and can be filtered and starred by users.

are small enough to avoid cognitive overload. However, once an automated information filtering mechanism of any type is applied, there is a real risk that useful, or perhaps critical information will never reach the end user. This problem is not new: Smyth and McClave argued in [208] that there is a sweet-spot between similarity and diversity in personalization, [174] refer to it as a general black-box problem with recommender systems, and more recently, Pariser [209] and Resnick [210] describe it as a filter bubble problem, wherein personalized filtering algorithms narrow a user’s window of information on the world.

In social networks such as Twitter, a user’s information feed is populated with content from the other users that they follow directly. Here, filtering can be seen as a two step process. First, the user must elect to follow another user, and second, that user acts as

an information curator by either authoring or propagating messages. Both steps in this process are subject to failures. For example, in the first step, Alice might follow Bob because she is interested in what he has to say about computer science. However, Bob might not post much information about that topic (as studied in [171]). In the second step, Bob, acting as an information filter, can propagate noisy or misinformation (studied by Morris in [211]).

Allowing people to see how their social network influences the information they receive may help alleviate these issues. We therefore propose and evaluate a novel approach to personalization in Twitter, which goes some way towards addressing filter bubble problems. We introduce HopTopics – a system that enables users to leverage their network to source novel and potentially relevant topics, and then to seek information on those topics from both the local and from the extended social network. The approach can be viewed as a hybrid of strong and weak ties [212] for personalized information seeking. By leveraging direct and/or strong ties for new topics, a user gets an idea about the topics that her networks see but she does not, and by leveraging weaker ties in the broader network for details about those topics, the system attempts to provide a broader perspective, unbiased by the particular views of a local clique of users.

6.4.2 Background

To frame this research in the context of related work, we look at three key areas. First, we discuss related work on interaction with intelligent systems for information retrieval, of which recommender systems are a key component. Second, we focus on interaction and control of data in microblogs – a topic which is central to our research and has also received much attention in recent years. Finally, we present a discussion of related work in the area of community based content discovery, which is the main focus and novelty

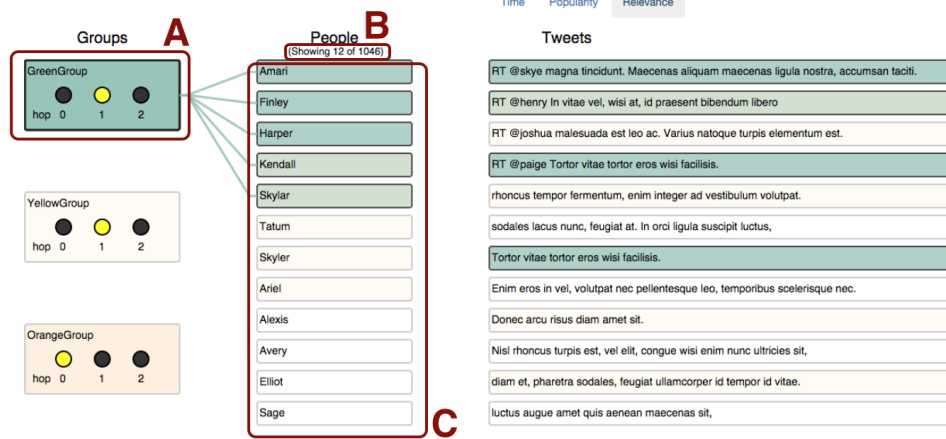


Figure 6.16: Initial UI design evaluated in the formative study. Annotation (A) shows changes to the number-of-hops selection. (B) shows the number of filtered users interactively in the form “m of n”, and (C) shows connectivity-based clustering and associated coloring of nodes in the “People” column.

of this study.

Inspectability and Control in Intelligent Systems Mechanisms for improving inspectability and control in intelligent systems have been introduced to different classes of systems from open learner models [213, 214], to autonomous systems [215, 216], decision support [217, 218] and recommender systems [219, 220]. These studies have found that inspectability and control can have a positive effect on user experience as well as improved mental models.

Over time, there has been a shift toward supporting more open-searches and users’ understanding of novel domains evolving through use [221]. This has also meant an evolution from static explanations to more dynamic forms of explanation such as interactive visualization.

For example, [222] has looked at how interaction visualization can be used to improve the effectiveness and probability of item selection when users are able to explore and interrelate multiple entities – i.e. items bookmarked by users, recommendations and

tags. Similarly, [223] found that in addition to receiving transparent and accurate item recommendations, users gained information about their peers, and about the underlying algorithm through interaction with a network visualization.

Inspectability and Control in Microblogs In order to better deal with the vast amounts of user-generated content in microblogs, a number of recommender systems researchers have studied user experiences through systems that provide transparency of and control over recommendation algorithms. Due to the brevity of microblog messages, many systems provide summary of events or trending topics with detailed explanations [166]. This unique aspect of microblogs makes both inspectability and control of recommender algorithms particularly important, since they help users to more efficiently and effectively deal with fine-grained data. For example, experimental evidence to argue that inspectability and control improve recommendation systems is presented for microblogs in [223], via a commuter traffic analysis experiment, and more generally in [224] using music preference data in their *TasteWeights* system.

Community-based Content Discovery *Serendipity* is defined as the act of unexpectedly encountering something fortunate. In the domain of recommender systems, one definition has been the extent to which recommended items are both useful and surprising to a user [9]. This study investigates how exploration can be supported in a way that improves serendipity, and maintains a sense of inspectability and control, for example through the interfaces in Figures 6.15 and 6.16.

The intuitions guiding the studies in this study are based on findings in the area of social recommendations, that is based on people's relationships in online social networks (e.g., [225]) in addition to more classical recommendation algorithms.

The *first intuition* is that weak rather than strong ties are important for content

discovery. This intuition is informed by the findings of the cohesive power of weak ties in social networks, and that some information producers are more influential than others in terms of bridging communities and content [212]. Results in the area of social-based explanations also suggest that mentioning which friend(s) influence a recommendation can be beneficial (e.g, [226, 227]). In this case, we support exploring immediate connections or friends, as well as friends-of-friends.

The *second intuition* is that the intersection of groups may be particularly fortuitous for the discovery of new content. This is informed by exploitation of cross-domain model inspiration as a means for serendipitous recommendations, e.g., [228].

6.4.3 Formative User Study

In this section we briefly summarize the (previously published in a workshop) findings from a formative user study (N=12) to evaluate the interface and interaction design. Information feeds, network structures and individual users in Twitter can influence user experience with a system such as HopTopics when it is evaluated in on real world data. The design of the initial interface is shown in Figure 6.16. Informed by these studies, an iterated design of the interface and interaction was applied and used for the real world experiment described in the next section. In this figure, the column on the left shows icons that represent different communities that a user is potentially interested in. Let's say ACM DIS, and two other related conferences, for example. The radio buttons in each group specify a hop distance to traverse from that groups' page, with a 0 value essentially disabling that group. The middle column shows prominent users from these groups, color coded by their source group. Content is shown in the right column, and is again color coded according to source. If a user hovers over a given node, edges appear dynamically to show all of its connections, as illustrated between A and B in the figure.

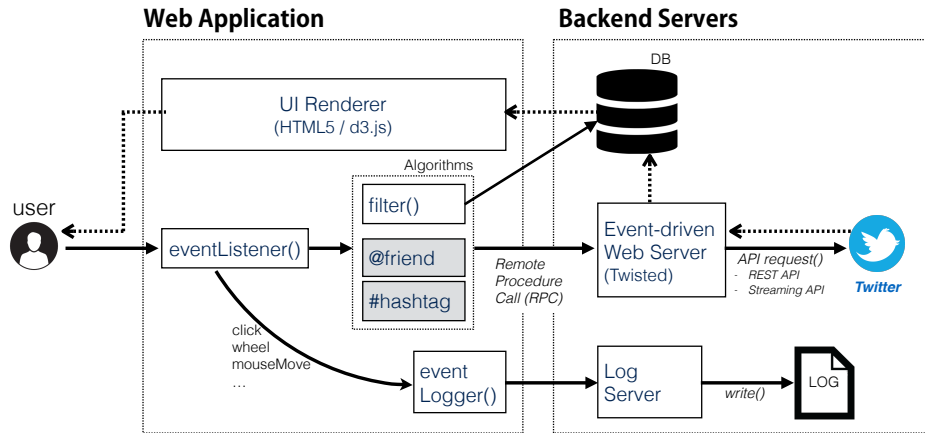


Figure 6.17: Architecture diagram of the HopTopics system. Solid and dashed arrows represent input or request sent and response in return, respectively.

In the initial formative study we used a layered evaluation approach [229], focusing on the decision of an adaptation and how it was applied (in contrast to which data was collected or how it was analyzed). So, to isolate some aspects of user interface and interaction design, we chose to perform an initial evaluation of the HopTopics interface using obfuscated (Lorem Ipsum) data. Participants in the two countries interacted with the interface in semi-structured interviews. In two iterations of the same study ($n=4$, $n=8$), we found that the interface gave users a sense of control. Users were asked for an active selection of communities, and a more fine-grained functionality for saving individual ‘favorite’ users. Users found the community-based exploration feature to be particularly useful and also highlighted unanticipated uses of the interface such as iteratively discovering new communities to follow, and organizing events.

6.4.4 HopTopics System

To recap, the goal of the HopTopics system is to support community-based content discovery in microblogs such as Twitter. The overarching concept behind the tool is to allow the user to interactively explore content in real-time, sourced through channels of

strong rather than weak ties –for identification of potentially relevant and novel topics, and then through weak rather than strong ties for information about these topics, to mitigate biases that may be introduced by opinions in tightly connected cliques. Of course, a user might be specifically interested in what a particular clique is saying about a target topic, so the system is designed to allow for both types of interactive search, either independently, or in a hybrid result set.

To achieve the goal of real-time interactive exploration of network data, several design and engineering challenges must be addressed, most notably given the strict data access limitations imposed on the Twitter data end point, commonly known as “rate limits”⁷. In this section, we first describe the UI design behind the HopTopics system, from the initial design used in the formative study, to the final design, shown in condition D of the main study, and in Figure 6.15. Next, we describe the interaction design, with a focus on the trade-offs between information and cognitive overload, recall of relevant topics, and practical rate limitations. Finally, we describe the novel architecture that supports real-time network-based, topic-specific data exploration the HopTopics.

User Interface Design

Figure 6.15 shows a screen shot of the training screen for the system and indicates the various components. The dark grey speech boxes illustrate the basic components of the system. The system has two core components. First, a *Network Dashboard* shows the active user’s one and two hop network along with the topics/hashtags that are prominent in them. Second, a *Content Viewer* panel shows a collection of the messages that are derived from the current set of selections made in the dashboard view. The content viewer shows an iconized combination of messages from the different network regions: one hop messages, two hop, and global. Messages from each group are shown with a source

⁷<https://dev.twitter.com/rest/public/rate-limiting>

provenance icon, shown in the left side of Figure 6.3. Within this viewer, participants can elect to filter messages based on source type. For example, by clicking on the “one-hop” filter button on the right side, the viewer will only show messages related to the hashtags that come from the one-hop group. Due to limited screen real estate for most web users, an important feature of the system is the ability to retract the Network Dashboard – which is essentially a navigation mechanism – and focus only on the Content Viewer – which contains the material they are typically interested in. Last, a color coding scheme is applied to the Network Dashboard to indicate links between groups of topics/hashtags and the network regions they originate from, as shown in dark blue and cyan in the Network Dashboard of Figure 6.3.

Complexity and Limitations Since nodes expand exponentially as one traverses the Twitter network, query complexity and data relevance were primary design considerations. In an ideal scenario, the HopTopics system would be connected via a firehose⁸ connection where complex queries would not pose quite as much of a constraint. However, given limited bandwidth for our experimental setup, node selections were limited to three selections for each column. The selection limit is shown dynamically on the top right of the view window.

Interaction Design

HopTopics supports a number of different workflows, depending on the data exploration task. A user begins by typing a query into the system. This is typically their own Twitter ID, but could also be another user whose network they are interested in exploring. The API is queried and network and content information are displayed in the appropriate panels as described above. A user can mouse over any of the entities listed

⁸<https://dev.twitter.com/streaming/firehose>

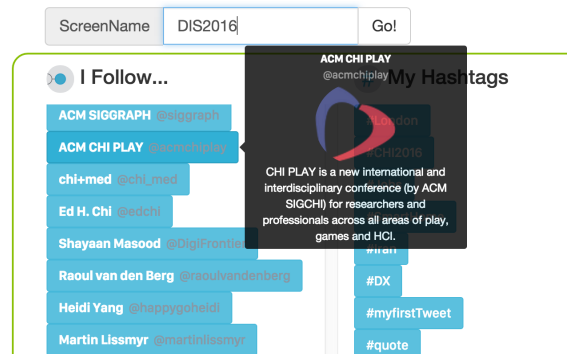


Figure 6.18: Example of a mouse-over query in the HopTopics Network Dashboard. Asynchronous queries allow for millisecond delay time only.

in columns to find out more detail. An example of this mouse-over result is shown in Figure 6.18. Users interact with the dashboard by first selecting up to 3 people in the left column, which consists of all the people they follow (1-Hop follows). As a user clicks on one of the people in the first column, it immediately populates the third column: people who they follow (2-Hop follows, or friends-of-friends). When the user select 2-Hop follows, further hashtags get shown in the “their hashtags” column furthest to the right. This also adds more tweets at the bottom, the message panel, that were authored by the selected people. As in typical Twitter feed interfaces, users can “star” or favorite tweets in the Content Viewer. The system takes into account several of the suggestions mentioned in the formative study. For example, to maintain information provenance, icons are used to annotate whether a tweet comes from someone the user follows, or if it is from someone two hops away. No information is discarded, and users can scroll down to see e.g. the full list of people they follow. Additionally, users can see how many tweets are available to them, giving a sense of the degree of filtering being performed by the current view of the system. At any point in a data exploration session, a user can hit the ‘reset’ button to return the system to its default view of the network.

| | Baseline | Augmented Data | Inspectable | Controllable |
|------------------|----------|-------------------|-------------|--------------|
| Non-personalized | A1 | B1 | C1 | D1 |
| Personalized | A2 | B2 | C2 | D2 |

Table 6.3: Overview of conditions. Degree of personalization is within participants, system type is between participants.

System Architecture

The HopTopics architecture is described in Figure 6.17. The system has three main components: First, a front end application which includes the user interface renderer. These run in the client browser. Second, an event logger and, third the remote back end server. Remote procedure calls are used for the communication between these components. Python’s Twisted⁹ framework has been used for the back end server and the event logger because of the ease of use for asynchronous event and data handling.

The main considerations in this particular design are scalability and user experience. In particular, due to the rate limitation policy of the Twitter API, we exploit caching algorithms wherever possible, in order to prevent exceeding the given rate limit. On each API request, the system stores every returned data into a heap memory, which has been prepared for the use of cache, and this can be used when another API request needs the same data without consuming the limited quota. We also integrated a cluster of back end servers so as to avoid the same issue. When there is a new session opened, the system allocates the user to one of the idle servers in order to balance concurrent API requests. This design is particularly critical since part of our experimentation was performed online, and resulted in many parallel sessions with the system.

⁹An event-driven networking engine written in Python. <https://twistedmatrix.com>

6.4.5 Main Experiment

In this section we describe an experiment to evaluate the interface using real world data. We evaluated it in terms of its ability to support content discovery, the perceived quality of discoveries, the correctness of this assessment, as well as perceived transparency and control.

The experimental toolkit was deployed as a web service and the link was made available on Amazon Mechanical Turk (AMT). The AMT web service is attractive for researchers who require large participant pools and low cost overhead for their experiments. However, there is valid concern that data collected online may be of low quality and require robust methods of validation. Numerous experiments, such as [230] have attempted to show the validity of using the service for the collection of data intended for academic and applied research. These studies have generally found that the quality of data collected from AMT is comparable to what would be collected from supervised laboratory experiments, if studies are carefully set up, explained, and controlled. We carefully follow recommended best practices in our AMT experimental design and procedures.

Experiment Design

The experiment used a mixed design, as shown in Table 6.3. The system variant was assigned between participants and was one of: A) Baseline - standard Twitter feed only, B) Data - augmented feed including topics mentioned by friends of friends, C) Inspectable - dashboard visible but not interactive, D) Controllable - interaction with dashboard, this is the full system introduced in the previous section. The Twitter API has a limitation on the number of accesses per time unit. To minimize the impact ceiling, the experimental sessions were run consecutively with one concurrent session at a time.

These conditions were compared between rather than within participants, in order

to avoid learning and ordering effects for a specific Twitter account. Instead we compared within participants the effect of personalization of the data, comparing 1) a non-personalized id (always the same id: @ACMIUI) with 2) a personalized (using their own Twitter ID).

The condition using the data for the non-personalized Twitter id was always shown first. While this data was retrieved live, it was not currently in progress. This is also a relatively small community on Twitter, so the dataset is relatively static, and contains many topics that may be unfamiliar to the average Twitter user.

The motivation for using such a dataset is i) to create familiarity with the interface through training, and ii) to have a condition where we expect participants to have a consistent level of familiarity (low) with the content, as it is not personalized. This design means for example that participants assigned to the Augmented Data condition would always see first B1 and then B2.

In addition to the responses we collected and computed the following indirect measures:

- Number of people saved
- Number of tweets starred
- Number of hashtags saved
- Correlation between perceived novelty and number of hashtags identified as novel.

Hypotheses

We hypothesized that the system will help users discover more unexpected and useful content, and lead to better usability perceptions when the system was controllable and

when the content was personalized. Specifically, our hypotheses were:

H1: Perceived serendipity.

“Compared to your regular twitter feed, how much does this interface help you find relevant and surprising items that you did not know about yet? (0=not helpful, 100=very helpful)”

H1a: Perceived serendipity will be higher in more interactive and transparent conditions (Baseline < Data < Inspectable < Controllable).

H1b: Perceived serendipity will be (slightly) higher in the personalized compared to the non-personalized conditions, across types of system.

H2: Perceived familiarity.

“Compared to your regular twitter feed, how helpful is this interface for finding information that is both relevant and familiar?”

H2a: Perceived familiarity will be higher in more static and opaque conditions. (Baseline > Data + Inspectable > Controllable).

H2b: Perceived familiarity will be higher in the personalized compared to the non-personalized conditions, across types of system.

H3: Perceived transparency.

Perceived transparency will be higher in more interactive and transparent conditions. (Baseline < Data < Inspectable < Controllable). We do not anticipate a difference in perceived transparency between the personalized and non-personalized conditions.

H4: Perceived control.

“The interface helped me change the tweets that are recommended to me. (0=not

helpful, 100=very helpful)” Perceived control will be higher in more interactive and transparent conditions (Baseline = Data \leq Inspectable < Controllable). We do not anticipate a difference in perceived control between the personalized and non-personalized conditions.

H5: Content discovery.

H5a: Degree of content discovery (sum of people + hashtags + tweets saved) will be greater in the interactive conditions (Inspectable < Controllable).

H5b: Degree of content discovery will be greater in the personalized than the non-personalized conditions.

H6: Correctness of mental model.

H6a: The correlation between perceived serendipity (subjective) and content discovery (objective) will be higher for the interactive conditions (Inspectable < Controllable).

H5b: The correlation between perceived serendipity (subjective) and content discovery (objective) will be higher for the personalized compared to the non-personalized condition.

H7: Perceived diversity

Perceived diversity will be higher in more interactive and transparent conditions (Baseline < Data + Inspectable + Controllable).

We do not anticipate a difference in perceived diversity between the personalized and non-personalized conditions.

H8: Increased Interaction.

H8a: There will be more interactions (e.g. click, double click, mouseover (hovering),

mouse wheel action, cursor trajectory) in the interactive conditions (Baseline \leq Data < Inspectable < Controllable).

H8b: There will be more interactions for the personalized data.

Participants

Participants were recruited from the US only and were required to have a Mechanical Turk acceptance rate of greater than 90% (at least 90% of their HITs are considered of good quality by other requesters). They were required to correctly answer some filler questions, and to have a minimal degree of interaction with the system (2 minutes and 1 interaction).

155 participants completed the full study, however 33 participants were excluded from analysis as they had technical issues (most likely they interacted with the system beyond Twitter's rate limitations). Out of the remaining 122, the distribution across the 4 versions of the system was (A=32, B=32, C=27, D=31). The lower number of participants in conditions C and D is due to a lower completion rate in these conditions. User comments suggest that this is due to the system being slow.

The majority of participants (50%) were aged 25-35, with a similar proportion of participants aged 18-25 (24%) and participants aged 35-50 (22%). Only 4% were aged 50-65. Participants were balanced across genders (49% male vs 51% female). The mean response for the personality trait openness to experience was 5.23 (SD=1.11) which is typical of the average population. 91% of the participants reported that they used Twitter "Sometimes" (27%), "Often" (39%), or "All of the Time" (25%).

Materials

Tweets were retrieved live at the time that the experiment was run and used to populate the interface described in Section 6.4.4. Tweets were either retrieved for the

@ACMIUI account (in the non-personalized condition), or the user's own Twitter id (personalized condition).

Procedure

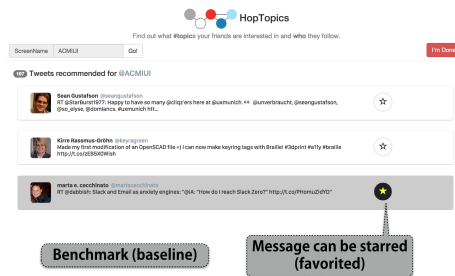
The procedure contained the following steps, described in detail below: Pre-survey ⇒ Instructions ⇒ Hoptopics Non-personalized ⇒ Post-survey1 ⇒ Instructions ⇒ Hoptopics personalized ⇒ Post-survey 2. Participants started the experiment with a pre-survey, including demographics, and the personality trait of openness to experience [231]. This survey can be viewed online at <http://anonymized>.

They were then taken to an instruction screen (see Figure 6.15). Here, they were given an open-ended task:

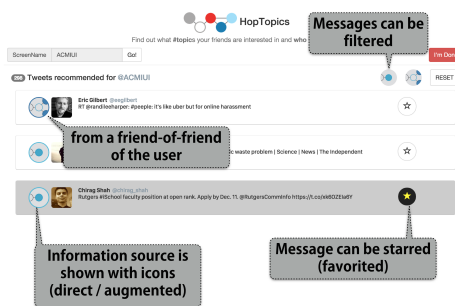
Imagine that you have just taken on a new role as a freelance journalist. You need to write a few pieces for a client. You can write them on any topics you find interesting and surprising. However, you need to send your boss a short summary on these topics by tomorrow! Your job is to find people and topics that help you with your task:

- Save people and hashtags by clicking on them.
- Star any tweets that you would use as the basis of the articles you are going to write.

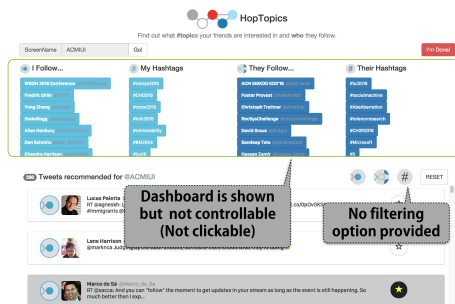
Once they closed the introduction screen, the main interface became visible with the non-personalized content. Participants could not move on to the next screen if they had interacted with the system for less than 2 minutes. In A and B conditions only interactions with tweets could be logged (Only for user-driven responses. System-wide logger recorded all interactions across all conditions separately.), while in conditions C and D also interactions with people and hashtags could be logged.



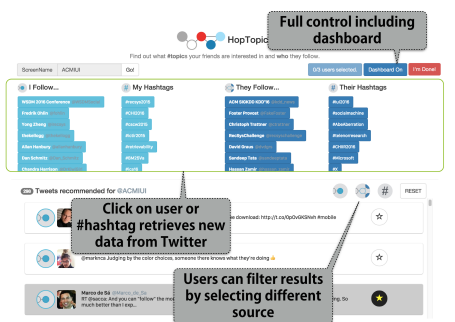
(A) Baseline: This is the typical message list view.



(B) Augmented Data: Augmented with new messages.



(C) Inspectable HopTopics with dashboard.



(D) Controllable HopTopics with interaction.

Figure 6.19: Screenshots of the four between subjects UI versions in the experiment.

Participants moved forward to a post-survey after they selected the “I’m Done!” button. Here they were asked about their perceptions of the system and the contents. The post-survey can also be viewed online at <http://anonymized>.

Next, participants were taken to the personalized variant in their condition where they were asked to enter their own Twitter ID. They performed the same task a second time, and were taken to the same post-study for this second interaction.

Results

The distribution of responses were not normally distributed for any of the variables, and non-parametric tests (Kruskal-Wallis or Wilcoxon) are used consistently to compare between conditions.

H1: Perceived serendipity There was no significant difference between versions of the system with regard to the degree of perceived serendipity (Kruskal-Wallis chi-squared = 5.8, df = 3, p-value = 0.12). There was also no significant difference between the perceived serendipity for the non-personalized and personalized conditions (Wilcoxon rank, $W = 5876.5$, p-value = 0.71). Tables 6.4 and 6.5 summarize the means. We observe a trend for the perceived serendipity to be lower for the augmented data (B), but comparable for the other versions of the system (A, C, and D). One possible explanation is that perceived serendipity decreases for the augmented data, but that perceptions increase once participants can identify the provenance of the topics and people recommended.

| A | B | C | D | All |
|---------|---------|---------|---------|---------|
| 59.81 | 50.97 | 62.59 | 60.38 | 57.95 |
| (30.58) | (32.54) | (28.95) | (26.33) | (30.34) |

Table 6.4: Mean (SD) of Perceived serendipity across the different levels of the system: A=Baseline, B=Augmented Data, C=Inspectable, D=Controllable. (0=low, 100=high)

| Non-personalized | Personalized | All |
|------------------|---------------|---------------|
| 57.69 (29.38) | 58.22 (31.39) | 57.95 (30.34) |

Table 6.5: Mean (SD) of Perceived serendipity for personalized and non-personalized conditions.

H2: Perceived familiarity Comparing between conditions we did not find a significant effect of interface condition on the perception of being able to find familiar tweets (Wilcoxon rank, $W = 1944.5$, $p\text{-value} = 0.07$). However, the low p value merited further investigation. Post-hoc tests were applied to investigate participants' ability to find familiar and relevant tweets for the non-personalized (Wilcoxon rank, $W = 542.5$, $p\text{-value} = 0.03$) and personalized (Wilcoxon rank, $W = 425$, $p\text{-value} = 0.76$) data. That is, we found an effect of condition in the non-personalized condition only. There was no significant difference w.r.t. perceived familiarity comparing the personalized compared to the non-personalized data, across types of system (Wilcoxon, $W = 1497.5$, $p\text{-value} = 0.47$).

Observing the means in Table 6.6 we see that the mean is lowest in for the fully controllable condition (D) for the non-personalized data. This suggest that with richer interaction users can (correctly) identify when tweets are not familiar and relevant to them. Interestingly, in the Inspectable condition (C) and for the non-personalized data, participants this effect does not occur.

| | Non-Personalized | Personalized | Both |
|----------|------------------|---------------|---------------|
| C | 59.96 (28.03) | 55.37 (27.69) | 57.67 (27.69) |
| D | 38.67 (34.40) | 52.07 (34.18) | 45.37 (34.66) |
| Combined | 48.75 (33.05) | 53.63 (31.04) | 51.19 (32.02) |

Table 6.6: Mean (SD) of content identified as relevant and familiar. (1=low, 100=high)

H3: Perceived transparency There was a significant difference between the versions of the system (A-D) w.r.t. the degree of perceived transparency (Kruskal-Wallis chi-

squared = 8.5456, $df = 3$, $p\text{-value} < 0.05$). The means in Table 6.7 show higher means for the more interactive and transparent conditions (pair-wise comparisons were not significant after correction was applied). We did not anticipate a difference in perceived transparency between the personalized and non-personalized conditions.

| A | B | C | D | All |
|--------|--------|--------|--------|--------|
| 4.45 | 5.12 | 5.41 | 5.45 | 5.09 |
| (2.05) | (1.82) | (1.52) | (1.37) | (1.76) |

Table 6.7: Mean (SD) of Perceived Transparency across the different levels of the system: A=Baseline, B=Augmented Data, C=Inspectable, D=Controllable. (1=low, 7=high)

H4: Perceived control There was a strong and significant difference between the versions of the system (A-D) w.r.t. the degree of perceived control (Kruskal-Wallis chi-squared = 13.562, $df = 3$, $p\text{-value} \ll 0.01$). Table 6.8 summarizes the means per condition, demonstrating a greater sense of control in the Inspectable (C) and Controllable (D) conditions. Post-hoc pairwise comparisons show a significant effect of conditions between the Baseline and both the Inspectable (C) and Controllable (D) conditions (Wilcoxon, $p < 0.01$, Bonferroni corrected).

We did not anticipate a difference in perceived control between the personalized and non-personalized conditions.

| A | B | C | D | All |
|--------|--------|--------|--------|--------|
| 3.97 | 4.38 | 4.88 | 5.03 | 4.49 |
| (1.67) | (1.88) | (1.63) | (1.29) | (1.71) |

Table 6.8: Mean (SD) of Perceived Control across the different levels of the system: A=Baseline, B=Augmented Data, C=Inspectable, D=Controllable. (1=low, 7=high)

H5: Content discovery Participants could not select topics or people in conditions A and B, so we compared conditions C and D only. This measure is the total sum of the number of hashtags and people selected/saved in both the direct and extended network. There was a trend toward greater content discovery in the Inspectable (C) condition compared to the Controllable (D) condition (Wilcoxon rank, $W = 1168.5$, $p\text{-value} = 0.07$), but the Inspectable condition also had a much larger standard deviation. There was also no significant difference for the degree of content discovery between the non-personalized and personalized conditions (Wilcoxon rank, $W = 6131.5$, $p\text{-value} = 0.84$), means are shown in Table 6.10.

| C | D | All |
|--------|--------|--------|
| 7.04 | 4.44 | 6.06 |
| (6.71) | (3.30) | (5.79) |

Table 6.9: Mean (SD) of content discovered across the different levels of the system: A=Baseline, B=Augmented Data, C=Inspectable, D=Controllable. (1=low, 7=high)

| Non-personalized | Personalized | All |
|------------------|--------------|-------------|
| 3.20 (4.80) | 2.94 (4.09) | 3.07 (4.45) |

Table 6.10: Mean (SD) of content discovered for personalized and non-personalized conditions. (1=low, 7=high)

H6: Correctness of mental model There were significant correlations between perceived serendipity and degree of content discovery in the personalized condition (Table 6.11). Participants could not select topics or people in conditions A and B, so we compared conditions C and D only. Post-hoc comparisons show that for the Inspectable interface and the non-personalized condition this correlation was negative and significant (Spearman, $p < 0.05$, $\rho = -0.47$, Bonferroni corrected).

| Comparison | p | rho |
|--------------|------|-------|
| Condition C | 0.16 | -0.19 |
| Condition D | 0.57 | 0.10 |
| Personalized | 0.02 | 0.15 |
| Non-Pers. | 0.16 | 0.08 |

Table 6.11: Correlations between perceived serendipity and degree of content discovery, Spearman rho.

H7: Perceived diversity There was no significant difference between the versions of the system (A-D) w.r.t. the degree of perceived diversity (Kruskal-Wallis chi-squared = 3.8267, $df = 3$, $p\text{-value} = 0.28$).

H8: Increased Interaction There were significant difference between versions of the system with regard to the number of double clicks for both personalized (Kruskal-Wallis chi-squared= 10.2, $df = 3$, $p\text{-value} < 0.05$) and non-personalized data (Kruskal-Wallis chi-squared= 16.7, $df = 3$, $p\text{-value} \ll 0.05$). We also found significant difference between the versions of the system w.r.t. the number of (single) clicks for the non-personalized data (Kruskal-Wallis chi-squared= 8.2, $df = 3$, $p\text{-value} < 0.05$).

Table 6.13 summarizes the number of interactions with the system. These numbers are computed as the total number of consecutive actions on the interface. As shown in Table 6.13, participants performed more clicks with the system in the Inspectable condition (C). It is interesting to note that users in Controllable condition (D) have shown the least degree of interaction in most event types. This partially contradicts our initial hypothesis (H8a). This might be an indicator which implies that users can easily get to relevant information with the minimum amount of interaction when rich controllability is provided. The result also contradicts the hypothesis (H8b) in both click and double click behaviors. Except for the Controllable condition (D), users had more clicks for the non-personalized data. The findings for mouseover and wheel action,

however confirm our initial hypothesis – there was more interaction in the personalized condition.

| condition | | Hashtag-1hop | | Hashtag-2hop | | User-2hop | |
|-----------|---|--------------|------|--------------|------|-----------|------|
| | | mean | std | mean | std | mean | std |
| Non-Pers. | C | 2.33 | 1.8 | 2.81 | 2.5 | 1.52 | 0.85 |
| | D | 2.06 | 1.82 | 2.06 | 2.22 | 1.23 | 0.76 |
| Pers | C | 2.07 | 2.07 | 1.7 | 1.61 | 2.59 | 3.1 |
| | D | 2.0 | 1.6 | 2.7 | 2.69 | 1.26 | 0.73 |

Table 6.12: Number of clicks across two versions of the system: C=Inspectable, D=Controllable. Hashtag-1hop is the number of topics that were clicked on in the participants' immediate network (1-hop), and Hashtag-2hop topics mentioned by their broader network (2-hop). Similarly, User-2hop denotes the number of users in the extended network the participant clicked on.

Summary We found that the both the Inspectable (C) and Controllable (D) versions of the system had significant impact on the degree of perceived control as well as transparency. However, with richer interaction (D) users can (correctly) identify when tweets are not familiar and relevant to them, while this does not seem to be the case for the Inspectable condition (C). In contrast, the Inspectable condition (C) showed a trend toward a greater degree of content discovery compared to the Controllable (D) condition, but this result was not significant due to a very large variance.

The Inspectable condition (C) also led to a poorer mental model for the non-personalized data: participant's perceptions of content discovery were negatively correlated with the actual degree of content discovery. I.e., participants may have underestimated how much content they were actually discovering. Across system versions (A-D), the personalized data had a weak, but positive and significant correlation for the mental model. Surprisingly, there was no effect of the degree of personalization (comparing personalized versus non-personalized data) on perceived serendipity, familiarity or the amount of content discovered when comparing across all four interface versions.

| condition | | # clicks | | | | # dbl-clicks | | | |
|-----------|---|-------------|------|------|-----|----------------|------|------|-----|
| | | sum | mean | std | usr | sum | mean | std | usr |
| Non-P. | A | 452 | 14.1 | 17.8 | 32 | 13 | 0.4 | 0.8 | 32 |
| | B | 501 | 15.7 | 24.1 | 32 | 17 | 0.5 | 2.7 | 32 |
| | C | 682 | 25.3 | 27.2 | 27 | 36 | 1.3 | 1.8 | 27 |
| | D | 288 | 9.3 | 12.6 | 31 | 10 | 0.3 | 1.0 | 31 |
| Pers. | A | 422 | 13.2 | 16.1 | 32 | 9 | 0.3 | 0.9 | 32 |
| | B | 418 | 13.1 | 14.9 | 32 | 4 | 0.1 | 0.3 | 32 |
| | C | 549 | 20.3 | 18.7 | 27 | 19 | 0.7 | 1.1 | 27 |
| | D | 374 | 12.1 | 9.9 | 31 | 11 | 0.4 | 1.3 | 31 |
| condition | | # mouseover | | | | # wheel action | | | |
| | | sum | mean | std | usr | sum | mean | std | usr |
| Non-P. | A | 1376 | 43.0 | 50.1 | 32 | 945 | 29.5 | 43.1 | 32 |
| | B | 1477 | 46.2 | 51.5 | 32 | 1052 | 32.9 | 45.5 | 32 |
| | C | 1622 | 60.1 | 64.9 | 27 | 987 | 36.6 | 45.9 | 27 |
| | D | 902 | 29.1 | 45.9 | 31 | 658 | 21.2 | 38.9 | 31 |
| Pers. | A | 1534 | 47.9 | 67.4 | 32 | 1135 | 35.5 | 60.3 | 32 |
| | B | 1517 | 47.4 | 64.3 | 32 | 1140 | 35.6 | 59.9 | 32 |
| | C | 1701 | 63.0 | 92.2 | 27 | 1195 | 44.3 | 81.1 | 27 |
| | D | 1187 | 38.3 | 63.6 | 31 | 840 | 27.1 | 58.8 | 31 |

Table 6.13: Number of user interactions across the different versions of the system: A=Baseline, B=Augmented Data, C=Inspectable, D=Controllable.

6.4.6 Discussion and Future Work

In this study we designed a novel interactive tool, called HopTopics, for social content discovery. The HopTopics system was developed with an improved interface and interaction design, based on a formative lab-based user study. In the main online user study (N=122) interface and interaction design of the system has been evaluated in both quantitative and qualitative ways using a layered evaluation approach.

Specifically, in the main online user study, we found that the Inspectable (C) and Controllable (D) versions of the system had a significant impact on the degree of perceived control and transparency. In this experiment we compared the results between users. This meant that the results may have been affected by individual differences and difference in the Twitter content for these users. The high standard deviations suggest this is the case, and we are therefore planning a more targeted study to compare the Inspectable and Controllable interfaces.

Another thing we plan to investigate is the effect of the global use of topics mentioned in the network on novel and relevant content discovery. That is, the trending, or most popular uses of a hashtag that has been selected by a user. We acknowledge that a limitation of evaluations with real-time and real-world data is the speed for loading content from API service providers in real time. This is a key aspect that we are considering for future versions and evaluations of the system.

Acknowledgments This research has been carried out within the project “Scrutable Autonomous Systems” (SAsSY), funded by the UK Engineering and Physical Sciences Research Council, grant ref. EP/J012084/1. This work was also partially supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053; The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied,

of ARL, NSF, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

6.5 Summary

In this chapter, we have discussed the importance of effective visual communication interfaces in information search and knowledge discovery in the Social Web and the role of visualization to achieve our goal. An interactive visualization framework or rich user interfaces can be a valuable means to gain insight or knowledge from reliable information. This can be maximized when those systems or interfaces are amalgamated with intelligent algorithms and reliability models. Such a powerful visual analytics tool can foster data-informed decision-making practices in both academia and industries.

Chapter 7

Conclusions and Future Work

7.1 Introduction

In this dissertation, we began with our core motivation and problem statements of the research, followed by a range of definitions of information quality metrics including information reliability and related work in computer science and other disciplines. Reliable information and knowledge have been keys to success in many fields and are recognized as indispensable resources that people seek for their various contextual needs. However, every user on the Web is likely to encounter low-quality information. Relying on such information can result in an irreversible and critical problems, especially when it comes to follow-up decision-making tasks. Moreover, non-expert users often experience particular difficulty with gauging information quality during their information-seeking tasks, and as the amount of available information grows, the need for an automated algorithm that retrieves high-quality, reliable information has been exponentially increased. As the title of this dissertation implies, our studies aim to find the best models for identifying such reliable information, especially for tasks that involves information-based decision-making. Additionally, we study effective communication that enables users and data analysts to

explore social web data with ease.

7.2 Reliable Information on the Social Web

In Chapter 1, we presented the motivation and explained why we focus on the Social Web as the backdrop of our research. Afterwards, we proposed a definition of information reliability in order to understand the complex spectrum of information reliability attributes.

In this thesis, we assume that information reliability is a superordinate concept which encompasses various information attributes. Since multiple information metrics construct a multidimensional space that is comprised of numerous features, it is extremely challenging to unveil underlying or hidden patterns and identify the most influential set of patterns or features among them. We have conducted a number of experiments to find the best features and modeled important information attributes on the social network that are portable across contexts and different platforms, such as credibility, competence, and influence.

7.3 Objectives and Contributions

As we highlighted throughout this dissertation, the main objectives of this thesis are:

1. Novel intelligent algorithms to identify high-quality information in large and diverse datasets
2. Visualization of data quality among large and/or heterogeneous datasets, for decision making
3. Intelligent user interfaces for information retrieval and content recommendation

In this thesis, we have modeled different attributes of reliable information in contexts that involve decision-making tasks. Furthermore, we have developed a range of algorithms and novel interfaces to identify, recommend and visualize reliable information. The list below summarizes the main contributions of this thesis.

- We provide a literature survey of information quality frameworks and a wide range of methods that filter, detect and predict information reliability.
- We propose a modified framework of high-quality (reliable) information based on an established framework of information quality.
- We provide different models of either intrinsic or perceived social web information reliability across tasks and contexts.
- To validate information reliability models, we provide effective methods to find ground truth of information reliability.
- In order to effectively communicate with extracted reliable information, we design, implement, and evaluate several visualization techniques and frameworks.

These contributions can be grouped into two larger categories of research focus, which we will summarize in the following two subsections.

7.4 Identifying Reliable Information with Intelligent Algorithms and Robust Ground Truth

The first main area of contributions from this thesis is formed by the computational algorithms and statistical models we developed that automatically detect reliable information, including credible, trustworthy information, and identify influential users a topic of interest.

Similar to all the previous work many scientists have proposed and implemented methods to automatically detect reliable information or objects, our work presented in this thesis faces a major difficulty inherent in the field: *subjectivity*. There is still no globally accepted unified definition of information quality (credibility, trustworthiness, interestingness, etc.) Accordingly, many challenges in this field are just beginning to emerge. The subjective aspects of reliable information characteristics made human factor analysis a core focus in the evaluation of the underlying intelligent algorithms and models that we developed. Typically, modeling, classification, and recommendation algorithms have been evaluated in terms of speed, accuracy, or other automated metrics. Understanding how the human in the loop changes the information search and discovery process is a central motivation for us, and our studies on this challenging topic have been illustrated throughout the thesis. For example, we studied reliable ground-truth data in assessing intelligent algorithms (covered in Chapter 5).

Another interesting problem covered in this thesis is the study on how to identify influential users in social networks (Section 4.6). In viral marketing, timely identification of influential users is a key to success. We developed a computational composite algorithm that recommends top influential users in Twitter to corporate marketers. To evaluate the proposed algorithm, we also developed a web-based front-end to a cluster computing engine (Apache Spark) for large-scale data processing in real-time. This system has provided orders of magnitude performance gain in computing speed during the evaluation in multi-core workstation setup. Furthermore, the system received positive feedback from corporate marketing experts. In the future, we would like to extend this model by applying modern deep neural network algorithms (deep learning), such as non-linear convolutional neural networks (CNN).

7.5 Information Reliability and the End User

In recent years, researchers have studied scalability and quality issues caused by the information overload problem and proposed information search and filtering algorithms in response. However, such intelligent algorithms will not live up to their full potential if the resulting information is not presented in a comprehensible way. In this vein, our second main contribution is in the area of visualization and intelligent interactive user interfaces in order to tackle the following research questions.

- How to develop scalable visualizations that allow users to easily comprehend the high dimensional structure of socially connected data?
- What are effective ways to visualize social data streams in real time?
- How can visualization and user interfaces support decision-making and recommender systems?

Volume, High Dimensionality and Resource User-generated contents on social networks have unique characteristics among them: the four-V challenges of big data—volume, variety, velocity and veracity. These properties make it difficult to parse information and to gain useful knowledge or intuition. The contributions of this thesis include implementations and evaluations of interactive visualization frameworks (Chapter 6). Many visualization frameworks have been developed to support post-hoc data analysis tasks. Our work on real-time visualization, called the *TweetProbe*¹, builds on different methodologies, such as a time-window based approach at the back-end and animated transition techniques for the front-end, in conjunction with other novel visual components. Our recent work on real-time visualization builds on state-of-the-art cluster

¹TweetProbe: A Real-Time Microblog Stream Visualization Framework [168]

computing techniques and asynchronous event handling for large-scale data processing and analytics.

Visual Interfaces for Decision Making and Recommender Systems Among the variety of tasks for which visualization is used, in decision-making, information reliability models play a key role. Users make important decisions based on reliable information. To support this process, we have studied how to identify and represent reliable information through effective visualization and user interfaces, allowing users to easily communicate with the data. Since reliable information can vary and be accessed differently across users or contexts, we evaluated the impact of visualization and user interfaces under different conditions. Our studies on this topic provided general insights on intelligent user interface design for information filtering and recommender systems to maximize user experience and represent better quality of information.

7.6 Future Work

This thesis has discussed from several reliability models to a range of intelligent algorithms and interactive user interfaces. We also have shown or proposed possible applications to which these intelligent algorithms and interfaces can be applied. However, at the same time, our research has revealed additional open questions that we need to answer in the future. The subjectivity of human's perception that inevitably involves uncertainty and variance across people and contexts remains not answered. Perhaps, this might be the core challenging question that all scientists, including researchers in the field of artificial intelligence, want to uncover. We still believe that there must be a certain set of patterns underlying human behaviors and we can find and understand them.

Since our strong belief is that the opportunity of big data can shed light on this convoluted question, our future work will focus on understanding and predicting human behavior in both mass (mass behavior) and micro-level (personalization) from the cross-disciplinary perspective. More specifically, we aim to broaden the possibilities of big data with intelligent algorithms and realize these potentials into original applications through multimodal interaction. To achieve this goal, we plan to develop interdisciplinary studies that extends the topics covered in this thesis.

Bibliography

- [1] J. M. Pujol, V. Erramilli, G. Siganos, X. Yang, N. Laoutaris, P. Chhabra, and P. Rodriguez, *The little engine(s) that could: Scaling online social networks*, *SIGCOMM Comput. Commun. Rev.* **40** (Aug., 2010) 375–386.
- [2] M. C. Pham, Y. Cao, R. Klamma, and M. Jarke, *A clustering approach for collaborative filtering recommendation using social network analysis.*, *J. UCS* **17** (2011), no. 4 583–604.
- [3] C. Wang, W. Chen, and Y. Wang, *Scalable influence maximization for independent cascade model in large-scale social networks*, *Data Mining and Knowledge Discovery* **25** (2012), no. 3 545–576.
- [4] X. Jin, C. Lin, J. Luo, and J. Han, *A data mining-based spam detection system for social media networks*, *Proceedings of the VLDB Endowment* **4** (2011), no. 12.
- [5] S. E. Robertson and K. S. Jones, *Relevance weighting of search terms*, *Journal of the American Society for Information science* **27** (1976), no. 3 129–146.
- [6] S. Brin and L. Page, *The anatomy of a large-scale hypertextual web search engine*, *Computer networks and ISDN systems* **30** (1998), no. 1 107–117.
- [7] T. H. Haveliwala, *Topic-sensitive pagerank*, in *Proceedings of the 11th International Conference on World Wide Web*, WWW '02, (New York, NY, USA), pp. 517–526, ACM, 2002.
- [8] B. Shneiderman, J. Preece, and P. Pirolli, *Realizing the value of social media requires innovative computing research*, *Communications of the ACM* **54** (2011), no. 9 34–37.
- [9] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, *Evaluating collaborative filtering recommender systems*, *ACM Transactions on Information Systems (TOIS)* **22** (2004), no. 1 5–53.
- [10] M. Gomez-Rodriguez, K. P. Gummadi, and B. Schölkopf, *Quantifying information overload in social media and its impact on social contagions*, *arXiv preprint arXiv:1403.6838* (2014).

- [11] C. W. Anderson, *Rebuilding the News: Metropolitan Journalism in the Digital Age*. Temple University, Philadelphia, PA, USA, 2013.
- [12] L. Willnat and D. H. Weaver, *The american journalist in the digital age*, tech. rep., School of Journalism, Indiana University, 2014.
- [13] R. S. Taylor, *Value-added processes in information systems*. Greenwood Publishing Group, 1986.
- [14] D. M. S. Richard Y. Wang, *Beyond accuracy: What data quality means to data consumers*, *Journal of Management Information Systems* **12** (1996), no. 4 5–33.
- [15] R. Y. Wang, *A product perspective on total data quality management*, *Commun. ACM* **41** (Feb., 1998) 58–65.
- [16] S. Y. Rieh, *Judgment of information quality and cognitive authority in the web*, *Journal of the American Society for Information Science and Technology* **53** (2002), no. 2 145–161.
- [17] M. Parker, V. Moleshe, R. D. la Harpe, and G. Wills, *An evaluation of information quality frameworks for the world wide web*, in *8th Annual Conference on WWW Applications*, 2006. Event Dates: 6-8th September 2006.
- [18] N. Agarwal and Y. Yiliyasi, *Information quality challenges in social media*, in *International Conference on Information Quality (ICIQ 2010)*, Little Rock, Arkansas, 2010.
- [19] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, *Finding high-quality content in social media*, in *the international conference*, (New York, New York, USA), p. 183, ACM Press, 2008.
- [20] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, *Tweeting is believing?: understanding microblog credibility perceptions*, in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, CSCW '12*, (New York, NY, USA), pp. 441–450, ACM, 2012.
- [21] A. Hermida, *Twittering the news*, *Journalism Practice* **4** (2010), no. 3 297–308, [<http://dx.doi.org/10.1080/17512781003640703>].
- [22] M. Mendoza, B. Poblete, and C. Castillo, *Twitter under crisis: can we trust what we rt?*, in *Proceedings of the First Workshop on Social Media Analytics, SOMA '10*, (New York, NY, USA), pp. 71–79, ACM, 2010.
- [23] R. Pierce, *Research methods in politics*. Sage, 2008.

- [24] S. A. Adams, *Revisiting the online health information reliability debate in the wake of “web 2.0”: an inter-disciplinary literature and website review*, *International journal of medical informatics* **79** (2010), no. 6 391–400.
- [25] S. Adams, *Under construction: Reviewing and producing information reliability on the web*. Erasmus Universiteit Rotterdam, 2006.
- [26] S. Y. Rieh and D. R. Danielson, *Credibility: A multidisciplinary framework*, *Annual Review of Information Science and Technology* **41** (2007), no. 1 307–364.
- [27] A. J. Flanagin and M. J. Metzger, *The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information*, *New Media & Society* **9** (2007), no. 2 319–342.
- [28] M. J. Metzger, A. J. Flanagin, and R. B. Medders, *Social and heuristic approaches to credibility evaluation online*, *Journal of Communication* **60** (2010), no. 3 413–439.
- [29] C. Castillo, M. Mendoza, and B. Poblete, *Information credibility on twitter.*, in *WWW* (S. Srinivasan, K. Ramamritham, A. Kumar, M. P. Ravindra, E. Bertino, and R. Kumar, eds.), pp. 675–684, ACM, 2011.
- [30] B. Kang, J. O’Donovan, and T. Höllerer, *Modeling topic specific credibility on twitter*, in *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, IUI ’12, (New York, NY, USA), pp. 179–188, ACM, 2012.
- [31] G. Eysenbach and C. Köhler, *How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews*, *Bmj* **324** (Mar., 2002) 573–577.
- [32] K. S. Freeman and J. H. Spyridakis, *An examination of factors that affect the credibility of online health information*, *Technical Communication* **51** (2004), no. 2 239–263.
- [33] M. J. Metzger, *Making sense of credibility on the web: Models for evaluating online information and recommendations for future research*, *Journal of the American Society for Information Science and Technology* **58** (2007), no. 13 2078–2091.
- [34] B. Fogg and H. Tseng, *The elements of computer credibility*, in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 80–87, ACM, 1999.
- [35] B. R. Bates, S. Romina, R. Ahmed, and D. Hopson, *The effect of source credibility on consumer’s perceptions of the quality of health information on the internet*, *Medical Informatics and the Internet in Medicine* **31** (2006), no. 1 45–52.

- [36] J. Golbeck and D. L. Hansen, *Computing political preference among twitter followers*, in *CHI*, pp. 1105–1108, 2011.
- [37] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, *Predicting elections with twitter: What 140 characters reveal about political sentiment.*, *ICWSM* **10** (2010) 178–185.
- [38] W. Zhang, T. J. Johnson, T. Seltzer, and S. L. Bichard, *The revolution will be networked: The influence of social networking sites on political attitudes and behavior*, *Social Science Computer Review* (2009).
- [39] A. Go, R. Bhayani, and L. Huang, *Twitter sentiment classification using distant supervision*, *CS224N Project Report, Stanford* (2009) 1–12.
- [40] Z. Ding, Y. Jia, B. Zhou, and Y. Han, *Mining tribe-leaders based on the frequent pattern of propagation*, in *Web Technologies and Applications*, pp. 143–153. Springer, 2012.
- [41] C. Edwards, P. R. Spence, C. J. Gentile, A. Edwards, and A. Edwards, *How much klout do you have a test of system generated cues on source credibility*, *Computers in Human Behavior* **29** (2013), no. 5 A12–A16.
- [42] N. Diakopoulos and A. Zubiaga, *Newsworthiness and network gatekeeping on twitter: The role of social deviance*, in *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014.*, 2014.
- [43] B. Kang, T. Höllerer, and J. O’Donovan, *The full story: Automatic detection of unique news content in microblogs*, in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, ASONAM ’15, (New York, NY, USA)*, pp. 1192–1199, ACM, 2015.
- [44] L. Geng and H. J. Hamilton, *Interestingness measures for data mining: A survey*, *ACM Comput. Surv.* **38** (Sept., 2006).
- [45] G. Piatetsky-Shapiro and C. J. Matheus, *The interestingness of deviations*, *Proc AAAI* (1994).
- [46] B. Liu, Y. Ma, and P. S. Yu, *Discovering unexpected information from your competitors’ web sites*, in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’01, (New York, NY, USA)*, pp. 144–153, ACM, 2001.
- [47] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi, *Bad news travel fast: A content-based analysis of interestingness on twitter*, in *Proceedings of the 3rd International Web Science Conference, WebSci ’11, (New York, NY, USA)*, pp. 8:1–8:7, ACM, 2011.

- [48] J. O'Donovan, B. Kang, and T. Höllerer, *Competence modeling in twitter: Mapping theory to practice*, .
- [49] S. Sikdar, B. Kang, J. O'Donovan, T. Höllerer, and S. Adalı, *Understanding information credibility on twitter*, in *2013 International Conference on Social Computing (SocialCom)*, pp. 19–24, IEEE, 2013.
- [50] S. K. Sikdar, B. Kang, J. O'Donovan, T. Höllerer, and S. Adalı, *Cutting through the noise: Defining ground truth in information credibility on twitter*, *ASE HUMAN* **2** (2013), no. 3 pp–151.
- [51] B. A. Nardi, D. J. Schiano, M. Gumbrecht, and L. Swartz, *Why we blog*, *Commun. ACM* **47** (Dec., 2004) 41–46.
- [52] C.-L. Hsu and J. C.-C. Lin, *Acceptance of blog usage: The roles of technology acceptance, social influence and knowledge sharing motivation*, *Information & Management* **45** (2008), no. 1 65–74.
- [53] T. J. Johnson and B. K. Kaye, *Wag the blog: How reliance on traditional media and the internet influence credibility perceptions of weblogs among blog users*, *Journalism & Mass Communication Quarterly* **81** (2004), no. 3 622–642.
- [54] B. Krishnamurthy, P. Gill, and M. Arlitt, *A few chirps about twitter*, in *Proceedings of the first workshop on Online social networks*, pp. 19–24, ACM, 2008.
- [55] H. Kwak, C. Lee, H. Park, and S. Moon, *What is twitter, a social network or a news media?*, in *Proceedings of the 19th international conference on World wide web*, WWW '10, (New York, NY, USA), pp. 591–600, ACM, 2010.
- [56] R. Campos, G. Dias, A. M. Jorge, and A. Jatowt, *Survey of temporal information retrieval and related applications*, *ACM Computing Surveys (CSUR)* **47** (2014), no. 2 15.
- [57] S. Bakhshi, D. A. Shamma, and E. Gilbert, *Faces engage us: photos with faces attract more likes and comments on instagram*, in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pp. 965–974, ACM, 2014.
- [58] L. Manikonda, Y. Hu, and S. Kambhampati, *Analyzing user activities, demographics, social network structure and user-generated content on instagram*, *arXiv preprint arXiv:1410.8099* (2014).
- [59] J. E. Rowley, *The wisdom hierarchy: representations of the dikw hierarchy*, *Journal of Information Science* (2007).

- [60] C. Zins, *Conceptual approaches for defining data, information, and knowledge*, *Journal of the American Society for Information Science and Technology* **58** (2007), no. 4 479–493.
- [61] M. Zeleny, *Management support systems: towards integrated knowledge management*, *Human systems management* **7** (1987), no. 1 59–70.
- [62] M. L. Silberman, *The handbook of experiential learning*. John Wiley & Sons, 2007.
- [63] C. W. Choo, B. Detlor, and D. Turnbull, *Web work: Information seeking and knowledge work on the world wide web (information science and knowledge management) aut*, .
- [64] R. L. Ackoff, *From data to wisdom*, *Journal of applied systems analysis* **16** (1989), no. 1 3–9.
- [65] G. Bellinger, D. Castro, and A. Mills, *Data, information, knowledge, and wisdom*, URL: <http://www.systems-thinking.org/dikw/dikw.htm> (2004) 47.
- [66] A. Liew, *Understanding data, information, knowledge and their inter-relationships*, *Journal of Knowledge Management Practice* **8** (2007), no. 2.
- [67] T. H. Davenport and L. Prusak, *Working knowledge: How organizations manage what they know*. Harvard Business Press, 1998.
- [68] D. P. Wallace, *Knowledge management: Historical and cross-disciplinary themes*. Libraries unlimited, 2007.
- [69] U. Lee, Z. Liu, and J. Cho, *Automatic identification of user goals in web search*, in *Proceedings of the 14th international conference on World Wide Web*, pp. 391–400, ACM, 2005.
- [70] D. M. Strong, Y. W. Lee, and R. Y. Wang, *Data quality in context*, *Communications of the ACM* **40** (1997), no. 5 103–110.
- [71] P. Sondhi, V. V. Vydiswaran, and C. Zhai, *Reliability prediction of webpages in the medical domain*, in *Advances in Information Retrieval*, pp. 219–231. Springer, 2012.
- [72] P. Stavri, D. Freeman, and C. Burroughs, *Perception of quality and trustworthiness of internet resources by personal health information seekers*, in *AMIA Annual Symposium Proceedings*, 2003.
- [73] S. Xu, Y. Zhu, H. Jiang, and F. C. Lau, *A user-oriented webpage ranking algorithm based on user attention time.*, in *AAAI*, vol. 8, pp. 1255–1260, 2008.

- [74] G. Dupret, V. Murdock, and B. Piwowarski, *Web search engine evaluation using clickthrough data and a user model*, in *WWW2007 workshop Query Log Analysis: Social and Technological Challenges*, 2007.
- [75] T. Joachims, *Optimizing search engines using clickthrough data*, in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 133–142, ACM, 2002.
- [76] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, *Accurately interpreting clickthrough data as implicit feedback*, in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 154–161, ACM, 2005.
- [77] J.-T. Sun, H.-J. Zeng, H. Liu, Y. Lu, and Z. Chen, *Cubesvd: a novel approach to personalized web search*, in *Proceedings of the 14th international conference on World Wide Web*, pp. 382–390, ACM, 2005.
- [78] K. E. Schmidt, Y. Liu, and S. Sridharan, *Webpage aesthetics, performance and usability: Design variables and their effects*, *Ergonomics* **52** (2009), no. 6 631–643.
- [79] L. J. Kensicki, *Building credibility for non-profit organizations through webpage interface design*, .
- [80] E. F. Can, H. Oktay, and R. Manmatha, *Predicting retweet count using visual cues*, in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, CIKM '13*, (New York, NY, USA), pp. 1481–1484, ACM, 2013.
- [81] S. Kumar, F. Morstatter, R. Zafarani, and H. Liu, *Whom should i follow?: identifying relevant users during crises*, in *Proceedings of the 24th ACM conference on Hypertext and social media*, pp. 139–147, ACM, 2013.
- [82] A. Zubiaga, H. Ji, and K. Knight, *Curating and contextualizing twitter stories to assist with social newsgathering*, in *Proceedings of the 2013 International Conference on Intelligent User Interfaces, IUI '13*, (New York, NY, USA), pp. 213–224, ACM, 2013.
- [83] J. A. Golbeck, *Computing and Applying Trust in Web-based Social Networks*. dissertation, University of Maryland (College Park, Md.), 2005.
- [84] B. Fogg, *Prominence-interpretation theory: Explaining how people assess credibility online*, in *CHI'03 extended abstracts on human factors in computing systems*, pp. 722–723, ACM, 2003.

- [85] C. Wagner, V. Liao, P. Pirolli, L. Nelson, and M. Strohmaier, *It's not in their tweets: Modeling topical expertise of twitter users*, in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*, pp. 91–100, IEEE, 2012.
- [86] S. Kioussis, *Public trust or mistrust? perceptions of media credibility in the information age*, *Mass Communication & Society* **4** (2001), no. 4 381–403.
- [87] B. J. Fogg, C. Soohoo, D. R. Danielson, L. Marable, J. Stanford, and E. R. Tauber, *How do users evaluate the credibility of web sites?: A study with over 2,500 participants*, in *Proceedings of the 2003 Conference on Designing for User Experiences, DUX '03*, (New York, NY, USA), pp. 1–15, ACM, 2003.
- [88] J. Yang, S. Counts, M. R. Morris, and A. Hoff, *Microblog credibility perceptions: comparing the usa and china*, in *Proceedings of the 2013 conference on Computer supported cooperative work*, pp. 575–586, ACM, 2013.
- [89] J. Mahmud, M. X. Zhou, N. Megiddo, J. Nichols, and C. Drews, *Recommending targeted strangers from whom to solicit information on social media*, in *Proceedings of the 2013 International Conference on Intelligent User Interfaces, IUI '13*, (New York, NY, USA), pp. 37–48, ACM, 2013.
- [90] S. S. Sundar, *Effect of source attribution on perception of online news stories*, *Journalism & Mass Communication Quarterly* **75** (1998), no. 1 55–68.
- [91] S. E. Dreyfus and H. L. Dreyfus, *A five-stage model of the mental activities involved in directed skill acquisition*, tech. rep., DTIC Document, 1980.
- [92] D. Pool and M. Kochen, *Contacts and influence*, *Social Networks* **1** (1978), no. 1 5–51.
- [93] S. Milgram, *The small world problem*, *Psychology Today* **1** (May, 1967) 61–67.
- [94] K. McNally, M. P. O'Mahony, B. Smyth, M. Coyle, and P. Briggs, *Towards a reputation-based model of social web search*, in *Proceedings of the 15th international conference on Intelligent user interfaces, IUI '10*, (New York, NY, USA), pp. 179–188, ACM, 2010.
- [95] J. Golbeck, *Computing with Social Trust*. Springer Publishing Company, Incorporated, 2010.
- [96] T. G. F. W. Haifeng Zhao, William Kallander, *Read what you trust: An open wiki model enhanced by social context*, in *Proceedings of the third IEEE International Conference on Social Computing (SocialCom)*, 2011.
- [97] D. Houser and J. Wooders, *Reputation in auctions: Theory, and evidence from ebay*, *Journal of Economics and Management Strategy* **15** (2006), no. 2 353–369.

- [98] P. Resnick and R. Zeckhauser, *Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system.*, *The Economics of the Internet and E-Commerce. Volume 11 of Advances in Applied Microeconomics*. (December, 2002).
- [99] J. O'Donovan, B. Smyth, V. Evrim, and D. McLeod, *Extracting and visualizing trust relationships from online auction feedback comments*, in *IJCAI*, pp. 2826–2831, 2007.
- [100] M. E. K. T. Jennifer Golbeck, Cristina Robles, *Predicting personality from twitter*, in *Proceedings of the third IEEE International Conference on Social Computing (SocialCom)*, 2011.
- [101] S. Adalı, R. Escriva, M. K. Goldberg, M. Hayvanovych, M. Magdon-Ismael, B. K. Szymanski, W. A. Wallace, and G. T. Williams, *Measuring behavioral trust in social networks*, in *ISI*, pp. 150–152, 2010.
- [102] Y. Suzuki, *A credibility assessment for message streams on microblogs*, in *Proceedings of the 2010 International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 3PGCIC '10*, (Washington, DC, USA), pp. 527–530, IEEE Computer Society, 2010.
- [103] K. Canini, B. Suh, and P. Pirolli, *Finding credible information sources in social networks based on content and social structure*, in *2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing (SocialCom)*, 2011.
- [104] J. O'Donovan and B. Smyth, *Trust in recommender systems*, in *IUI '05: Proceedings of the 10th international conference on Intelligent user interfaces*, pp. 167–174, ACM Press, 2005.
- [105] R. Burke, B. Mobasher, R. Zabicki, and R. Bhaumik, *Identifying attack models for secure recommendation*, in *Beyond Personalisation Workshop at the International Conference on Intelligent User Interfaces*, (San Deigo, USA.), pp. 347–361, ACM Press, 2005.
- [106] B. Mobasher, R. Burke, R. Bhaumik, and C. Williams, *Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness*, *ACM Trans. Inter. Tech.* **7** (2007), no. 4 23.
- [107] P. Victor, C. Cornelis, M. D. Cock, and E. Herrera-Viedma, *Practical aggregation operators for gradual trust and distrust*, *Fuzzy Sets and Systems* **184** (2011), no. 1 126–147.

- [108] T. DuBois, J. Golbeck, and A. Srinivasan, *Predicting trust and distrust in social networks*, in *Proceedings of the third IEEE International Conference on Social Computing (SocialCom)*, 2011.
- [109] F. Al Zamal, W. Liu, and D. Ruths, *Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors.*, in *ICWSM*, 2012.
- [110] J. O'Donovan, B. Kang, G. Meyer, T. Höllerer, and S. Adalı, *Credibility in context: An analysis of feature distributions in twitter*, in *2012 International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2012 International Conference on Social Computing (SocialCom)*, pp. 293–301, IEEE, 2012.
- [111] R. Thomson, N. Ito, H. Suda, F. Lin, Y. Liu, R. Hayasaka, R. Isochi, and Z. Wang, *Trusting tweets: The fukushima disaster and information source credibility on twitter*, in *Proceedings of the 9th International ISCRAM Conference*, 2012.
- [112] S. C. Herring and J. C. Paolillo, *Gender and genre variation in weblogs*, *Journal of Sociolinguistics* **10** (2006), no. 4 439–459.
- [113] S. Singh, *A pilot study on gender differences in conversational speech on lexical richness measures*, *Literary and Linguistic Computing* **16** (2001), no. 3 251–264.
- [114] J. D. Burger and J. C. Henderson, *An exploration of observable features related to blogger age.*, in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 15–20, 2006.
- [115] N. Garera and D. Yarowsky, *Modeling latent biographic attributes in conversational genres*, in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pp. 710–718, Association for Computational Linguistics, 2009.
- [116] R. Jones, R. Kumar, B. Pang, and A. Tomkins, *I know what you did last summer: query logs and user privacy*, in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 909–914, ACM, 2007.
- [117] I. Weber and C. Castillo, *The demographics of web search*, in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 523–530, ACM, 2010.
- [118] J. Otterbacher, *Inferring gender of movie reviewers: exploiting writing style, content and metadata*, in *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 369–378, ACM, 2010.

- [119] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, *Classifying latent user attributes in twitter*, in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pp. 37–44, ACM, 2010.
- [120] D. Rao, M. J. Paul, C. Fink, D. Yarowsky, T. Oates, and G. Coppersmith, *Hierarchical bayesian models for latent attribute detection in social media.*, in *ICWSM*, 2011.
- [121] Z. Cheng, J. Caverlee, and K. Lee, *You are where you tweet: a content-based approach to geo-locating twitter users*, in *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 759–768, ACM, 2010.
- [122] C. Fink, C. D. Piatko, J. Mayfield, T. Finin, and J. Martineau, *Geolocating blogs from their textual content.*, in *AAAI Spring Symposium: Social Semantic Web: Where Web 2.0 Meets Web 3.0*, pp. 25–26, 2009.
- [123] M. Thomas, B. Pang, and L. Lee, *Get out the vote: Determining support or opposition from congressional floor-debate transcripts*, in *Proceedings of the 2006 conference on empirical methods in natural language processing*, pp. 327–335, Association for Computational Linguistics, 2006.
- [124] M. Pennacchiotti and A.-M. Popescu, *A machine learning approach to twitter user classification.*, in *ICWSM*, 2011.
- [125] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, *Short text classification in twitter to improve information filtering*, in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 841–842, ACM, 2010.
- [126] R. B. Cialdini and M. R. Trost, *Social influence: Social norms, conformity and compliance.*, McGraw-Hill (1998).
- [127] P. F. Lazarsfeld and E. Katz, *Personal influence: the part played by people in the flow of mass communications*, Glencoe, Illinois (1955).
- [128] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, *Measuring user influence in twitter: The million follower fallacy.*, *ICWSM* **10** (2010) 10–17.
- [129] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, *Everyone’s an influencer: Quantifying influence on twitter*, in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM ’11*, (New York, NY, USA), pp. 65–74, ACM, 2011.

- [130] H. Bao and E. Y. Chang, *Adheat: An influence-based diffusion model for propagating hints to match ads*, in *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, (New York, NY, USA), pp. 71–80, ACM, 2010.
- [131] S. Singh, N. Mishra, and S. Sharma, *Survey of various techniques for determining influential users in social networks*, in *Emerging Trends in Computing, Communication and Nanotechnology (ICE-CCN), 2013 International Conference on*, pp. 398–403, IEEE, 2013.
- [132] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum, *Robust disambiguation of named entities in text*, in *EMNLP*, pp. 782–792, 2011.
- [133] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti, *Collective annotation of wikipedia entities in web text*, in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09*, (New York, NY, USA), pp. 457–466, ACM, 2009.
- [134] R. Mihalcea and A. Csomai, *Wikify!: Linking documents to encyclopedic knowledge*, in *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM '07*, (New York, NY, USA), pp. 233–242, ACM, 2007.
- [135] D. Paranjpe, *Learning document aboutness from implicit user feedback and document structure*, in *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09*, (New York, NY, USA), pp. 365–374, ACM, 2009.
- [136] D. Milne and I. H. Witten, *Learning to link with wikipedia*, in *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08*, (New York, NY, USA), pp. 509–518, ACM, 2008.
- [137] K. Chakrabarti, V. Ganti, J. Han, and D. Xin, *Ranking objects based on relationships*, in *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, SIGMOD '06*, (New York, NY, USA), pp. 371–382, ACM, 2006.
- [138] T. Cheng, X. Yan, and K. C.-C. Chang, *Entityrank: Searching entities directly and holistically*, in *Proceedings of the 33rd International Conference on Very Large Data Bases, VLDB '07*, pp. 387–398, VLDB Endowment, 2007.
- [139] N. Craswell and M. Szummer, *Random walks on the click graph*, in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07*, (New York, NY, USA), pp. 239–246, ACM, 2007.

- [140] G. Jeh and J. Widom, *Scaling personalized web search*, in *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, (New York, NY, USA), pp. 271–279, ACM, 2003.
- [141] H. Bota, K. Zhou, J. M. Jose, and M. Lalmas, *Composite retrieval of heterogeneous web search*, in *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, (New York, NY, USA), pp. 119–130, ACM, 2014.
- [142] P. J. Shoemaker, *News and newsworthiness: A commentary*, *Communications* **31** (2006), no. 1 105–111.
- [143] P. J. Shoemaker and A. A. Cohen, *News around the world: Practitioners, content and the public*, 2006.
- [144] P. J. Shoemaker, D. Martin Eichholz, E. Kim, and B. Wrigley, *Individual and Routine Forces in Gatekeeping*, *Journalism & Mass Communication Quarterly* **78** (June, 2001) 233–246.
- [145] J. Weng, E.-P. Lim, Q. He, and C.-K. Leung, *What do people want in microblogs? measuring interestingness of hashtags in twitter*, in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pp. 1121–1126, Dec, 2010.
- [146] V. Milicic, G. Rizzo, J. L. Redondo Garcia, R. Troncy, and T. Steiner, *Live topic generation from event streams*, in *Proceedings of the 22nd international conference on World Wide Web companion*, WWW '13 Companion, (Republic and Canton of Geneva, Switzerland), pp. 285–288, International World Wide Web Conferences Steering Committee, 2013.
- [147] M. Mathioudakis and N. Koudas, *Twittermonitor: trend detection over the twitter stream*, in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, SIGMOD '10, (New York, NY, USA), pp. 1155–1158, ACM, 2010.
- [148] S. Garcia Esparza, M. P. O'Mahony, and B. Smyth, *Catstream: categorising tweets for user profiling and stream filtering*, in *Proceedings of the 2013 international conference on Intelligent user interfaces*, IUI '13, (New York, NY, USA), pp. 25–36, ACM, 2013.
- [149] I. Guy, T. Steier, M. Barnea, I. Ronen, and T. Daniel, *Swimming against the streamz: search and analytics over the enterprise activity stream*, in *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, (New York, NY, USA), pp. 1587–1591, ACM, 2012.
- [150] Y. Zhao, F. Zhou, X. Fan, X. Liang, and Y. Liu, *Idsradar: a real-time visualization framework for ids alerts*, *Science China Information Sciences* (2013) 1–12.

- [151] X. Yin, W. Yurcik, M. Treaster, Y. Li, and K. Lakkaraju, *Visflowconnect: netflow visualizations of link relationships for security situational awareness*, in *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security, VizSEC/DMSEC '04*, (New York, NY, USA), pp. 26–34, ACM, 2004.
- [152] K. Abdullah, C. Lee, G. Conti, J. A. Copeland, and J. Stasko, *Ids rainstorm: Visualizing ids alarms*, in *Proceedings of the IEEE Workshops on Visualization for Computer Security, VIZSEC '05*, (Washington, DC, USA), pp. 1–, IEEE Computer Society, 2005.
- [153] S. D. Kamvar and J. Harris, *We feel fine and searching the emotional web*, in *Proceedings of the fourth ACM international conference on Web search and data mining, WSDM '11*, (New York, NY, USA), pp. 117–126, ACM, 2011.
- [154] M. S. Eastin, *Credibility assessments of online health information: The effects of source expertise and knowledge of content*, *Journal of Computer-Mediated Communication* **6** (2001), no. 4 0–0.
- [155] A. Abdul-Rahman and S. Hailes, *Using recommendations for managing trust in distributed systems*, in *In Proceedings of IEEE Malaysia International Conference on Communication97 (MICC97)*, Kuala Lumpur, Malaysia, Citeseer, 1997.
- [156] A. Abdul-Rahman and S. Hailes, *A distributed trust model*, in *Proceedings of the 1997 Workshop on New Security Paradigms, NSPW '97*, (New York, NY, USA), pp. 48–60, ACM, 1997.
- [157] B. Yu and M. P. Singh, *An evidential model of distributed reputation management*, in *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1, AAMAS '02*, (New York, NY, USA), pp. 294–301, ACM, 2002.
- [158] C. M. Jonker, J. J. Schalken, J. Theeuwes, and J. Treur, *Human experiments in trust dynamics*, in *Trust Management*, pp. 206–220. Springer, 2004.
- [159] S. Few, *Data visualization for human perception*, *The Encyclopedia of Human-Computer Interaction, 2nd Ed.* (2013).
- [160] A. Java, X. Song, T. Finin, and B. 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, WebKDD/SNA-KDD '07*, (New York, NY, USA), pp. 56–65, ACM, 2007.
- [161] K. Starbird and J. Stamberger, *Tweak the tweet: Leveraging microblogging proliferation with a prescriptive syntax to support citizen reporting*, *Proceedings of the 7th International ISCRAM Conference Seattle, USA* (2010).

- [162] A. Burns and B. Eltham, *Twitter free iran: An evaluation of twitter's role in public diplomacy and information operations in iran's 2009 election crisis*, *Communications Policy & Research Forum 2009, 19th-20th November 2009*, University of Technology, Sydney (2009).
- [163] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, *Microblogging during two natural hazards events: What twitter may contribute to situational awareness*, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, (New York, NY, USA), pp. 1079–1088, ACM, 2010.
- [164] D. L. Lasorsa, S. C. Lewis, and A. E. Holton, *Normalizing twitter: Journalism practice in an emerging communication space*, *Journalism Studies* **13** (2012), no. 1 19–36.
- [165] J. Holcomb, J. Gottfried, and A. Mitchell, *News use across social media platforms*, 2013.
- [166] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller, *Twitinfo: Aggregating and visualizing microblogs for event exploration*, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, (New York, NY, USA), pp. 227–236, ACM, 2011.
- [167] M. Rios and J. Lin, *Visualizing the "pulse" of world cities on twitter*, in *Proceedings of the Seventh International Conference on Weblogs and Social Media, Cambridge, MA, USA, July 8-11, 2013.*, 2013.
- [168] B. Kang, G. Legrady, and T. Höllerer, *Tweetprobe: A real-time microblog stream visualization framework*, *Proceedings of the IEEE VIS Arts Program (VISAP)* (2013).
- [169] B. Kang, T. Höllerer, and J. O'Donovan, *Believe it or not? analyzing information credibility in microblogs*, in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, (New York, NY, USA), pp. 611–616, ACM, 2015.
- [170] S. Goel, J. Hofman, and M. Siner, *Who does what on the web: A large-scale study of browsing behavior*, 2012.
- [171] S. Garcia Esparza, M. P. O'Mahony, and B. Smyth, *Catstream: Categorising tweets for user profiling and stream filtering*, in *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, (New York, NY, USA), pp. 25–36, ACM, 2013.
- [172] L. P. Morton and J. Warren, *News elements and editors' choices*, *Public Relations Review* **18** (Mar., 1992) 47–52.

- [173] P. Melville, R. Mooney, and R. Nagarajan, *Content-boosted collaborative filtering*, in *In Proceedings of the Eighteenth National Conference on Artificial Intelligence*, 2002.
- [174] J. L. Herlocker, J. A. Konstan, and J. Riedl, *Explaining collaborative filtering recommendations*, in *Proceedings of ACM CSCW'00 Conference on Computer-Supported Cooperative Work*, pp. 241–250, 2000.
- [175] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, *GroupLens: An open architecture for collaborative filtering of netnews*, in *Proceedings of ACM CSCW'94 Conference on Computer-Supported Cooperative Work*, pp. 175–186, 1994.
- [176] O. Phelan, K. McCarthy, and B. Smyth, *Using twitter to recommend real-time topical news*, in *Proceedings of the third ACM conference on Recommender systems*, pp. 385–388, ACM, 2009.
- [177] J. C. McCroskey, *Scales for the measurement of ethos*, Taylor & Francis (1966).
- [178] S. Epstein, R. Pacini, V. Denes-Raj, and H. Heier, *Individual differences in intuitive-experiential and analytical-rational thinking styles.*, *J Pers Soc Psychol* **71** (1996), no. 2 390–405.
- [179] M. O'Connor, D. Cosley, J. A. Konstan, and J. Riedl, *Polylens: a recommender system for groups of users*, in *ECSCW 2001*, pp. 199–218, Springer, 2001.
- [180] P. T. Metaxas and E. Mustafaraj, *From obscurity to prominence in minutes: Political speech and real-time search*, in *Proceedings of Web Science Conference*, 2010.
- [181] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan, *Opinionfinder: A system for subjectivity analysis*, in *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pp. 34–35, Association for Computational Linguistics, 2005.
- [182] *Natural language toolkit*, Jan., 2014.
- [183] A. Hermida, F. Fletcher, D. Korell, and D. Logan, *Share, like, recommend: Decoding the social media news consumer*, *Journalism Studies* **13** (2012), no. 5-6 815–824.
- [184] C. Budak, S. Goel, and J. M. Rao, *Fair and balanced? quantifying media bias through crowdsourced content analysis*, in *Proceedings of the Ninth International Conference on Weblogs and Social Media, Oxford, UK, AAAI*, 2015.

- [185] D. Bär, C. Biemann, I. Gurevych, and T. Zesch, *Ukp: Computing semantic textual similarity by combining multiple content similarity measures*, in *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics, SemEval '12*, (Stroudsburg, PA, USA), pp. 435–440, Association for Computational Linguistics, 2012.
- [186] M. A. Yildirim and M. Coscia, *Using random walks to generate associations between objects.*, *PLoS ONE* **9** (2014), no. 8.
- [187] D. M. Blei, A. Y. Ng, and M. I. Jordan, *Latent dirichlet allocation*, *J. Mach. Learn. Res.* **3** (Mar., 2003) 993–1022.
- [188] A. A. Benczúr and D. R. Karger, *Approximating st minimum cuts in $\tilde{O}(n^2)$ time*, in *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pp. 47–55, ACM, 1996.
- [189] B. Hilligoss and S. Y. Rieh, *Developing a unifying framework of credibility assessment: Construct, heuristics and interaction in context*, *Information Processing and Management* **44** (2008) 1467–1484.
- [190] S. Adalı, *Modeling Trust Context in Networks*. Springer Briefs, 2013.
- [191] D. Boyd, S. Golder, and G. Lotan, *Tweet, tweet, retweet: Conversational aspects of retweeting on twitter*, in *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences, HICSS '10*, (Washington, DC, USA), pp. 1–10, IEEE Computer Society, 2010.
- [192] J. Keller, “How truth and lies spread on twitter.” <http://www.bloomberg.com/bw/articles/2012-10-31/how-truth-and-lies-spread-on-twitter>, 2012. ”October 31, 2012.”.
- [193] S. Adalı, M. Magdon-Ismail, and F. Sisenda, *Actions speak as loud as words: Predicting relationships from social behavior data*, in *Proceedings of the WWW Conference*, 2012.
- [194] D. Z. Levin and R. Cross, *The strength of weak ties you can trust: The mediating role of trust in effective knowledge transfer*, *Academy of Management Journal* **50** (2002), no. 11 1477–1490.
- [195] J. Baumes, M. Goldberg, M. Hayvanovych, M. Magdon-Ismail, W. Wallace, and M. Zaki, *Finding hidden group structure in a stream of communications*, *Intel. and Sec. Inform. (ISI)* (2006).
- [196] A. Todorov and N. N. Oosterhof, *Modeling social perception of faces*, *IEEE Signal Processing Magazine* **117** (2011).

- [197] V. Wout and Sanfey, *Friend or foe: the effect of implicit trustworthiness judgments in social decision-making*, *Cognition* **108** (2008), no. 3 796–803.
- [198] A. Todorov, A. N. Mandisodza, A. Goren, and C. C. Hall, *Inferences of competence from faces predict election outcomes*, *Science* **308** (2005) 1623–1626.
- [199] L. J. Chang, B. B. Doll, M. van’t Wout, M. J. Frank, and A. G. Sanfey, *Seeing is believing: Trustworthiness as a dynamic belief*, *Cognitive Psychology* **61** (2010), no. 2 87–105.
- [200] M. Chen, D. Ebert, H. Hagen, R. Laramee, R. van Liere, K.-L. Ma, W. Ribarsky, G. Scheuermann, and D. Silver, *Data, information, and knowledge in visualization*, *Computer Graphics and Applications, IEEE* **29** (Jan, 2009) 12–19.
- [201] D. A. Keim, *Information visualization and visual data mining*, *IEEE Transactions on Visualization and Computer Graphics* **8** (2002), no. 1 1–8.
- [202] T. H. H. L. James Schaffer, Byungkyu Kang and J. O’Donovan, *Interactive interfaces for complex network analysis: An information credibility perspective*, in *IEEE International Conference on Pervasive Computing and Communications (PERCOM) 2013*, pp. 464–469, 2013.
- [203] B. Gretarsson, J. O’Donovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman, and P. Smyth, *Topicnets: Visual analysis of large text corpora with topic modeling*, *ACM Trans. Intell. Syst. Technol.* **3** (Feb., 2012) 23:1–23:26.
- [204] J. Schaffer, P. Giridhar, D. Jones, T. Höllerer, T. Abdelzaher, and J. O’Donovan, *Getting the message?: A study of explanation interfaces for microblog data analysis*, in *Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI ’15*, (New York, NY, USA), pp. 345–356, ACM, 2015.
- [205] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher, *On truth discovery in social sensing: a maximum likelihood estimation approach*, in *Proceedings of the 11th international conference on Information Processing in Sensor Networks, IPSN ’12*, (New York, NY, USA), pp. 233–244, ACM, 2012.
- [206] N. Cawthon and A. Moere, *The effect of aesthetic on the usability of data visualization*, in *Information Visualization, 2007. IV ’07. 11th International Conference*, pp. 637–648, 2007.
- [207] J. Sparks, *The Histomap of Evolution. [A Chart. With "Foreword, Bibliography and Recommended Books."]*. Histomap, 1932.
- [208] B. Smyth and P. McClave, *Similarity vs. diversity*, in *Proceedings of the 4th International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development, ICCBR ’01*, (London, UK, UK), pp. 347–361, Springer-Verlag, 2001.

- [209] E. Pariser, *The filter bubble: What the Internet is hiding from you*. Penguin Books, 2011.
- [210] P. Resnick, R. K. Garrett, T. Kriplean, S. A. Munson, and N. J. Stroud, *Bursting your (filter) bubble: Strategies for promoting diverse exposure*, in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion, CSCW '13*, (New York, NY, USA), pp. 95–100, ACM, 2013.
- [211] J. Teevan, M. R. Morris, and S. Azenkot, *Supporting interpersonal interaction during collaborative mobile search*, *Computer* **47** (2014), no. 3 54–57.
- [212] M. S. Granovetter, *The Strength of Weak Ties*, *The American Journal of Sociology* **78** (1973), no. 6 1360–1380.
- [213] P. Brusilovsky, E. Schwarz, and G. Weber, *Elm-art: An intelligent tutoring system on world wide web*, in *Intelligent Tutoring Systems*, 1996.
- [214] V. Dimitrova, *Style-olm: Interactive open learner modelling*, *International Journal of Artificial Intelligence in Education* **17(2)** (2003) 35–78.
- [215] F. Cerutti, N. Tintarev, and N. Oren, *Formal arguments, preferences, and natural language interfaces to humans: an empirical evaluation*, in *ECAI*, pp. 207–212, 2014.
- [216] N. Tintarev and R. Kutlak, *Explanations - making plans scrutable with argumentation and natural language generation*, in *Intelligent User Interfaces (demo track)*, 2014.
- [217] S. W. Bennett and A. C. Scott., *The Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, ch. 19 - Specialized Explanations for Dosage Selection, pp. 363–370. Addison-Wesley Publishing Company, 1985.
- [218] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, *Principles of explanatory debugging to personalize interactive machine learning*, in *IUI*, 2015.
- [219] N. Tintarev and J. Masthoff, *Recommender Systems Handbook (second ed., in print)*, ch. Explaining Recommendations: Design and Evaluation. Springer, 2015.
- [220] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell, *Explaining the user experience of recommender systems*, *User Modeling and User-Adapted Interaction* **22** (2012), no. 4-5 441–504.
- [221] R. A. Amar and J. T. Stasko, *Knowledge precepts for design and evaluation of information visualization*, *IEEE Trans. Visualization and Computer Graphics* **11** (2005) 432–442.

- [222] K. Verbert, D. Parra, P. Brusilovsky, and E. Duval, *Visualizing recommendations to support exploration, transparency and controllability*, in *Proceedings of the 2013 International Conference on Intelligent User Interfaces, IUI '13*, (New York, NY, USA), pp. 351–362, ACM, 2013.
- [223] J. Schaffer, P. Giridhar, D. Jones, T. Höllerer, T. Abdelzaher, and J. O'Donovan, *Getting the message?: A study of explanation interfaces for microblog data analysis*, in *Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI '15*, (New York, NY, USA), pp. 345–356, ACM, 2015.
- [224] B. P. Knijnenburg, S. Bostandjiev, J. O'Donovan, and A. Kobsa, *Inspectability and control in social recommenders*, in *Proceedings of the Sixth ACM Conference on Recommender Systems, RecSys '12*, (New York, NY, USA), pp. 43–50, ACM, 2012.
- [225] S. Nagulendra and J. Vassileva, *Providing awareness, understanding and control of personalized stream filtering in a p2p social network*, in *Conference on Collaboration and Technology (CRIWG)*, 2013.
- [226] A. Sharma and D. Cosley, *Do social explanations work? studying and modeling the effects of social explanations in recommender systems*, in *World Wide Web (WWW)*, 2013.
- [227] B. Wang, M. Ester, J. Bu, and D. Cai, *Who also likes it? generating the most persuasive social explanations in recommender systems*, in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.
- [228] P. André, m.c. schraefel, J. Teevan, and S. T. Dumais, *Discovery is never by chance: Designing for (un)serendipity*, in *ACM Creativity & Cognition*, 2009.
- [229] A. Paramythis, S. Weibelzahl, and J. Masthoff, *Layered evaluation of interactive adaptive systems: Framework and formative methods.*, *User Modeling and User-Adapted Interaction* **20** (2010).
- [230] J. D. Weinberg, J. Freese, and D. McElhattan, *Comparing data characteristics and results of an online factorial survey between a population-based and a crowdsource-recruited sample*, *Sociological Science* **1** (2014) 292–310.
- [231] S. D. Gosling, P. J. Rentfrow, and W. B. Swann, *A very brief measure of the big-five personality domains*, *Journal of Research in personality* **37** (2003), no. 6 504–528.