Health Misinformation in Search and Social Media

by

Amira Ghenai

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2019

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Joan Bartlett
Associate Professor, School of Information Studies
University of McGill

Supervisor(s): Charles L.A Clarke
Professor, School of Computer Science
University of Waterloo
Mark D. Smucker
Associate Professor, Department of Management Sciences
University of Waterloo

Internal Members: Robin Cohen
Professor, School of Computer Science
University of Waterloo
Maura R. Grossman
Research Professor, School of Computer Science
University of Waterloo

Internal-External Member: Lukasz Golab
Associate Professor, Department of Management Sciences
University of Waterloo

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Related to the work conducted about health misinformation in social media, I collaborated with Yelena Mejova in the parts presented in Chapters 3 and 4. Yelena was a scientist in the Qatar Computing Research Institute, a member of the Qatar Foundation organization where I received my doctoral scholarship, during the time we were collaborating. Yelena's main role was to advise and provide guidance on the work we conducted. She further conducted specific research tasks including collecting the tweets, facilitating the access to computing platforms, performing the crowdsourcing tasks, helping with the data analysis parts, and contributing to the writing process of the related conference papers. I collected the historical data for users and the list of rumors, performed the medical and URL lexicon creation, implemented the NLP and data parsing tools during the data processing phase, applied LDA modeling, built most of the features and the classification tasks, maintained thorough documentation on the project website, contributed to the writing of the associated conference papers, and wrote Chapters 3 and 4.

In the first study related to the research area of misinformation for online health search presented in Chapter 5, I collaborated with my colleague Frances Pogacar in conducting this research. Frances originated the idea for the study, designed the study, partially completed the work needed to conduct the study, helped with the data analysis, and contributed to the writing of the associated conference paper. I refined the study, completed the work needed to conduct the study, ran the study and collected all data, analyzed data, contributed to the writing of the associated conference paper, and wrote Chapter 5.

I am the sole contributor on Chapter 6.

My supervisors Charles L.A Clarke and Mark D. Smucker provided supervisory guidance throughout the whole research presented in this PhD thesis.

Abstract

People increasingly rely on the Internet in order to search for and share health-related information. Indeed, searching for and sharing information about medical treatments are among the most frequent uses of online data. While this is a convenient and fast method to collect information, online sources may contain incorrect information that has the potential to cause harm, especially if people believe what they read without further research or professional medical advice.

The goal of this thesis is to address the misinformation problem in two of the most commonly used online services: search engines and social media platforms. We examined how people use these platforms to search for and share health information. To achieve this, we designed controlled laboratory user studies and employed large-scale social media data analysis tools. The solutions proposed in this thesis can be used to build systems that better support people's health-related decisions.

The techniques described in this thesis addressed online searching and social media sharing in the following manner. First, with respect to search engines, we aimed to determine the extent to which people can be influenced by search engine results when trying to learn about the efficacy of various medical treatments. We conducted a controlled laboratory study wherein we biased the search results towards either correct or incorrect information. We then asked participants to determine the efficacy of different medical treatments. Results showed that people were significantly influenced both positively and negatively by search results bias. More importantly, when the subjects were exposed to incorrect information, they made more incorrect decisions than when they had no interaction with the search results.

Following from this work, we extended the study to gain insights into strategies people use during this decision-making process, via the think-aloud method. We found that, even with verbalization, people were strongly influenced by the search results bias. We also noted that people paid attention to what the majority states, authoritativeness, and content quality when evaluating online content. Understanding the effects of cognitive biases that can arise during online search is a complex undertaking because of the presence of unconscious biases (such as the search results ranking) that the think-aloud method fails to show.

Moving to social media, we first proposed a solution to detect and track misinformation in social media. Using Zika as a case study, we developed a tool for tracking misinformation on Twitter. We collected 13 million tweets regarding the Zika outbreak and tracked rumors outlined by the World Health Organization and the Snopes fact-checking website. We incorporated health professionals, crowdsourcing, and machine learning to capture health-related rumors as well as clarification communications. In this way, we illustrated insights that the proposed tools provide into potentially harmful information on social media, allowing public health researchers and practitioners to respond with targeted and timely action.

From identifying rumor-bearing tweets, we examined individuals on social media who are posting questionable health-related information, in particular those promoting cancer treatments that have been shown to be ineffective. Specifically, we studied 4,212 Twitter users who have posted about one of 139 ineffective "treatments" and compared them to a baseline of users generally interested in cancer. Considering features that capture user attributes, writing style, and sentiment, we built a classifier that is able to identify users prone to propagating such misinformation. This classifier achieved an accuracy of over 90%, providing a potential tool for public health officials to identify such individuals for preventive intervention.

Acknowledgements

First, I would like to thank my supervisors, Prof. Charles L. A. Clarke and Prof. Mark D. Smucker, for their support throughout the progress I have achieved so far in my PhD research. My academic achievements would not have been possible without their help and guidance.

Besides my advisors, my deepest gratitude goes to Dr. Yelena Mejova from the ISI Foundation & ISI Global Science Foundation for her knowledge, encouragement and feedback. Her guidance helped me improve throughout my academic and professional life.

Many thanks to my colleague Frances Pogacar for the fruitful collaboration. Thanks to my IR lab colleagues, specially Nimesh Ghelani and Gaurav Baruah, for their help during my research.

Throughout my published work, I had the privilege of collaborating with experts from different fields. Specifically, I would like to thank Dr. Ryen White, Applied Science Manager, Microsoft Cortana, for his assistance and knowledge. I would also like to thank Luis Fernandez-Luque, scientist, Qatar Computing Research Institute, William Schulz, research assistant, London School of Hygiene & Tropical Medicine, Clarissa Simas, research assistant, London School of Hygiene & Tropical Medicine and Per Egil Kummervold, research scientist, Norut Northern Research Institute for the valuable discussions and their health-related expertise. I would further like to express my gratitude to Dr. Jeremie Arash Rafii Tabrizi, who assisted with misinformation topic evaluation, query validation, and overall advice.

I would like to thank the Qatar Science and Leadership scholarship program at the Qatar Foundation for Education, Science and Community Development for their financial support. I also thank the Natural Sciences and Engineering Research Council of Canada (Grants CRDPJ 468812-14 and RGPIN-2014-03642), Google, and the University of Waterloo for supporting projects that are part of my PhD research. Further, I would like to thank Dania Sheldon for taking the time to provide copy and stylistic editing of this thesis.

Finally, I wish to offer my special thanks to my parents for their moral and emotional support and for always giving me the liberty to choose what I desire. Most importantly, I am immensely grateful to my husband, the love of my life, for showing strong faith in me and for always being around during hard times. Thank you!

Dedication

I would like to dedicate this thesis to:
my husband, Mr. Mohamed Nadjib Salmi,
my parents, Mr. Salah Eddine Ghenai and Mrs. Naziha Boudra,
and my lovely daughter, Yara Salmi.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Statement

In the past, medical professionals as well as patients have relied on medical reports from hospitals and health clinics to search for health-related topics. Not until the 1960s, were steps taken towards automating data collection (de Lusignan and van Weel, 2006). As much as the old data resources seemed to be reliable and focused, this kind of data does not cover all aspects of public health but instead focuses more on illnesses (Yom-Tov, 2016). As a result, people find the old medical data resources limited and hard to access.

Nowadays, people choose to search for health-related topics online because of the massive amount of information available and the ease with which they can access it. Indeed, using search engines to look for health-related news is the third most common online news (66% of Internet news users) after the weather (81%) and national events (73%) (Purcell et al., 2010) according to a study done by Pew Internet and American Life Project. Additionally, Internet usage estimates show that 80% of adult Internet users in the US seek health advice online (Fox, 2011). A telephone-based survey conducted by the Pew Internet and American Life Project showed that 44% of the participants had changed their decisions about how to treat an illness or a condition after consulting search results (Rice, 2006).

In addition to online searching, sharing health information through online platforms is becoming a common practice. In 2011, a survey conducted by Pew Internet & American Life Project indicated that as many as 62% of adult Internet users in the U.S. used social network sites for health-related topics, from following friends' health experiences or updates to gathering health-related information (Fox, 2011). Some manage their health via general platforms such as PatientsLikeMe, while others join condition-specific communities such as TuDiabetes, yet others share their experiences in general-purpose social media (De Choudhury et al., 2014; Korda and Itani, 2013). For instance, Paul et al. (2011) showed that a significant number of personal and health-related questions were being asked on the microblogging platform Twitter, which is becoming the top destination for both patients and healthcare professionals (Antheunis et al., 2013). The social component of such interactions is especially important with respect to educating the population on health matters, as individuals' opinions can be strongly biased by their social network (Lau et al., 2011).

Using the massive amount of data available online, we can potentially help improve people's health in a direct and quantitative way (De Choudhury et al., 2013; Yom-Tov and Boyd, 2014; Yom-Tov and Gabrilovich, 2013; Yom-Tov et al., 2015b,a; Sadilek et al., 2012; Yom-Tov et al., 2014a,b). However, while being a convenient and fast method to collect information, online health search presents health-related facts without regard for their correctness. The presence of incorrect information in search results has the potential to cause harm, especially as the majority of Internet users are confident about and believe the information that they find online, according to the Pew Internet & American Life survey (Purcell et al., 2012).

Furthermore, while the use of social media for health management is growing, so are the concerns over the lack of accountability, dubious quality, and unreliable confidentiality (Greene et al., 2011; Moorhead et al., 2013). With few legal constraints imposed on what are frequently profit-seeking websites, social media provides a dynamic forum for propagating medical misinformation (Frish and Greenbaum, 2017).

If people believe what they read online without further research or professional medical advice, they are likely to make incorrect decisions that waste their money, negatively impact their health, or both. A notable example is the story of Wei Zexi, a 21-year-old Chinese student who died of synovial sarcoma, a form of cancer (Ouyang, 2016). At the beginning of his sickness, doctors treated Wei using the known standard treatment methods. However, when the illness did not improve, Wei's family looked for alternatives via the Baidu search engine and learned about an experimental trial treatment. Desperately seeking recovery for their child, the family decided to spend 200,000 yuan (US$30,650) so Wei could participate in the trial which had not yet been approved by the Chinese government. Later, the family learned from friends outside China that there was no scientific evidence supporting the new treatment. With time, Wei's health deteriorated, and he died on April 12, 2016. Before his death, he shared a web post explaining how Baidu had violated his trust by showing the treatment advertisement on a highly ranked search results page. The post caught significant public attention and resulted in the Chinese government imposing new regulations on search engines (Abkowitz, 2016). Among these new rules was a requirement that search engines clearly identify advertisements as different from natural or organic search results (Abkowitz and Chin, 2016).

Another relevant example of online health search implications lies in the story of three women, from the United States aged between 72 and 88, who were slowly losing their sight as a result of age-related macular degeneration (AMD). In June 2015, the patients found out about a trial that used tissue-derived stem cells to treat patients with AMD. They decided to pay US$ 5,000 to participate in the trial (Kuriyan et al., 2017). After undergoing the treatment, the patients experienced complete vision loss. This story is notable because the patients found out about the trial from the online source ClinicalTrials.gov. At that time, the trial had not yet been evaluated by the US Federal Government. Subsequently, in September 2015, the trial was withdrawn (Kuriyan et al., 2017). When the three patients had searched online for a medical solution, the search results had not mention that the treatment was unhelpful, and the so-called cure actually ended up harming their health.

The impact of online health usage extends beyond online search. In the context of people using social media for health purposes, a recent rise in vaccine hesitancy has been linked to an active movement on Twitter, promoting conspiratorial thinking and mistrust in the government (Mitra et al., 2016). Image-sharing platforms such as Flickr and Instagram have become battlegrounds between the pro-anorexia movement and physicians attempting to intervene (Yom-Tov et al., 2012; Chancellor et al., 2016). Uncertainty surrounding infectious disease outbreaks, such as the Zika epidemic of 2016, yielded rumors and speculations about its causes and consequences, and how to prevent it (Dredze et al., 2016; Ghenai and Mejova, 2017).

1.2 Proposed Solutions

In this dissertation, using Internet data, we aim to understand how information and, more specifically, misinformation affect people's health-related decisions where we define misinformation as a piece of information spread via the online media and confirmed to be false by reliable sources (Coady, 2006). The Internet data sources are varied and include search engines, social media, online forums, and mobile applications, among others. We will focus on the use of search engines as well as social media for health-related purposes. We choose search engines and social media because they are two of the most commonly used online services (De Choudhury et al., 2014). We next describe how we approach and tackle the misinformation problem in each chapter.

In Chapter 3 as well as Chapter 4, we aim at understanding whether we can we detect and track misinformation in social media. Specifically, in Chapter 3, we use Zika as a case study for building a tool to track the misinformation around health concerns on Twitter. We collect more than 13 million tweets regarding the Zika outbreak and track rumors outlined by the World Health Organization and the Snopes fact-checking website. The tool pipeline, which incorporates health professionals, crowdsourcing, and machine learning, allows us to capture health-related rumors around the world, as well as clarification campaigns by reputable health organizations. Moreover, the suggested solution allows public health researchers and practitioners to respond with timely and targeted actions to trending potentially harmful information spreading in social media.

Next, in Chapter 4, we examine individuals on social media who are posting questionable health-related information, and in particular promoting cancer treatments shown to be ineffective (making it a kind of misinformation, willful or not). Using a multi-stage user selection process, we study 4,212 Twitter users who have posted about one of 139 such “treatments” and compare them to a baseline of users generally interested in cancer. Considering features that capture user attributes, writing style, and sentiment, we build a classifier that is able to automatically identify users prone to propagate such misinformation, providing a potential tool for public health officials to identify such individuals for preventive intervention.

Chapters 5 and 6 focus on the effect of search result pages on people’s health-related decisions. In Chapter 5, we conduct a controlled laboratory study where we biased search results towards correct or incorrect information for 10 different medical treatments. The goal of the study was to see whether search results would influence people both positively and negatively. The findings of the study confirm that search engine designers and researchers must recognize that not all non-relevant information is harmless. Some non-relevant information is incorrect and potentially harmful when people use it to make decisions that may negatively impact their lives. One potential solution is to introduce a notion of negative gain to incorrect documents when search engines retrieve relevant documents.

We extend the study presented in Chapter 5 by designing a think-aloud study, described in Chapter 6, to understand the decision-making process for determining the efficacy of medical treatments using search result pages. As think-aloud protocols have been used in the past to build models for understanding cognitive processes (Van Someren et al., 1994), we use a think-aloud study to gain insights into strategies people use during online search for health-related topics. The discovered insights are potentially helpful for improving current search engines to better support people’s health-related decisions.

We review the related work in Chapter 2 and conclude the thesis in Chapter 7. In Appendix C, we list the definitions of the most frequent technical terms used in this report.

1.3 List of Contributions

In this thesis, we make the following contributions in terms of online health search:

- People can be significantly influenced by the information present in search result pages. When biased towards correct information, the accuracy of answering health-related questions increases from 43% to 65%. On the other hand, when there is a bias towards incorrect information, the accuracy decreases from 43% to 23%. (Chapter 5)

- The rank of the topmost correct result has some effect on people's accuracy: the accuracy was 59% only when the top two ranked results were incorrect, compared to 70% accuracy when the result placed at rank 1 in the search results was correct. (Chapter 5)

- As prior research shows (White and Hassan, 2014), people have an uncontrolled bias towards believing that treatments are helpful, regardless of the ground-truth. (Chapter 5)

- Knowledge of the medical treatment can protect searchers from the incorrect information potentially available in search results. In this regard, we found that more self-reported knowledge reduced the effect of incorrect information on accuracy ($p = 0.04$). (Chapter 5)

- Even with verbalization in the think-aloud study explained in Chapter 6, participants were heavily influenced by a search result bias. When biased towards correct information, participants' accuracy reached 67%, whereas the accuracy was reduced to 32% when search results were biased towards incorrect information. (Chapter 6)

- People pay a large amount of attention to what the majority of the search results state. Furthermore, to make their decisions about the efficacy of a medical treatment, people look for information related to the concepts of authoritativeness and quality. (Chapter 6)

- Even though prior research shows that people pay attention to highly ranked results (Allam et al., 2014; Haas and Unkel, 2017), when doing the think-aloud study, participants did not talk about the notion of rank. This shows that rank is a potentially subconscious bias that, the think-aloud study fails to reveal. (Chapter 6)

Further, we introduce the following new knowledge in terms of social media health usage:

- Extracting rumor content in social medial is not a trivial task: high-precision approaches such as keyword search capture roughly half of the actual rumor content. (Chapter 3)

- Rumors in social media are varied in terms of longevity and severity: everyday activities rumors stay longer and have a higher potential to propagate misinformation. (Chapter 3)

- We can successfully build an automatic rumor classifier. However, the features used could be easily manipulated. This suggests how important it is for health authorities to rectify incorrect facts during health crisis. (Chapter 3)

- Early health rumor detection in social media is challenging. Early detected rumors might in fact be true, as it takes time to access the veracity of health rumors. (Chapter 3)

- We can successfully build a tool to detect users prone to propagating health rumor-mongering. This tool has an accuracy of over 90%, providing a potential means for public health officials to identify such individuals for preventive intervention. (Chapter 4)

- The multifaceted proposed set of behavioral and content features, presented in the model explained in Chapter 4, suggest a foundation for examining behavioral characteristics and interests of people susceptible to rumors. (Chapter 4)

- The dataset collected in the work presented in Chapter 4 presents a highly curated resource for the research community's future studies on the topic of health misinformation. (Chapter 4)

Our results highlight the importance of studying the influence of health misinformation in both web search and social media. In online health search, people can be potentially harmed by incorrect search results. Possible factors that affect people's decisions are what the majority of the search results state, the search results' authoritativeness, and the pages' quality. Further, rank and helpfulness factors are examples of unconscious biases that affect decision making. In social media, people share inaccurate information and can thereby contribute to spreading rumors. We build a tool to identify users posting questionable information, which can help health authorities intervene to change individuals' views and quickly identify and limit the spread of misinformation.

The work presented in Chapters 3 and 4 contributes to the Social Computing community (such as the CSCW conference on Computer-Supported Cooperative Work and Social Computing and the International AAAI ICWSM Conference on Web and Social Media) in proposing a tool for monitoring health misinformation on a large scale. Specifically, the model built in these Chapters exemplifies specialized tools that can help address the spread of health misinformation on social media, mainly in (i) automatically detecting Twitter users who may be likely to post questionable information, (ii) attempting to change those individuals' view of the topic, and (iii) quickly identifying and limiting the spread of misinformation.

Further, the work conducted in Chapters 5 and 6 contributes to the Human Information Interaction community (such as the ACM SIGIR CHIIR conference on Human Information Interaction and Retrieval and the ACM CHI Conference on Human Factors in Computing Systems) in designing user-centered studies to evaluate searchers' performance when doing medical information retrieval tasks. Specifically, the studies explained in those Chapters show how search results can potentially lead people to make incorrect decisions. The findings where mainly as follows: (i) searchers are heavily influenced with search results when making health-related decisions (ii) the majority of what the search results state is a potential reason why people make incorrect decisions (iii) authoritativeness and quality are factors people look for when evaluating search results. These observations show that is is paramount to design ways to better support searchers to make more informed decisions. An example would be introducing a notion of negative gain to incorrect documents that might be harmful.

Chapter 2

Background and Related Work

As online health search is becoming a very critical part of our daily lives, the literature in this area is extensive. In this section, we will first outline the key references about online misinformation and rumor spreading in general. Next, we will explore specifically the health domain and identify the research about online health search. Later, we will summarize the work done on social media sites' role in public health.

2.1 Misinformation Tracking

In 2017, the term "fake news" was Collins' Word of the Year, referring to "false, often sensational, information."[1] Similarly, separating newsworthy stories from misinformation across online sites has been a popular research topic in recent years, as low-quality news has the potential for negative impact on individuals and societies (Shu et al., 2017). Here, we detail the work conducted on rumor tracking in social media. Later, we describe the work about misinformation in search results and its effects on people's decisions.

2.1.1 Microblogging Sites

There are three different types of misinformation that can spread: fraudulent journalistic writing, satirical fake news, and largescale hoaxes. The last type has been claimed to be the one best suited to be spread in social media as per Rubin et al. (2015). Different approaches

[1] https://ind.pn/2AnI2Bw

have been proposed in the literature to access the credibility in microblogging sites. A vast amount of work has been done in this area, which can be divided into three main categories: classification-based approaches, network-based approaches, and survey-based studies.

Classification-Based Approaches

A large amount of work focuses on identifying the credibility of news propagated in social media by building classification tools (Castillo et al., 2011; Leskovec et al., 2009; Yang et al., 2012; Wu et al., 2015), and responding to misinformation in real time (Liu et al., 2015). Often, machine learning models based on features related to either users or the content of propagated messages are employed. In their first work, Castillo et al. (2011) used multiple features extracted from trending topic tweets to classify the messages as either credible or not. The authors defined *message-based*, *topic-based*, and *user-based* features. The proposed technique is a supervised learning approach where authors prepared a dataset of predefined labels (using crowdsourcing) to train and test the suggested classifier. Later, the authors proposed another approach in which they built two classifiers to identify information cascades corresponding to "newsworthy" events (Information cascade is a list of all of the messages which usually accompany newsworthy events) (Gayo-Avello et al., 2013).

Kang et al. (2012) used different feature types to build different classifiers that recommend the credibility level of topics in Twitter. Similar to Castillo et al. (2011), the authors used *network-based* features, *topic-based* features, and a combination of both on a set of manually labeled tweets. Different from Castillo et al. (2011), we propose a model to access each tweet individually instead of classifying a group of "newsworthy" tweets. Both Castillo et al. (2011) and Kang et al. (2012) agree that network-based features are better predictors of information credibility than linguistic features.

Gupta and Kumaraguru (2012) used *user-based* features as well as *message-based* features to build a logistic linear regression model that identifies the credibility of Twitter messages. Later authors adopted an SVM ranking algorithm with relevance feedback to rank tweets based on their credibility. Gupta et al. (2014) extended the previous work to introduce a real-time system called *TweetCred* that works as a Chrome extension. The proposed system evaluates the credibility level of each tweet in real time and returns its credibility value on a scale of 1 (low credibility) to 7 (high credibility) using the ranking SVM model in the previous work. In terms of features, the authors used a larger set of features than in the previous work including *message-based* features, *user-based* features, and *network-based* features. In another work, instead of identifying tweets' credibility, Gupta et al. (2013) used visual feature extraction tools to identify fake images about Hurricane

Sandy (2012) in Twitter. They built a Decision Tree classifier model to distinguish fake images from real ones and reached the highest accuracy (97%) with *message-based* features.

Unlike these studies, in a more natural setting where no labeled data were available, Qazvinian et al. (2011) used content-based, network-based, and Twitter-specific features to track urban legends. They built Bayes classifiers using engineered features and then learned a linear function of these classifiers for rumor retrieval and classification.

Wu et al. (2015) and Yang et al. (2012) trained a graph-kernel based hybrid SVM classifier to automatically detect rumors on the Sina Weibo micro-blog site.[2] Instead of manually labeling the content, the authors used the Sina Weibo rumor-busting service[3] as ground-truth data for the classification. Similar to most of the prior work, the authors used *message-based* and *user-based* features to build the classifier.

Ma et al. (2015) argued that the previously explained models ignore the importance of variation in social and contextual features during message propagation over time. They proposed a novel approach to capture the temporal characteristics of these features based on the time series of a rumor's lifecycle. Specifically, they use both Twitter and Sina Weibo as case studies to detect rumors in social media. Using *Dynamic Series-Time Structure* features to build classifiers, they captured the temporal characteristics of the rumor detection features. The authors drew upon already labeled datasets (from Twitter (Castillo et al., 2011) and Weibo (Wu et al., 2015)) to train and evaluate the approach.

The most recent work in this domain has been done by Zellers et al. (2019) who have proposed a framework called *GROVER* to generate and detect neural fake news as an adversarial game where neural fake news are fake news generated with recent AI technologies. First, the authors built a tool to generate fake viral or persuasive stories that were hard for humans to distinguish from real news. Second, they used the previous model to build a classifier for identifying real news from fake stories. The first part of the work was a text-generation task for language modeling whereby the authors used specific fields (such as the domain, date, author, headline, and body) as an input to the language model to generate fake news propaganda. Using Amazon Mechanical Turk, the results of the disinformation generation were evaluated. Overall, the system generated fake news had high trustworthiness and, as a result, humans had a hard time distinguishing them from real news. In the next step, the authors used the automatically generated neural fake news to train a linear classifier that identified human from machine (fake) written text. The classifier identified human from machine-written stories with an accuracy of 92%.

[2] http://weibo.com

[3] http://weibo.com/weibopiyas

Another research area that can fall under the same umbrella of misinformation detection using a classification-based approach is detecting troll profiles (opinion spam detection) in social media platforms. The work of Galán-García et al. (2016) and Gupta and Kaushal (2015) offered among the first approaches attempting to detect spammers in social media. Aside from spammer detection, detecting bots is an active research area in this domain. Users play an important role, as misinformation sometimes originates from automated accounts working in synchrony – bot nets. Shao et al. (2017) claimed that millions of political tweets were spread this way during and following the 2016 U.S. Presidential election. Shao et al. (2016) introduced a tool to detect such bots known as "Hoaxy" (Shao et al., 2016) to track the spread of claims.

Filippo Menczer et al. have been actively working in the field of detecting social bots (Davis et al., 2016; Ferrara et al., 2016; Varol et al., 2017; Shao et al., 2017; Suárez-Serrato et al., 2016; Shao et al., 2018; Dong and Liu, 2018; Lou et al., 2019). In their work, Davis et al. (2016) introduced the tool called "Botometer" which is a publicly-available service that uses more than one thousand features to evaluate the extent to which a Twitter account exhibits similarity to the known characteristics of social bots. Later, Varol et al. (2017) used the "Botometer" to look deeper into social bots characteristics. Authors found out that social bots exhibit a human-like behavior (retweets, mentions etc.) to target human accounts. They, further distinguished several social bots types such as spammers, self promoters, and accounts that post content from connected applications. Social bots do not only aim to manipulate discussions, alter the popularity of users and pollute content for political propaganda (Suárez-Serrato et al., 2016), Shao et al. (2017) found that they play an effective role in mitigating the spread of online misinformation. Social bots amplify low-quality content and target users with a large number of followers (Shao et al., 2018).

Case studies of incidents such as the Ukrainian conflict (Khaldarova and Pantti, 2016) and mass shootings in the U.S. (Starbird, 2017) have examined human reactions to questionable information online. The provenance and motivation behind such information has increasingly become a contentious issue, as many experts have speculated that important political decisions, including the United Kingdom vote to leave the European Union, and the election of Donald Trump to the U.S. presidency, were potentially swayed by forces outside those nations (Hern, 2017), posing a danger to democracy itself.

Chen et al. (2015) looked at "clickbaiting" behavior as a form of deception where the main purpose is to encourage people to click on links regardless of their quality and credibility. Authors built classifier models to automatically detect clickbait using textual as well as non-textual features. Authors suggest that a hybrid approach may yield best results.

The work conducted in Chapters 3 and 4 uses classification-based approaches to detect tweets' credibility and identify users spreading rumors, respectively. Similar to the works proposed in the literature (Castillo et al., 2011; Gayo-Avello et al., 2013; Kang et al., 2012; Gupta et al., 2014), we used supervised learning techniques with labeled datasets to build classification models. We also used similar suggested features for building a classifier to access tweets' credibility (*message-based*, *user-based*). Unlike prior work, we introduce *medical-based* features to better detect health-related misinformation about medical topics in Twitter. In Chapter 4, instead of identifying the credibility of Twitter messages, we identify the probability of users posting about a fake fact in the future by looking at their historical timeline. This will help identify who spreads rumors and might be a tool to predict future potential rumor topic.

Propagation-Based Approaches

The other popular method to identify the credibility of information in social medial platforms is the use of propagation/network features. This approach considers the network structure (retweet activity, followers, and followees) as well as the trust propagation in the network.

Seo et al. (2012) proposed an approach for assessing the probability that a piece of information is a rumor and for detecting rumor sources. The authors modeled the social network as a directed graph (vertices are individuals, directed edges are information flow). Then, they relied on the use of network monitor nodes, where detecting rumors and their sources relied on which monitors received the information and which did not. Computing specific metrics about the monitor nodes, the authors sort the nodes in the network; the top node was the source of the rumor. The authors show that with a sufficient number of monitor nodes, it is possible to recognize most rumors and their sources with high accuracy.

Gupta et al. (2012) built a graph from users, tweets, and events to identify the credibility of an event. At each iteration, every node shared the value of credibility (computed by a classifier) to its direct neighbors. As the number of iterations increased, and with the help of propagation, the credibility values of credible sources increase while the value of non-credible sources decreased. The authors enhanced the suggested model by introducing another tool built on a graph of events (verticies) with edges as credibility values where similar events had similar credibility scores. The authors claim that this method outperforms the classifierbased approach discussed earlier.

Ratkiewicz et al. (2011) presented the real-time "Truthy" online web service, designed to detect fake political grassroots movements (dubbed "AstroTurf") in the context of

U.S. political elections. The authors used mining, visualizing, mapping, classifying, and modeling massive streams of public microblogging events.

The work of Zhao et al. (2016a) is an example of trust modeling whereby the authors evaluated the trustworthiness of each tweet and each user. To achieve this, the authors relied on the idea that a tweet is credible if its features are similar to those of a tweet from a trustworthy news sources. Then, by means of four propagation rules defined on the social graph, the trustworthiness of tweets and users was refined and propagated.

Survey-Based Studies

Survey-based approaches rely on conducting a questionnaire and/or executing tasks to shed light on the contribution that credibility factors make in the overall process of credibility judgment by users of microblogs in general. Examples are the large-scale online user surveys conducted on Amazon Mechanical Turk that offer direct ways to measure the credibility of information (Sikdar et al., 2013).

Morris et al. (2012) conducted a survey to understand the possible factors influencing the credibility assessment of tweets. The authors manipulated a set of features, and the results showed that people pay attention to user names or tweet topics more than to profile picture images when making credibility assessments.

Among the features that affect the credibility assessment of social media messages are cultural differences. Yang et al. (2013) pointed out that people in China tend to trust social media content more than people in the United States.

Sikdar et al. (2013) argued that survey-based studies used to evaluate the credibility of information in microblogging websites are extremely noisy due to the small amount of information with which one can reliably assess credibility. To prove this point, the authors designed two different surveys about the same tweets and asked participants to evaluate their credibility. They varied information related to the tweets, such as the retweeting count, the time of the message, etc. Results showed that the outcomes of a survey-based approach can be unreliable and unstable.

2.1.2 Online Search

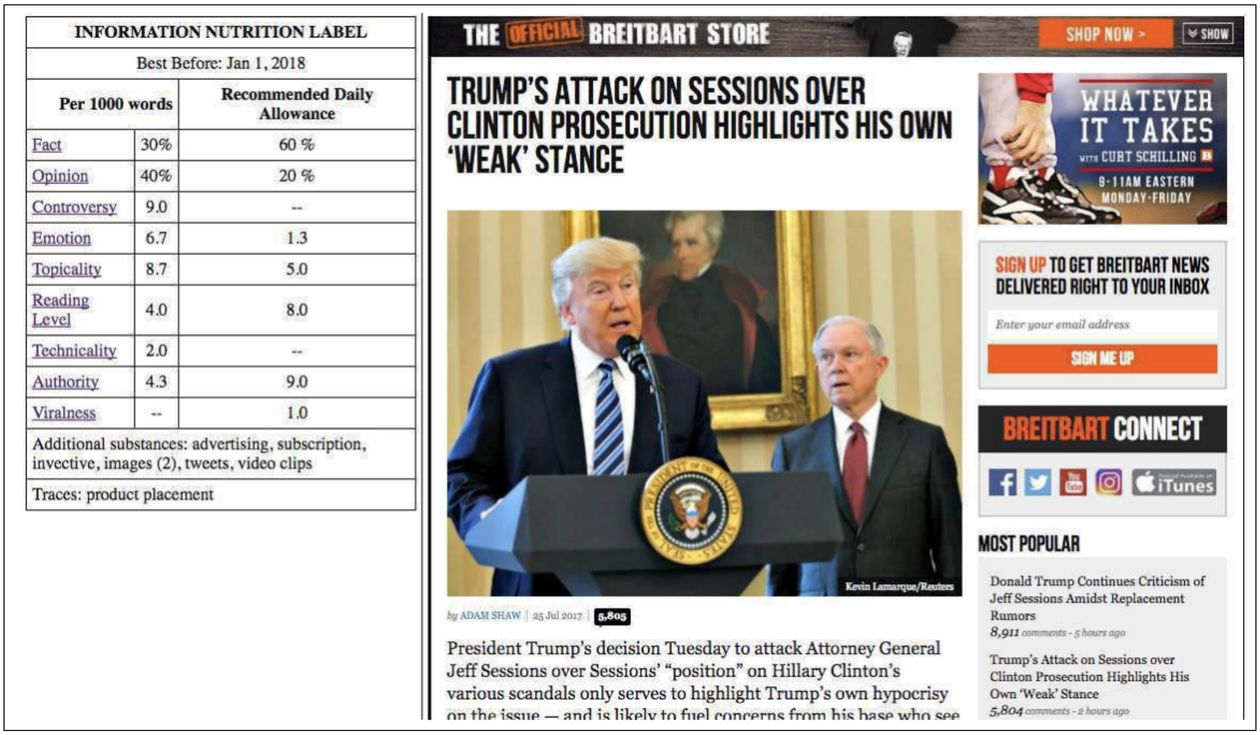
To help people distinguish between credible and fake sources, many studies have focused on identifying measures to detect misinformation and distinguish it from credible facts in search result pages.

For example, Aker et al. (2019) evaluated the writing style of 250 news articles by manually annotating the subjectivity scores at the article level. Results showed that articles containing fake news have significantly more subjective language than credible sources. The authors shared the news articles corpus with their subjectivity scores to the research community which can be used as a signal for automatically detecting and tracking misinformation in search result pages.

Kumar et al. (2016) looked at how Wikipedia pages can present false information. Focusing on hoax articles, the authors measured the real-world impact of such articles by looking at their survival period before they were debunked, the number of pageviews, and the amount of citations they received in online content. Kumar et al. (2016) found that even though a large portion of hoax articles are quickly deleted, a small amount survive for a long period and are cited on the Web. Further, hoax Wikipedia articles differ from correct articles in many ways, such as their structure, editors ,and embeddedness in Wikipedia. These findings can be applied to build automatic classification models to detect hoax content on the Web which could be particularly useful because humans are not good at identifying fake online information.

Even search in online social media platforms has been shown to introduce biases that might lead to incorrect information. In this regard, Kulshrestha et al. (2017) proposed a method for studying political information bias in search when using social media platforms. The authors investigated the effect of information bias in the Twitter search engine. Specifically, they introduced different bias aspects into the Twitter search system, such as query bias based on terms relevant to the query, output bias (cumulative bias introduced by ranked lists), and ranking bias (bias introduced by the ranking system). Their results showed that input and ranking bias (such as query topic or how the query is phrased) play an important role in producing biased search results.

Using natural-language processing tools, Fuhr et al. (2017) introduced matrices that can be computed automatically to evaluate the information quality in search results. Named the "information nutrition label" (see Figure 2.1 for an example), this measure includes the following criteria: factuality, readability, virality, emotion, opinion, controversy, authority / credibility / trust, technicality, and topicality. The suggested measures in this work can be used to automatically detect misinformation in online content. The work proposed in Chapter 3 uses some of the measures suggested in this work (readability and emotion) to measure content credibility of Twitter messages regarding the Zika outbreak. While the work of Fuhr et al. (2017) is geared towards general misinformation detection, the work in Chapter 3 focuses on measures to detect health-related misinformation. Specifically, in Chapter 3, in addition to general credibility measures, we introduce medically oriented matrices (such as a medical lexicon) to automatically compute the veracity of messages



INFORMATION NUTRITION LABEL		
Best Before: Jan 1, 2018		
Per 1000 words		Recommended Daily Allowance
Fact	30%	60 %
Opinion	40%	20 %
Controversy	9.0	--
Emotion	6.7	1.3
Topicality	8.7	5.0
Reading Level	4.0	8.0
Technicality	2.0	--
Authority	4.3	9.0
Viralness	--	1.0
Additional substances: advertising, subscription, invective, images (2), tweets, video clips		
Traces: product placement		

THE OFFICIAL BREITBART STORE SHOP NOW > SHOW

TRUMP'S ATTACK ON SESSIONS OVER CLINTON PROSECUTION HIGHLIGHTS HIS OWN 'WEAK' STANCE

Kevin Lamarque/Reuters

by ADAM SHAW | 25 Jul 2017 | 5,805

President Trump's decision Tuesday to attack Attorney General Jeff Sessions over Sessions' "position" on Hillary Clinton's various scandals only serves to highlight Trump's own hypocrisy on the issue — and is likely to fuel concerns from his base who see

WHATEVER IT TAKES WITH CURT SCHILLING 9-11AM EASTERN MONDAY-FRIDAY

SIGN UP TO GET BREITBART NEWS DELIVERED RIGHT TO YOUR INBOX

Enter your email address

SIGN ME UP

BREITBART CONNECT

MOST POPULAR

Donald Trump Continues Criticism of Jeff Sessions Amidst Replacement Rumors
8,911 comments - 5 hours ago

Trump's Attack on Sessions over Clinton Prosecution Highlights His Own 'Weak' Stance
5,804 comments - 2 hours ago

Figure 2.1: Information nutrition label computed for a news article (Fuhr et al., 2017).

between Twitter users.

The most recent and influential work in the political domain is a study by Epstein and Robertson (2015). The authors proposed a large-scale controlled-environment experiment to understand the influence of search results' rank on people's election preferences. Using search engine logs analysis and different statistical significance tests, they showed that people's voting preferences can be significantly shifted even in the case of undecided voters. As the study was large and involved diverse demographic characteristics (participants from USA and India), it is possible that the search engine effect was different for different groups. The authors also implemented different search ranking masking techniques (slightly swapped order of results in two specific positions) and concluded that even when the search rank pattern is hidden, people are still influenced by the search engine manipulation effect (Epstein and Robertson, 2015). This work is similar to ours in Chapter 5, as they are both controlled studies in which participants' behavior is interpreted by a set of log files. The difference between the two experiments is that Epstein and Robertson (2015)'s study focused on the political domain, while we explore the health domain. Additionally, our work computes the exact positive and negative influences of search results on participants' decisions about the efficacy of a set of medical treatments.

2.2 Online Health Information

More relevant to the research proposed in this thesis is the prior work conducted about online health information. People find online health information using various methods, including online health communities, mobile applications, etc. (Leavitt and Robinson, 2017). Because online search is the most commonly used tool for finding online health information (Lee et al., 2014), we focus on how people do online search for health-related topics. We further explore how social media platforms are being used in the health domain. We focus on these specific online resources as they are considered among the most heavily used online platforms (Purcell et al., 2012).

The work conducted by De Choudhury et al. (2014) is very relevant to this PhD research as the authors looked at how seeking and sharing information in terms of health activities is different in web search and in social media. Results of their survey data showed that the most common usage of search engine and social media health activities is to look for information about treatments. Further, people tend to share health information online to promote their health status and share news. Results show that people are motivated to use search engines and social media because of the ease and speed these tools provide while they tend to share information in Twitter to reach a large audience and benefit others. More interestingly, looking at search engine logs and Twitter data, the authors concluded that people tend to look for information about severe medical conditions more often in search engines than on social media. On the other hand, Twitter is more popular for seeking information about specific medical symptoms. Looking at social stigma, the authors found that search engines are more frequently used for high-stigma health topics, while Twitter is used for low-stigma issues. Authors argue that this is due to the fear of being judged on social medial platforms so people tend to avoid revealing sensitive information about their health. Finally, the authors noticed that the context used when searching for severe and high stigma health conditions is different than the one used for benign and low stigma issues. Both the work of De Choudhury et al. (2014) and the work proposed in this thesis focus on search engines and social medial platforms. However, while De Choudhury et al. (2014) focused on general health-related activities in search engines and social media, our proposed work focuses on the effects of health misinformation in those two platforms.

In the coming sections, we will first explain the work conducted about health search by listing examples of existing applications, evaluation of online content quality, and possible search biases. Then we will detail the work conducted on social media health usage including health-related attitudes and health misinformation in social media platforms.

2.3 Health in Online Web Search

Studying online health search activities includes studying the possible applications of using online search activities (such as predicting drugs side effects and a diagnosis tool for health conditions). Prior work studied the quality of web content, trust in health-related search results, and search behavior biases. We will deal with each of these in turn in the coming sections.

2.3.1 Health Applications Using Online Search

The literature is full about the work conducted on health online search. In this section, we indicate some of the notable work on using online search for health applications, such as pharmaceuticals, medical diagnosis, and monitoring tools.

First, in the **pharmaceutical domain**, Yom-Tov and Gabrilovich (2013), White et al. (2016), and Odgers et al. (2014) used a large set of web search logs to predict the side effects of drugs. Using people's search queries for side effects after being sick, the authors built a model to predict new side effects of different drugs that might take longer period of time to be discovered.

Yom-Tov (2017) used search engine queries to predict drug recalls. The author extracted queries located in the USA using the Bing search engine that mentioned 5,195 pharmaceutical drugs during the year of 2015 and all recall notifications issued by the Food and Drug Administration (FDA) during that year. Then, using the attributes that quantify changes in query volume (such as queries about drug symptoms and drug query spikes), the author built a model to predict whether a recall of a drug would be ordered by the FDA within one to 40 days. The model successfully predicted future drug recalls with an accuracy of 79.1%. The most predictive feature for future recall is a sudden spike in drug querying in every state. The findings of this work suggest that search logs can be used for early detection of harmful batches of medicines.

Yom-Tov and Lev-Ran (2017) used search engine queries to predict adverse reactions associated with cannabis consumption. Studying Cannabis-associated reactions is difficult because the substance is prohibited in many countries. For this reason, Yom-Tov and Lev-Ran (2017) collected US-based Bing queries over a six-month period from people making queries about cannabis (as well as 121 synonyms). Then, they compared the queries with the prevalence of cannabis use reported by the US National Survey on Drug Use in the Household. Their results showed that the search queries were correlated with the reported case of cannabis usage. Further, the queries revealed many of the adverse effects of cannabis reported in the medical literature.

Gahr et al. (2015) gathered search logs from Google trends and annual prescription volumes (APV) from statutory health insurance in Germany from 2004 to 2013. The aim was to determine whether there was any correlation between the APV of antidepressants and the web search query data. Their results indicated a significant and strong correlation between the APV and the annual search query volume for different antidepressants. The results of this study showed that there is the potential to predict the volume of prescription practice through web search query volume.

In addition to tools targeting the pharmaceutical field, search logs have been used for **general public health monitoring** to identify potential health outbreaks (Yom-Tov, 2015; Yom-Tov et al., 2014a). For example, Yom-Tov et al. (2014a) designed a study that used search engine queries as well as Twitter data to automatically detect potential outbreaks of communicable diseases. Yom-Tov et al. (2014a) looked at potential symptoms consistent between different online sources during major musical festivals. Their results showed that online sources might be indicative of a disease that some users attributed to being at the festival.

White and Horvitz (2009, 2012) focused on understanding how people use web search as a self-diagnosis tool. They explored the benefits of health-related web content for increasing knowledge, while also identifying the potential harmful effects of *cyberchondria* — "the unfounded escalation of concerns about common symptomatology, based on the review of search results and literature on the Web" (White and Horvitz, 2009, 2012). White and Horvitz (2009) showed that the anxiety level of the searcher can become more intense when web search is used to interpret symptoms for undiagnosed conditions. Using similar techniques, they studied how this anxiety persists and influences future behavior (White and Horvitz, 2012). The heightened anxiety induced by medical web results has a negative impact on the searcher's health.

Hochberg et al. (2019) looked at whether early diagnosis of diabetes is possible using search data. The authors collected queries made by people in the USA using Bing during a one-year time period that contained symptoms of diabetes. Four different predictive models were built (linear regression, logistic regression, decision tree and random forest) to distinguish between users who mentioned that they had been diagnosed with diabetes and people who did not refer to diabetes in their queries. The logistic regression and random forest models were able to distinguish between people diagnosed with diabetes and people not diagnosed with diabetes with an accuracy of 92%. The suggested model could identify the patients up to 240 days before they mentioned being diagnosed.

Soldaini and Yom-Tov (2017) used statistical information about the entire population generated from search logs as well as small sets of labeled examples to label unseen examples. This model was then used to identify users who might be suffering from cancer by looking at their search patterns. The same model can be used to predict the spread of disease; partial epidemiological data is used, showing the distribution of disease given the incidence in a subset of a population.

Agarwal et al. (2016) used geotagged mobile search logs to predict potential future patient visits to a medical facility. Gathering more than 42 days of search logs, the authors constructed a matrix of general, semantic, and location-based features. They then trained a random forest classifier to predict users' future visits with the help of advertising techniques.

2.3.2 Health Information Credibility in Online Search

For over a decade, medical sociologists have studied the Internet as a new component of health ecosystems (Chen and Siu, 2001). Patients' reliance on the Web has resulted in a patient–Web–physician "triangulation" (Wald et al., 2007), with benefits such as more efficient use of clinical time and additional support from online support groups, coupled with potential harms such as the dangers posed by the variable quality of information, unnecessary visits to a physician, and exacerbating existing socioeconomic health disparities. Early on, the use of the Internet by patients was shown to be problematic due to the questionable quality of content (Schmidt and Ernst, 2004).

The credibility of online health information has been an active topic of interest to researchers. Studies about the correctness of Internet content have reported that the quality of online health information is low or incomplete. This erroneous and incomplete information creates difficulties for users trying to navigate the search for truthful content and make informed decisions (Kaicker et al., 2013), thus creating potential harm to users (Benigeri and Pluye, 2003). Here we list examples of studies attempting to understand how people evaluate search results' quality.

A large amount of work has shown that specific features of a website can increase the perceived credibility of information in the health domain, such as a professional-appearing interface design, ease of navigation, and the presence of endorsements (Thompson, 2014).

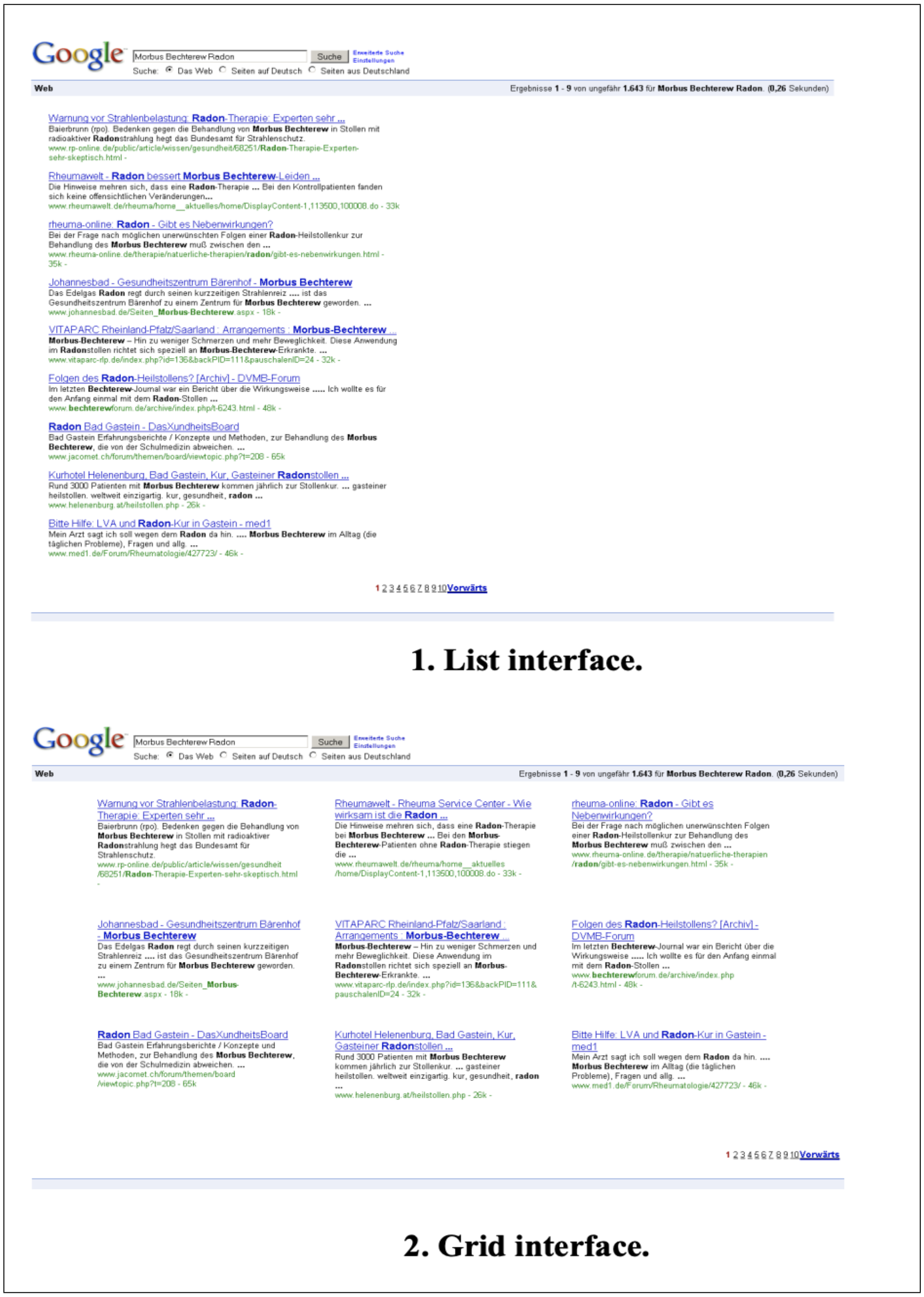
Freeman and Spyridakis (2004) were among the first to summarize the complete list of factors that influence how users judge the credibility of health-related websites.

Dhoju et al. (2019) looked at a collection of 44,064 health-related news articles, published by media outlets, to understand the differences between reliable and unreliable content spanning from 1 January 2015 to 2 April 2018, using Facebook Graph API. The authors found out that unreliable sources use clickbait headlines to catch users' attention (27% of the headlines are clickbait headlines in unreliable sources, compared to 40% clickbait headlines in reliable sources). Further, the use of fewer quotations and hyperlinks is associated with unreliable sources. These signals are helpful to automatically identify health disinformation.

Elsweiler and Kattenbeck (2019) designed a think-aloud user study to shed light on how people assess the credibility of search result pages. Their findings showed that people are not certain when assessing the credibility of online sources. People use ten different cues to access the credibility of sources, and the usage of these cues differs for each participant and each topic.

Jung et al. (2016) looked at potential factors influencing the perceived credibility of diet-nutrition information on web sites. The authors designed an online experiment with 575 subjects to measure the effect of source expertise cues and message accuracy on people's perception of credibility. Their findings showed that information accuracy increased the perceived credibility regardless of the level of source expertise. Furthermore, when readers had low prior knowledge, source expertise was an important factor affecting the their perception of credibility. Message accuracy had a higher impact on people who had experience with the nutrition issue compared to people who were not involved in the diet issue. Health practitioners should consider such findings when designing online content for users.

Kammerer et al. (2013) proposed a controlled laboratory study to understand the behavior and decision making of people when they evaluated web search sources about specific medical issues. The authors selected two different treatments for a certain health issue. Then they biased search results about the treatments by controlling for different source credibility levels (medical institutions, journals, forums). Later, the authors asked participants to evaluate which treatment was better. Using eye tracking, participants' logs, and verbal protocols, they found that people spend less time and effort evaluating search results when information sources seem accurate and reliable. Furthermore, people tend to be more certain and require less justifications of information when the source is trusted.



Morbus Bechterew Radon

Suche: ● Das Web ○ Seiten auf Deutsch ○ Seiten aus Deutschland

Web Ergebnisse **1** - **9** von ungefähr **1.643** für **Morbus Bechterew Radon**. (**0,26** Sekunden)

Warnung vor Strahlenbelastung: **Radon**-Therapie: Experten sehr ...
Baierbrunn (rpo). Bedenken gegen die Behandlung von **Morbus Bechterew** in Stollen mit radioaktiver **Radon**strahlung hegt das Bundesamt für Strahlenschutz.
www.rp-online.de/public/article/wissen/gesundheit/68251/**Radon**-Therapie-Experten-sehr-skeptisch.html -

Rheumawelt - **Radon** bessert **Morbus Bechterew**-Leiden ...
Die Hinweise mehren sich, dass eine **Radon**-Therapie **...** Bei den Kontrollpatienten fanden sich keine offensichtlichen Veränderungen**...**
www.rheumawelt.de/rheuma/home__aktuelles/home/DisplayContent-1,113500,100008.do - 33k

rheuma-online: **Radon** - Gibt es Nebenwirkungen?
Bei der Frage nach möglichen unerwünschten Folgen einer **Radon**-Heilstollenkur zur Behandlung des **Morbus Bechterew** muß zwischen den **...**
www.rheuma-online.de/therapie/natuerliche-therapien/**radon**/gibt-es-nebenwirkungen.html - 35k -

Johannesbad - Gesundheitszentrum Bärenhof - **Morbus Bechterew**
Das Edelgas **Radon** regt durch seinen kurzzeitigen Strahlenreiz **....** ist das Gesundheitszentrum Bärenhof zu einem Zentrum für **Morbus Bechterew** geworden. **...**
www.johannesbad.de/Seiten_**Morbus-Bechterew**.aspx - 18k -

VITAPARC Rheinland-Pfalz/Saarland : Arrangements : **Morbus-Bechterew** ...
Morbus-Bechterew – Hin zu weniger Schmerzen und mehr Beweglichkeit. Diese Anwendung im **Radon**stollen richtet sich speziell an **Morbus-Bechterew**-Erkrankte. **...**
www.vitaparc-rlp.de/index.php?id=136&backPID=111&pauschalenID=24 - 32k -

Folgen des **Radon**-Heilstollens? [Archiv] - DVMB-Forum
Im letzten **Bechterew**-Journal war ein Bericht über die Wirkungsweise **.....** Ich wollte es für den Anfang einmal mit dem **Radon**-Stollen **...**
www.**bechterew**forum.de/archive/index.php/t-6243.html - 48k -

Radon Bad Gastein - DasXundheitsBoard
Bad Gastein Erfahrungsberichte / Konzepte und Methoden, zur Behandlung des **Morbus Bechterew**, die von der Schulmedizin abweichen. **...**
www.jacomet.ch/forum/themen/board/viewtopic.php?t=208 - 65k

Kurhotel Helenenburg, Bad Gastein, Kur, Gasteiner **Radon**stollen ...
Rund 3000 Patienten mit **Morbus Bechterew** kommen jährlich zur Stollenkur. **...** gasteiner heilstollen. weltweit einzigartig. kur, gesundheit, **radon** **...**
www.helenenburg.at/heilstollen.php - 26k -

Bitte Hilfe: LVA und **Radon**-Kur in Gastein - med1
Mein Arzt sagt ich soll wegen dem **Radon** da hin. **....** **Morbus Bechterew** im Alltag (die täglichen Probleme), Fragen und allg. **...**
www.med1.de/Forum/Rheumatologie/427723/ - 46k -

1. List interface.

Morbus Bechterew Radon

Suche: ● Das Web ○ Seiten auf Deutsch ○ Seiten aus Deutschland

Web Ergebnisse **1** - **9** von ungefähr **1.643** für **Morbus Bechterew Radon**. (**0,26** Sekunden)

Warnung vor Strahlenbelastung: **Radon**-Therapie: Experten sehr ...
Baierbrunn (rpo). Bedenken gegen die Behandlung von **Morbus Bechterew** in Stollen mit radioaktiver **Radon**strahlung hegt das Bundesamt für Strahlenschutz.
www.rp-online.de/public/article/wissen/gesundheit/68251/**Radon**-Therapie-Experten-sehr-skeptisch.html -

Rheumawelt - Rheuma Service Center - Wie wirksam ist die **Radon** ...
Die Hinweise mehren sich, dass eine **Radon**-Therapie bei **Morbus Bechterew** **...** Bei den **Morbus-Bechterew**-Patienten ohne **Radon**-Therapie stiegen die **...**
www.rheumawelt.de/rheuma/home__aktuelles/home/DisplayContent-1,113500,100008.do - 33k -

rheuma-online: **Radon** - Gibt es Nebenwirkungen?
Bei der Frage nach möglichen unerwünschten Folgen einer **Radon**-Heilstollenkur zur Behandlung des **Morbus Bechterew** muß zwischen den **...**
www.rheuma-online.de/therapie/natuerliche-therapien/**radon**/gibt-es-nebenwirkungen.html - 35k -

Johannesbad - Gesundheitszentrum Bärenhof - **Morbus Bechterew**
Das Edelgas **Radon** regt durch seinen kurzzeitigen Strahlenreiz **....** ist das Gesundheitszentrum Bärenhof zu einem Zentrum für **Morbus Bechterew** geworden. **...**
www.johannesbad.de/Seiten_**Morbus-Bechterew**.aspx - 18k -

VITAPARC Rheinland-Pfalz/Saarland : Arrangements : **Morbus-Bechterew** ...
Morbus-Bechterew – Hin zu weniger Schmerzen und mehr Beweglichkeit. Diese Anwendung im **Radon**stollen richtet sich speziell an **Morbus-Bechterew**-Erkrankte. **...**
www.vitaparc-rlp.de/index.php?id=136&backPID=111&pauschalenID=24 - 32k -

Folgen des **Radon**-Heilstollens? [Archiv] - DVMB-Forum
Im letzten **Bechterew**-Journal war ein Bericht über die Wirkungsweise **.....** Ich wollte es für den Anfang einmal mit dem **Radon**-Stollen **...**
www.**bechterew**forum.de/archive/index.php/t-6243.html - 48k -

Radon Bad Gastein - DasXundheitsBoard
Bad Gastein Erfahrungsberichte / Konzepte und Methoden, zur Behandlung des **Morbus Bechterew**, die von der Schulmedizin abweichen. **...**
www.jacomet.ch/forum/themen/board/viewtopic.php?t=208 - 65k

Kurhotel Helenenburg, Bad Gastein, Kur, Gasteiner **Radon**stollen ...
Rund 3000 Patienten mit **Morbus Bechterew** kommen jährlich zur Stollenkur. **...** gasteiner heilstollen. weltweit einzigartig. kur, gesundheit, **radon** **...**
www.helenenburg.at/heilstollen.php - 26k -

Bitte Hilfe: LVA und **Radon**-Kur in Gastein - med1
Mein Arzt sagt ich soll wegen dem **Radon** da hin. **....** **Morbus Bechterew** im Alltag (die täglichen Probleme), Fragen und allg. **...**
www.med1.de/Forum/Rheumatologie/427723/ - 46k -

2. Grid interface.

Figure 2.2: (1) List versus (2) Grid interface (Kammerer and Gerjets, 2010).

Kammerer and Gerjets (2010) designed a laboratory study to measure the effect of search results' interface design on people's ability to evaluate the trustworthiness of search results. More specifically, the authors looked at how people evaluate the trustworthiness of search results for the standard list search results format (search results presented in a list) versus the grid format (search results presented in the form of a grid). Figure 2.2 shows the list and grid interfaces. In total, the study had four different experimental conditions with two main factors: the order of trustworthiness of the search engine results page (SERP) in ascending and descending order, and the interface design as list and grid. In terms of measuring the interaction with SERP pages, the authors measured eye movements as well as number of clicks. The study results showed that when dealing with a list interface, participants had a more homogeneous linear viewing sequence (i.e., viewing from top to bottom in the list interface or from one column to the other in the grid interface), and less linear, more varied viewing in the grid interface. Furthermore, participants paid more attention to highly ranked results in the list interface, compared to an equal spread of attention with the grid interface. As a result, in the experimental condition where a list was provided in ascending order of trust, participants interacted significantly more with the least trustworthy pages than with the most trustworthy ones. This behavior could be harmful, especially if participants are dealing with controversial medical topics and trustworthiness is important, but the best answers might not be found quickly in the top-ranked results (Kammerer and Gerjets, 2010). Based on this study, Kammerer and Gerjets (2010) claimed that a grid interface is a better setting for supporting users to evaluate the trustworthiness of search page results, as it provides more freedom and exposure to varied content.

Pam Briggs et al. conducted a significant amount of work looking at **trust in online health information** (Sillence et al., 2007, 2006, 2004). Specifically, Sillence et al. (2007, 2006) designed a three-stage model of trust when searching for information online. Sillence et al. (2007) conducted a "think-aloud" user experiment with menopausal patients, while Sillence et al. (2006, 2004) used a larger-scale experiment. Both studies showed the following: participants rejected sales sites as well as low-quality design content even though they were legitimate sources. Second, when looking at high-quality designed sites, participants trusted content coming from medical institutions or health experts but also personalized content from people similar to the searchers. Third, participants' decision-making process was influenced by online information: they used online content to reinforce a decision they had already made to find supporting facts and build confidence about their decisions (Sillence et al., 2007).

Other researchers have studied **the effect of credibility on people's decisions**. Lau and Coiera (2008) designed a controlled laboratory study to understand whether providing high quality search results improves people's accuracy when searching online for health information. Results showed that, when participants were provided with high-quality search results from reliable sources (such as PubMed, MedlinePlus, and HealthInsite), the participants' accuracy in answering health questions increased compared to when they were not provided with search results. Both this work and the work conducted in Chapters 5 and 6 aim to measure people's accuracy when using online search to answer health-related questions. While Lau and Coiera (2008)'s work focused on measuring the change in accuracy before and after searching, we focus on measuring people's accuracy when search results pages have been biased.

2.3.3 Search Biases

As search engines apply information retrieval algorithms to return the most relevant documents, the search result pages come with a number of algorithmic biases. In this section, we first present work about search results bias, then we list the literature about search behavior bias.

Fu et al. (2016) evaluated **the quality** and rank of search result pages when the efficacy of HPV vaccination was being researched. The authors collected 116 Google search engine pages using 20 terms to search for the HPV vaccine. Then they measured the web page bias by manually annotating the pages' content to either being critical (stating concerns about vaccines) or non-critical. Web pages that included content with a bias against vaccination were categorized as overall noncritical if they allotted roughly equal space or more to viewpoints supportive of vaccination. Web pages that presented only evidence-based content without editorial comment regardless of the focus (e.g., vaccine side effects) were also categorized as noncritical. Authors measured the web page quality by looking at the Journal of the American Medical Association (JAMA) benchmarks which were the basis for the present study's assessment of HPV Web page quality. Their results showed that search engine returned more frequent pages with low-quality scores that were critical about vaccines. Tang et al. (2006) compared the search results of Google to those of a domain-specific health and depression search engine. Their findings showed that while Google returned more relevant documents, the domain-specific search engine returned more correct search result pages.

Another study conducted by Lau and Coiera (2009, 2007), looked at the effect of **rank** on people's accuracy in answering health-related questions when using online search. To test the effect of rank, the authors introduced three different online search interfaces: one baseline user interface and two modified interfaces specifically designed to debias the anchoring or order effect. Their results showed that people were influenced by order bias when using the baseline interface. On the other hand, the order effect disappeared with the use of a debiased interface, but no improvement was seen in participants' accuracy. Further, participants preferred using a debiased interface whereby they conducted fewer searches and accessed more documents.

Coiera and Vickland (2008) determined that judging the relevance of search engine documents is not useful when searching for answers to health-related questions. Better search results (highly relevant) did not seem to increase people's ability to answer health-related questions. Coiera and Vickland (2008)'s work highlights the importance of introducing new natural metrics for search engines to evaluate and display search results to users.

Another study, conducted by Venkatraman et al. (2016), explored possible uncontrolled biases in search results. The authors analyzed the top 20 search results of two popular search engines (Google, Bing) for the search term "Zika virus." Manual labeling by medically trained experts showed that out of 20 pages, six results from the Google search engine contained incorrect information, whereas only one result was incorrect using the Bing search engine.

Looking at the topic of distal radius fractures, Dy et al. (2012) aimed to determine whether varying the search terms would affect the search results' quality, accuracy, and readability levels. Different from the work conducted by Fuhr et al. (2017), Dy found out that the readability score had no correlation with the quality or accuracy of the search result content. The authors further noted that the quality and accuracy of information presented in search results varies depending on the level of sophistication of the search terms. Similarly, Dy et al. (2012) found that the quality of the search results varied for differently worded medical queries.

Related to the potential biases in medical web pages as well as among searchers, the vast amount of work conducted by White (2013), White and Hassan (2014), White (2014), White and Horvitz (2015) showed that searchers, as well as search engines, strongly favor positive information over negative information regardless of the truth. In his work, White looked at two different types of questions: medical queries with a yes/no question form (White, 2013; White and Hassan, 2014) as well as medical queries about the efficacy of medical treatments (White, 2014; White and Horvitz, 2015).

When search results are biased towards one answer (yes/no) and results are all ranked above the contradictory answer (all yes above no or all no above yes), people choose the dominant answer (White, 2014). When results were biased towards the correct answer and all *yes* results were ranked above *no* results, the accuracy was 74.9%. However, when the bias was towards incorrect answer and all *no* results were ranked above the *yes* results, the accuracy was reduced to 63.1%. This work is different than what we present in Chapters 5 and 6 with respect to the type of search query: while White focused on yes/no questions, we designed the study to examine the efficacy of medical treatment questions. Further, the notion of rank is different: in White (2014)'s work, yes/no answers were ranked above or bellow yes/no answers. In our work, explained in Chapters 5 and 6, we biased the rank of the top most correct document.

White and Horvitz (2015) conducted work similar to the study explained in Chapter 5. However, while White focused on measuring the search accuracy in organic search, we were interested in measuring the search accuracy when the search results were biased with respect to correctness and rank. Specifically, White examined the effect of organic search as well as controlled search with 50/50 mixture of answers (White and Horvitz, 2015).

In their studies, White (2014) and White and Horvitz (2015) focused on measuring beliefs before and after searchers were exposed to search results. To measure changes in beliefs, the authors asked about prior beliefs before the search task. Their results showed that, if a searcher holds a strong belief before searching, they are unlikely to change their opinion following their search. Conversely, if the searcher is uncertain before the search, they are twice as likely to move towards a positive answer (White, 2013, 2014; White and Horvitz, 2015).

Similar to White (2013), Kayhan (2013) investigated whether people are biased towards believing positive information over negative information regardless of the truth. They also tried to implement different techniques to reduce positive bias, using recommendation (suggest content with negative information) and incorporation (incorporate negative results with a set of positive results). The authors concluded that incorporating negative information with the available positive results of search pages helped reduce the positive uncontrolled bias people have when searching for health-related information online.

Allam et al. (2014) looked at the influence of search results' selection and ranking on users' knowledge and beliefs about vaccination. Their first experiment involved three experimental conditions: the control condition (uncontrolled search results returned by Google), the pro biased condition (all websites pro vaccination from trusted sources) and the anti biased condition (all websites against vaccination from untrusted sources). Their results showed that participants exposed to only highly trusted pro-vaccination content experienced increased knowledge (acknowledged the importance of vaccines), while the anti-vaccination content caused greater concerns about vaccines. In the second experiment, in addition to the three previously explained experiential conditions, the authors added the following search results' pro-anti vaccination ratios: 4:6, 6:4, 8:2. Their results showed that knowledge was significantly affected by different exposure to anti vaccination versus pro-vaccination search results. Additionally, Google-offered websites (control) created as much fear of vaccination as any of the customized search engines. As the amount of anti-vaccination results increased, people gained less knowledge and more fear. Finally, having more trusted search results encouraged participants to explore more content in the search result pages. While Allam et al. (2014)'s proposed work focused on measuring knowledge and belief change, in Chapter 5, we are interested in measuring the accuracy of participants when answering questions. We also did not control for the trustworthiness of the search results when designing the study. However, we controlled for rank and ratio of correct to incorrect results.

Ludolph et al. (2016) investigated the effect of the content of the knowledge graph box tool available in Google search on people's beliefs about vaccination. The authors had three experimental conditions. The first one was the trustworthiness of the search results (comprehensive from WHO and non-comprehensive from Wikipedia). The second one was the presence or absence of warning indicating that one would encounter false information about vaccines on the Web. The third was the control condition (no warning and no trustworthiness bias). Their results showed that knowledge increased when the knowledge graph box contained comprehensive information compared to non-comprehensive information from Wikipedia. Furthermore, knowledge decreased when there was a false information warning in the comprehensive content. Third, the comprehensive content with a warning reduced fear, whereas the warning increased fear with non-comprehensive content. Fourth, people appreciated the benefits of vaccination when exposed to comprehensive information in the knowledge graph box. This study shows that the presence of reliable content in the knowledge graph box positively affects people's knowledge about vaccines, and that a small change in current search engines could lead to a valuable difference in online health information search.

Looking at the behavioral search bias, White et al. (2008) performed log-based analysis to investigate the different search behaviors corresponding to domain expertise. Their study showed that experts use different vocabulary and patterns of interaction when writing queries and selecting sources. Based on this work, the authors claimed that automatic detection of expertise may improve search results and aid non-expert users with query suggestions and site recommendations reflective of expert users.

Further, Cartright et al. (2011) studied searcher behavior from log files to automatically identify medically related search sessions. The authors claimed that identifying health search intentions automatically can help better understand health search behavior and thereby better support it (Cartright et al., 2011).

Finally, Lau et al. (2010, 2011) looked at the effect of social feedback while doing online search for health topics. The authors conducted a controlled study wherein they asked participants to answer health-related questions using online search before and after showing the majority answers of other participants. In addition to looking at changes in the final decision, the authors tracked the participants level of confidence throughout the study. Their results showed a statistically significant difference between answers before and after being exposed to social feedback. Participants were more likely to change their answer whenever it did not agree with the majority. This study provides an empirical evidence that social feedback is important for decision making in the health domain. In the next section, we will look more deeply into the influence of social feedback presents in social media platforms.

2.4 Health in Social Media

In the domain of using social media for health, previous work has focused on studying different health-related attitudes and behaviors using social media, such as suicidal ideation and mental well-being. Another area of research has involved understanding how incorrect health-related information is spread and propagated in social media sites. In this section, we will list examples of studies in these areas that are directly related to this PhD research.

2.4.1 Health-Related Attitudes on Social Media

As social media is heavily used on a daily basis, it captures the actions of millions of users. In the health domain specifically, social media provides ample resources for health-related decision making, capturing behaviors and attitudes impacting individual health. Facebook is the most utilized social media platform, followed by health-specific social media sites, and Twitter (Laranjo et al., 2015). Recent work has focused on using this content to extract health-related behaviors and attitudes.

Recently, Ginart et al. (2016) suggested building a classifier to detect the use of marijuana using Twitter. Attitudes toward legal drugs, including Xanax and Adderall, have been studied by Seaman and Giraud-Carrier (2016). Further, Yang and Yang (2013) used association mining of health communities to discover adverse effects of drug interactions. Likewise, behaviors related to lifestyle diseases such as type 2 diabetes and obesity have been tracked using Twitter (Abbar et al., 2015), Instagram (a photo sharing platform) (Mejova et al., 2015a), and Facebook (Araújo et al., 2017a), along with attitudes toward food and diet (Mejova et al., 2016). These studies often combine expert health knowledge with big data analytics (including machine learning, in the case of the work of Ginart et al. (2016)) to provide insight into attitudes captured in social media interactions.

A study of a community promoting anorexia on Flickr (another photo-sharing platform) (Yom-Tov et al., 2012) showed that attempts of anti-anorexia programs to infiltrate the community with intervention messages tagged with pro-anorexia tags were counterproductive in the long run (with users exposed to such attempts remaining in the pro-anorexia community longer). Further uses of social media to gauge the efficacy of health communication include a recent study about breast cancer mammography advisory on Twitter which found many users to be confused by it rather than approving of it (Nastasi et al., 2017).

Using social media to gather insights about its users' mental well-being is a common research endeavor (Amir et al., 2017, 2019; De Choudhury et al., 2016; Wongkoblap et al., 2018). De Choudhury et al. (2013) tracked significant signals in Twitter that help detect signs of depression in Twitter users. The authors used tweets to measure behavioral changes in people in terms of feelings, mood swings, and drug usage. Later, the authors built a classifier to detect people suffering from depression and predict whether someone was likely to have depression in the future (De Choudhury et al., 2013).

De Choudhury et al. (2016) focused on support communities in the Reddit as a data source and explored possible outcomes of getting involved in suicidal discussion such as self-attention focus, linguistic changes, and social engagement. They then suggested a statistical methodology to predict whether users, involved in mental health conversations, are at risk of suicidal ideation. Their results showed that the presence of some tokens in the post comments statistically significantly affect (either increase or decrease) the likelihood of posting from the mental health community to suicidal ideation community in Reddit. In other words, positive words in social media communities tend to push people away from suicidal ideation, while negative words tend to increase the likelihood of people thinking about suicide (De Choudhury et al., 2016).

Using the dataset collected in the work of De Choudhury et al. (2016), De Choudhury and Kıcıman (2017) further examined how much receiving social support in the form of comments on Reddit influenced the risk of suicidal ideation. The authors used the propensity score matching technique to measure the effect of social support language on individuals in mental health communities' likelihood of moving towards suicidal ideation. Studying social support with respect to a theoretical model in the clinical literature, the authors concluded that esteem as well as network support are helpful in reducing the risk of suicidal ideation.

Bagroy et al. (2017) studied the effect of social media on mental well-being in the college student population. The authors used propensity score matching to identify possible outcomes of using alcohol on college students in Reddit platform. Specifically, they looked at pre-defined measures of college success factors, such as peer-group interactions, negative academic outcome etc. Results showed that mentioning alcohol had a statistically significant effect on first-year college students' success. Further, the rate of posting/commenting on Reddit increased after mentioning alcohol usage, which is possibly due to social and networking reasons (Bagroy et al., 2017).

Newman et al. (2011) studied health-related information sharing behavior on Facebook. The authors conducted interviews about sharing health information in online health communities. Focusing on weight loss and diabetes, they showed that people share health information in social online communities to achieve a set of goals. Emotional support is considered the most common reason for participants to use online health communities, as they find sympathy, positive support, and fast responses. Next, sharing health information on Facebook caused a significant number of participants to feel more accountable about committing to a treatment plan that is visible to others. Furthermore, the findings of this study showed that people seek role models to follow in online health communities as a source of motivation for improving their health. Finally, the authors found that few Facebook users look for advice in online health communities. On the other hand, the study revealed a list of challenges participants faced when sharing health-related information. The most common challenge was impression management, where users find themselves stuck between sharing information to improve their health and how they wish to represent themselves within the community.

Finally, public awareness of health-related topics has recently been gauged through the advertising platforms provided by social media – for instance, using Facebook Advertising Manager to estimate the number of Facebook users interested in diabetes-related topics (Araújo et al., 2017b). Thus, as social media use increases, so does health-related discussion and information seeking on these platforms, allowing for large-scale analysis and tracking.

2.4.2 Health Misinformation in Social Media

WHO's white paper on risk communication urges the research community to“build capacity to quickly transform new information into usable, culturally-appropriate and easily understood risk communication resources that can be disseminated on multiple platforms," including social media (Organization et al., 2016). In this section, we will highlight the work conducted on evaluating health-related content credibility in social media, then we will detail examples of tools to detect and track health misinformation. Finally we will list examples of work on the possible implications of health misinformation spread via social media platforms.

Health Information Credibility in Social Media

Rapid, easy access to health information in social media raises a particular issue: direct access to medical information, and the absence of credible intermediaries, makes it hard to verify the information present in social media platforms. In this regard, we will describe work conducted on evaluating content credibility and automatically detecting information credibility in social media.

Some researchers have evaluated the quality of health-related content present in social media and have found a large amount of the content to be potentially harmful. Democratization of content publishing may be exacerbating quality concerns, as YouTube videos have been found to contain instances of the public display of harmful or unhealthy behaviors, promotion of tobacco to consumers, and distort on of policy and research funding agendas (Lau et al., 2012).

Further, Syed-Abdul et al. (2013) identified 29.3% of YouTube about anorexia to be pro-anorexia. While not constituting the majority of the material on this topic, pro-anorexia content is highly rated and favored by viewers.

Sharma et al. (2017) studied the popularity of Zika-related information on Facebook. Their results showed that the most popular posts about Zika on Facebook contained misinformation about the Zika virus. From the results, the authors concluded that people tend to view and share content about misleading/misguiding information more frequently than sharing useful content.

McGregor et al. (2014) evaluated glaucoma-related content explored by patients on five different social media platforms (the International Glaucoma Association forum, Facebook, Twitter, YouTube, and PatientOpinion.org.uk). A total of 3,785 items were collected, analyzed, and coded. The results showed that many unmoderated sites contained misleading information. More importantly, complementary therapies and treatments with poor evidential basis were more represented than evidence-based treatments. With the increase in the amount of misleading content comes an increased risk of exposure to incorrect information, which might pose a threat to the online community.

Lederman et al. (2014) looked at how people assess the credibility of health information in online health forums. Their results showed that when the message contains scientific references, people do not seek other users' opinions. However, when the information is subjective or contradicts reliable sources, people rely on the crowd's consensus to make their decisions.

It is important to note that credibility in social media is different than other credibility concepts because the original health information source does not have a great impact on the perceived information credibility. People might trust content generated by laypersons due to homophily as per Ma and Atkin (2017). This is what makes credibility assessment in health-related social media essential and challenging.

Despite the importance of credibility detection, few approaches have so far been proposed to automatically assess the credibility of healthrelated information in social media. Existing techniques involve the crawling method (Abbasi et al., 2013), examining social interactions in social media platforms (Weitzel et al., 2014), and the RetweetNetwork model (Freeman, 1978).

A notable work in this area has been conducted by Mukherjee et al. (2014), who have proposed a tool for automatically assessing the credibility of medical statements in health communities. The authors used linguistic as well as user-based features in an a Markov Random Field model where the random variables were constituted by users, their posts, and the medical statements contained within the posts. The suggested model showed better detection accuracy than other state-of-the-art baselines.

Health Misinformation Detection and Tracking in Social Media

In the previous sections, we have talked about misinformation detection, using social media in different domains such as news and politics. In this section, we specifically focus on misinformation in the health domain using social media platforms. The problem of misinformation in social media is important because the presence of fake, misleading health-related news is a potential threat for public health.

Waszak et al. (2018) quantified the amount of fake medical news in social media. Their results showed that around 40% of the most frequently shared links in social media were actually fake news (shared more than 450,000 times). The most widespread topic with fake content was about vaccines, while cardiovascular diseases were also well represented. Further, 20% of the fake content was shared by a single source. These findings highlight the importance of identifying fake medical information in social media.

The dynamic nature of social media platforms allows for the rapid spread of misinformation during an ongoing epidemic, and machine-learning techniques have been deployed to track such content. To help social media users, tools are being developed to ease the verification of health claims via natural language processing and the retrieval of reliable medical literature (Samuel and Zaıane, 2018).

Recently, Dredze et al. (2016) analyzed the characteristics of non-scientific claims about vaccines, spreading from the vaccine refusal community to the rest of the world. Specifically, the authors analyzed the two most prominent misleading theories about Zika vaccination on Twitter using supervised machine learning technique and observed the effect of vaccine-skeptical communities on other users' opinions about vaccination. While Dredze et al. (2016) looked at two Zika vaccine-related memes, in the work described in Chapter 3, we propose a more general methodological pipeline to track health-related rumors. Taking Zika as a case study and with the help of health professionals, we expand the list of rumors to six and examine the behavior of rumors, as well as clarification efforts.

Theng et al. (2013) designed a survey to understand how people in different geographical locations respond to health misinformation in social media. Specifically, the authors looked at 10 different tuberculosis myths and observed social media users' level of belief. Their results showed that age and country had a statistically significant effect on the level of belief in rumors. For example , younger people were more susceptible to rumors than older people. Further, the results showed that different social media platforms influenced the level of belief in rumors (Theng et al., 2013).

Kostkova et al. (2016) created the "VAC Medi+ board" online interactive visualization framework integrating heterogeneous real-time data streams with Twitter data. They tracked the spread of vaccine-related information on Twitter and the sources of information spread. Both of these systems rely on keyword-based rumor identification.

Kinsora et al. (2017) suggested building a medical misinformation-labeled dataset to help build machine learning classifiers that automatically detect fake medically related information. The dataset is a collection of misinformation and true facts gathered from comments in an online health forum. To achieve this, the authors used information retrieval techniques. Later, they designed a coding scheme to label and annotate the dataset. Using nine engineered features in the generated dataset, the authors built a classifier that can identify medical misinformation with an accuracy of 90.1%. The dataset proposed by Kinsora et al. (2017)'s work can be used to build more sophistical tools to identify medical misinformation in user-generated content systems such as medical forums.

A potential framework for engaging expert knowledge in a real-time crisis situation is described in Imran et al. (2014)'s work, where content was selected to be annotated via crowdsourcing into predefined classes. These could then be used to train a classifier, and update it as necessary with active learning data selection.

It is important to note that, in the literature, little attention has been paid to modeling and understanding the spread of cancer "complementary and alternative medicine" (CAM) on the Internet. A 2008 survey of 80 cancer patients found that when going online, respondents dealt with the emotional stress of being reminded about their prognosis, and they scought a second opinion from another doctor before using CAM promoted online (Broom and Tovey, 2008). However, more recent surveys have found the Internet to be increasingly important source of information on CAM, with around half of patients using alternative medicines, as well as their relatives getting health advice online (Huebner et al., 2014; Ebel et al., 2015). Thus, in Chapter 4, we focus on the kinds of individuals who are susceptible to propagating unverified information about cancer treatments that have been found to be ineffective at treating cancer.

While a large amount of work has focused on detecting and tracking fake health-related news in social media, another part of the literature focuses on correcting health misinformation (Bode and Vraga, 2015; Vraga and Bode, 2017; Bode and Vraga, 2018; Vraga and Bode, 2018) to eliminate its dangerous effects. While correcting misleading facts might not be effective if done in an algorithmic way or by peer users (Vraga and Bode, 2017; Bode and Vraga, 2018), correcting facts via trusted sources (such as the CDC) is a more effective way to reduce misconceptions (Bode and Vraga, 2015; Vraga and Bode, 2017). Further, different platforms may require different correction mechanisms (Vraga and Bode, 2018). Roozenbeek and van der Linden (2019) introduced a very novel method for correctly misconceptions, using a fake-news game where users were exposed to strategies used by fake-news sources and were trained to spot misinformation.

However, in two studies on facts about the MMR vaccine (measles, mumps, and rubella) and the seasonal flue vaccine (Nyhan et al., 2014; Nyhan and Reifler, 2015), when people were presented with pro-vaccine information from the CDC website, a stronger negative opinion about vaccines was formed (called the "backfire" effect). More work needs to be done to evaluate how effective it is to correct health misinformation in social media when the aim is to reduce incorrect beliefs.

Health Misinformation Implications in Social Media

The detection and tracking tools suggested in the previous paragraph have been shown to impact the future opinions of social media users, and potentially their subsequent behaviors. While some studies (Yom-Tov et al., 2012; Laranjo et al., 2014, 2015; Bode and Vraga, 2018) have shown that social media content has either a positive or no significant effect on people's behavior, other work has demonstrated that many communities succeed in spreading their beliefs, and that misinformation spread by social media has a larger potential to reach the public and is more popular than correct information (Oyeyemi et al., 2014b; Sharma et al., 2017).

A notable example is the presence of anti-vaccine groups posting fake claims about vaccines to influence decision making within some communities (Dunn et al., 2015; Salathé and Khandelwal, 2011). In this regard, Dunn et al. (2015) found that people who were more exposed to negative opinions about HPV (Human papilloma virus) vaccines were more likely to post negative opinions about the topic. Salathé and Khandelwal (2011) found a correlation between sentiments expressed on Twitter about a new vaccine and the CDC-estimated vaccination rates by region. More importantly, if negative vaccine sentiment communities reach new individuals, there is a high likelihood that disease rates will increase in the targeted area. Moreover, concentrated efforts promoting doubts about the medical establishment, such as the "anti-vaxxer" movement on Twitter, play into larger skepticism about government, and conspiratorial thinking, as Mitra et al. (2016) showed.

Looking at the vaccination topic in social media blogs, Tangherlini et al. (2016) designed a machine learning approach to understand the effect of story aggregation on parents in "mommy blogs". Their results showed that anti-vaccine content was highly represented in parenting social media blogs. The major arguments anti-vaccine parents used in their story aggregation involved religious beliefs and adverse reactions to vaccines (such as autism, pain, and even death). Although parents joined and left the blogs frequently, the anti-vaccine amount of content was persistent and robust regardless of these membership changes. More importantly, when new parents join the blog, they are exposed to a large content from the anti-vaccine community, which had the potential to negatively influence new parents' health decision making.

Similar to these are pro-eating disorder (ED) online communities and their efforts to support ED lifestyles, which are associated with negative health consequences. Arseniev-Koehler et al. (2016) designed a study to investigated Pro-ED profiles in Twitter and their followers. To identify the targeted profiles, authors performed content analysis of the profile content using a codebook based ED screening guidelines. The results showed that self identified pro-ED profiles mention eating disorders through tweets to their audience of followers. While the presence of ED socialization in Twitter mention social support, pro-ED content can reinforce disordered eating behaviors and their associated negative health consequences..

These findings highlight the dangers of misinformation spread on social media, and the urgency of developing research that looks at factors affecting the adoption of beliefs about public health interventions. We suggest tools (using machine learning and crowdsourcing) to detect and tract health misinformation facts and who is responsible of propagating them in Chapters 3 and 4.

Chapter 3

Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter

This chapter proposes a suite of tools for tracking health-related misinformation, and describes a case study of tracking a health crisis on Twitter. We provide a methodology for uncovering the streams of tweets spreading rumors about the 2016 Zika outbreak identified by the WHO. In particular, we track rumors outlined by the WHO (along with Snopes.com[1]) in the stream of nearly 13 million tweets. We employ both automated LDA-based topic discovery as well as a high-precision expert-led Information Retrieval approach to identify the relevant tweets in this stream. Using crowdsourcing, we distinguish between rumor and clarification tweets, which we then use to build automatic classifiers. Here, we present in-depth temporal analysis of the found rumors, their origins, and interactions with informational sources.

In this chapter, we first explain the data collection steps then we detail the process of selecting the list of rumors. Later, we describe the suggested solution about rumor tracking from query construction, crowdsourcing to the rumor classification task. Finally, we present a discussion section and ' then we conclude.

[1] http://www.snopes.com/

3.1 Research Questions

Focusing on the Zika virus and looking at Twitter data, we aim at answering the following research questions:

- R1. What are the best features to assess the credibility of a medical topic in Twitter?
- R2. What is the correlation between the clarification campaigns and the raising health-related rumors in Twitter?
- R3. Can we automatically detect tweets containing rumors about a specific health condition in Twitter?

3.2 Data Collection

The Zika virus has been known for decades; it was discovered in Uganda in the 1940s and until recently it has been unnoticed. Things changed dramatically in 2015 when this mosquito-borne disease started to spread quickly across Brazil and then most of the American continent, becoming a major global health crisis. This crisis became more dramatic as the link between the Zika infection and serious brain malformation (i.e. microcephally) started to emerge. Furthermore, fears of a global pandemic started to emerge, since Zika is spread by a mosquito from the *Aedes* family, which is present in many countries. Another source of concern were the Rio summer Olympics Games, which brought international travelers to the affected areas. As at the time there was no cure or vaccine for the Zika viral infection, communication with the public was one of the most important tools to control this outbreak. These communication efforts – dealing with the detection and prevention of Zika, and also the reduction of mosquito breeding – have been challenged by the appearance of rumors that, in the best of cases, were interfering with the public health campaigns (WHO, 2016).

In this chapter, communications about the Zika virus was collected using the Artificial Intelligence for Disaster Response (AIDR)[2] platform, which taps into Twitter Streaming Application Program Interface (API). The keywords' list contained the following (searched as quoted strings): zika, microcefalia, microcephaly, #zika, zika virus, *Aedes*, zika fever, *Spondweni* virus, *Aedes albopictus*, *maculopapular* rash. We aimed to cover both everyday wording as well as medical jargon which may be associated with the topic. Furthermore,

[2] http://aidr.qcri.org/

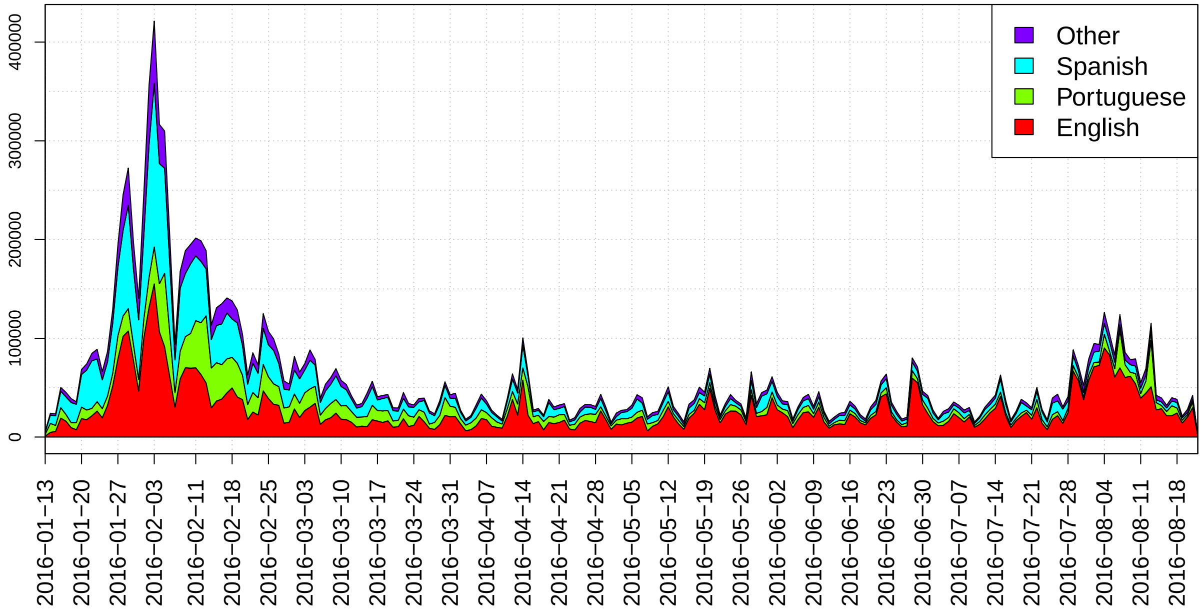


Figure 3.1: Zika-related Twitter data volume, separated by language.

"zika" word is used in English, Portuguese, and Spanish, which are the major languages of the populations affected by the virus. The resulting collection of 13,728,215 tweets spans January 13 - August 22, 2016, and includes the peak of interest in Zika (in early February) and the Olympic Games in Brazil (August 5-21). Figure 3.1 shows the volume of the data by language.

Since no language restriction was imposed during data collection (besides some bias English keywords introduced), we captured a plurality of languages, with three dominant ones which represent more than half of the dataset – English, Spanish and Portuguese (46%, 27% and 17% respectively). The language was determined using the language tag in the meta-data of the tweet provided by the Twitter API. Table 3.1 summarizes the global statistics of the language distributions. It illustrates the international nature of the Zika crisis, with each language identifying a population and its affected diasporas. In this chapter, we focus on the English data, and discuss future work involving other languages below.

To understand the geographic distribution of tweets, we use several sources to geo-locate them. We begin with the GPS attributes (*latitude*, *longitude*) of the tweet meta-data and convert them to the corresponding country name using World Borders API[3]. As 99% of

[3] http://thematicmapping.org/downloads/world_borders.php

Table 3.1: Data statistics by main language groups.

Language	**Total tweets**	**Users**	**Tweets/user**
English	6,267,173	1,318,293	4.75
Spanish	3,689,292	727,105	5.07
Portuguese	2,296,611	623,968	3.68
Other	1,475,139	593,221	2.49
Total	13,728,215	3,262,587 ⋆	4.21

⋆ or 2,546,851 unique users

tweets have missing GPS attributes, we look at location names in the *place* attribute of the tweet and convert them to exact country names. Where there is no location mentioned, we assume that the tweet location is where the user is located and obtain the corresponding location from the user's profile. It is worth mentioning that users' locations are messy, as they are written by the users themselves. Thus, we use Yahoo Placemaker API[4] to map the users' place fields to GPS locations. Finally, where no user location is mentioned, we get the tweet location by looking into other already identified location tweets tweeted by the same user (resorting to this noisy approach only if all other geo-location attempts fail). Implementing these steps, we achieve 68% coverage for English data. The top locations in decreasing number of tweets are the United States, United Kingdom, India, Canada, Nigeria, and Brazil, indicating a highly international dataset.

Implementing these steps, we achieve relatively high coverage for English, Spanish and Portuguese tweets (68%, 63% and 64% respectively). Appendix A shows the world maps of the tweet volume distribution in these three languages, normalized by the number of Internet users per country (U.S. Census Bureau, 2016).

English tweets are spread over various locations, with 26.7% of them from the United States. For Spanish, most tweets are originated from South America (Venezuela 11.2%, Mexico 6.4%, Colombia 3.3%) and Spain (2.9%). Finally, most Portuguese tweets are located in Brazil (34.75%). Note the worldwide popularity of Portuguese, suggesting the topic is relevant to the diasporas around the world.

[4] http://www.programmableweb.com/api/yahoo-placemaker

3.3 Rumor Selection

Our rumor selection process begins with a reliable list of information from trusted sources. We chose the WHO website as an authority for detecting and verifying rumors about Zika. As Zika was spreading further in the world, WHO provided a source listing major international rumors and misinformation about the virus. At the time of writing, the WHO website (World Health Organization, 2016) listed eight statements debunking ongoing rumors. Out of these, four were unsuitable, as they were not topically cohesive. For instance, one explained "Fish can help stop Zika" but did not explicitly state what the rumor which this statement would debunk would be. The process in understanding whether a topic is a rumor involved writing out an unambiguous description of the rumor, such that a tweet may be easily classified as being one. If no such description could be written, the topic was discarded. Additionally, we employed Snopes.com, which is an online authority for detecting and verifying rumors in social media, emails, and other online networks (Nourbakhsh et al., 2015), based on expert sourcing. From its Zika-related articles, we selected rumors which are not specific to the US. The final list of rumors, shown in Table 3.2, along with example tweets which propagate them, includes a total of six Zika rumor stories (four from WHO and two from Snopes).

Note that the selection of these Zika rumor topics was supervised by health experts (William Schulz, Clarissa Simas, and Per Egil Kummervold listed in the acknowledgement section) in order to ensure the coverage of the most important and influential topics related to the Zika outbreak.

3.4 Rumor Tracking

We first attempt to discover these rumors using an automated technique, such as topic discovery which has been used to identify rumors in social media (Ma et al., 2015; Wu et al., 2015). We train a Latent Dirichlet Allocation (LDA) (David et al., 2003) model on the English-language tweets, which then produces n "topics," grouping words which appear in similar contexts together in a topic. However, after a manual examination of n topics (varying from 5 to 50), we did not find any topics pertaining to the above-selected rumors. The vast majority of topics were informational, followed by spam and jokes. Examples of the LDA discovered topics are: the link between Zika virus and microcephali, media warnings about Zika traveling risks, the Olympic public threat of spreading Zika etc. Thus, we illustrate the necessity of incorporating expert knowledge in order to achieve a high-precision view of the data for our purpose.

Table 3.2: Zika rumor descriptions and example tweets. First four come from WHO and last two from Snopes.

Rumor Description	**Example tweets**
R1) Zika virus is linked to genetically modified mosquitoes	*BIOWEAPON! #Zika Virus Is Being Spread by #GMO #Mosquitoes Funded by Gates!*
R2) Zika virus symptoms are similar to seasonal flu	*The affects of Zika are same symptoms as the Common Cold. #StopSpreadingGMOMosquitos*
R3) Vaccines cause microcephaly in babies	*Government document confirms tdap vaccine causes microcephaly.. https://t.co/4ZVLbaabbG*
R4) Pyriproxyfen insecticide causes microcephaly	*"Argentine and Brazilian doctors suspect mosquito insecticide as cause of microcephaly"*
R5) Americans are immune to Zika virus	*Yup and Americans R immune to Zika, so why fund a response to it?*
R6) Coffee as mosquito-repellent to protect against Zika	*Bring on the Cuban coffee. Say Goodbye to Zika mosquitoes. Dee Lundy-Charles Fredric Sweeney Joshua Oates Laure... http://fb.me/tArL595b*

3.4.1 Query Construction

We consider the task of extracting tweets relevant to our rumors as a standard Information Retrieval task. We first index the collected tweets using Indri[5], and submit a set of handcrafted interactively designed search queries (similarly to Qazvinian et al. (2011)). Each query is a boolean string consisting of a list of keywords that best describe the rumor. These keywords are first identified then connected using the AND, OR, and NOT operators. Each keyword is then replaced with a series of possible synonyms and replacements, all connected via the OR operator. For instance, consider the rumor saying that "vaccines cause microcephaly" (R3). Transforming this story to a query language would include several common ways of referring to vaccines, as shown in Table 3.3. The queries are hand-crafted over at least 3 iterations of labeling the top 10 returned results.

Designing the queries to extract the tweets was not a trivial task. One of the challenges is that many medical term synonyms needed to be added to the query to get the highest coverage. We did not rely on automatic query expansion techniques such as Pseudo-relevance feedback as these automatic algorithms perform well in medical articles and not in informal unstructured text such as Twitter messages (Ruthven and Lalmas, 2003). Additionally, we added words that distinguish general information tweets from rumors.

[5] http://www.lemurproject.org/indri.php

Table 3.3: Rumor queries and the number of tweets retrieved.

No	Regular Expression Query	# tweets
R1	genetically | GMO	73,832
R2	(symptom & (flu | cold)) & (not(rash))	469
R3	(tdap | MMR | Measles | Mumps | Rubella) & vaccine & microcephaly) | (vaccine &(cause | link | relate) & microcephaly)	4,329
R4	(montsanto | pesticide | pyriproxyfen | insecticide) & microcephaly	10,389
R5	american & immune	351
R6	((coffee | java | jive) & (repellent | protect)) & (java & jive) & (coffee & mosquito))	202
Total	-	89,572

For example, in R2, to distinguish a rumor from a general information, we need to add (NOT rash) to the query because this is the symptom that differs between Zika symptoms and the seasonal flu ones.

The final retrieval resulted in 89,572 tweets varying greatly by rumor, with a maximum of 73,832 to 202 (Table 3.3). These tweets, however, still may contain false positive tweets that match the query but are not a rumor. For example, the following tweets are all about vaccines and microcephaly in babies (R3). The first tweet is stating that Zika vaccine causes microcephaly **(rumor)**, but the second tweet clarifies that there is no evidence suggesting Zika vaccine causes microcephaly **(clarification)**, and the third does not mention anything specific about the relationship between Zika vaccine and microcephaly **(other)**.

> **(rumor)** *Government document confirms tdap vaccine causes microcephaly.. https://t.co/4ZVLbaabbG*
> **(clarification)** *Anti-vaccination extremists falsely claim that Tdap #vaccine causes microcephaly suspected to be caused by.. https://t.co/yvfHlAFKhw*
> **(other)** *No cure, no vaccine for a virus that scientists believe to cause microcephaly! #microcephaly #ZikaVirus https://t.co/EuG9b1AJVw*

In the coming section, we explain the approach we take in order to distinguish between the three different types of information available in our dataset.

Table 3.4: Crowdflower label statistics of unique tweets in each category (propagated labels to duplicates in parentheses).

	Labeled	**Rumor**	**Clarification**	**Other**
R1	1,000 (42,432)	253 (11,773)	50 (1,912)	697 (28,747)
R2	302 (469)	217 (348)	71 (100)	14 (21)
R3	796 (4,329)	478 (2,853)	88 (846)	230 (630)
R4	1,000 (8,085)	749 (5,586)	221 (2,338)	30 (161)
R5	131 (351)	17 (22)	99 (17)	15 (312)
R6	114 (202)	72 (129)	5 (25)	37 (48)

3.4.2 Crowdsourced Annotation

To annotate the tweets as to whether they are indeed rumors, we employ the crowdsourcing platform "Crowdflower"[6]. Previous studies have shown that using crowds (anonymous workers) for health-related annotation is an effective way to label large amounts of data without employing experts (Yu et al., 2013; Zhai et al., 2013). We begin by creating a task for each topic with clear instructions on the labeling of the tweets as either supporting the rumor (by outright statement or ambiguity), debunking the rumor (by clarification), or doing neither. Also for each task we create a set of no fewer than 20 "gold standard" tweets (those with known classifications) in order to test the quality of annotations throughout the jobs. If an annotator did not pass the threshold of 70% accuracy, he/she would be banned from the task and the annotations would be discarded. Each tweet was labeled at least 3 times and a majority vote determined its classification.

The tweets were first de-duplicated by stripping tweet-specific elements such as RT (standing for "re-tweet"), special characters, and mentions, such that only one copy of each tweet was to be labeled. A maximum of 1,000 tweets were annotated per rumor. For those which had more than 1,000 unique tweets (R1 and R4), we first selected 700 most re-tweeted tweets, and sampled 300 from the rest. After the labeling of these unique ones, the label was then propagated to the duplicates within the set.

Table 3.4 shows the distribution of classes for the six rumors, with the number of tweets with propagated labels in parentheses. Although the queries were hand-crafted (manually generated by us) to capture rumors, only 51% of final tweets were rumors (an

[6] http://www.crowdflower.com/

average percentage across topics, such that no one topic dominates the statistic), and 15% clarifications, attempting to debunk these rumors. The annotator agreement (as measured in label overlap) ranged between 76% (R2) and 93% (R5) with an average of 87.7%, indicating the task differs in difficulty, but is overall clear to the annotators.

3.4.3 Temporal Tracking

Next, we examine the "paths" these rumors have taken in the story line of Zika in our dataset. Figure 3.2 illustrates the bursty nature of these rumors. The plots also show Pearson product-moment correlation r between the rumor and clarification volumes. For R4,5,6, the volume of clarification corresponds rather closely to that of the rumor with r of around 0.5. However, R1,2,3 display a mismatch between clarification attempts and the rumors. We define the "origin" tweets for rumors or clarifications as the most prominent tweets at that time for the corresponding class and we explain Figure 3.2 in detail as follows:

The case of mutant mosquitoes: the case of R1 is interesting, since it carries over a concern about the dangers of genetically modified organisms (GMO) which has been popular for several years. For instance, the spike in July was due to an article published on The Real Strategy website[7] claiming a link between "chemical exposure" and microcephaly, which gained thousands of retweets within days. However, without any interference that we detected from authoritative sources, the rumor quickly died out.

Have you got Zika? Flu and cold are very common diseases, therefore confusion between flu or cold and Zika might pose a serious problem for health authorities. This case is addressed in R2. Often the tweets appear to be jokes of users who feel flu-like symptoms such as:

> *RT @arzel: my friend had a small cold and I caught him googling "zika virus symptoms"*

Thus, although there are regular tweets on the true symptoms of Zika, there is a large proportion among these tweets that are jokes or lighthearted statements.

The killer vaccines: Similar to R1, R3's peak originated in April with an article on another advocacy website www.march-against-monsanto.com (which argues that Monsanto, an agricultural biotechnology corporation, threatens the environment and the farmers) ti-

[7] For more on this rumor see http://www.huffingtonpost.com/entry/zika-monsanto-pyriproxyfen-microcephaly_us_56c2712de4b0b40245c79f7c

tled “1991 Government Document Confirms Tdap Vaccine Causes Microcephaly”[8]. The article was readily believable to people who already view Monsanto negatively and might be spread by pharmaceutical companies to create an opportunity to sell new Zika virus vaccines as Dredze et al. (2016) suggested in his paper. The post happened after a major WHO campaign in February and March saying “No evidence that vaccines cause microcephaly”[9]. Interestingly, the April spike receded just as quickly without any clarifications from authoritative sources.

Pesticides, immunities and coffee grounds: Others, however, did have a strong interaction between the rumor and a quick reaction with clarifications. For instance, the most retweeted stories of R4 are those coming from mainstream media including CNN and WHO stating there is “No link between pesticide and microcephaly”. At the top three of R5 are stories on the “crazy and dangerous story [that] Americans are immune to Zika” and links to the debunking website Snopes. Similar to R2, in R6 is a case of hyperbole and exaggeration of a story saying mosquito larvae do not thrive in coffee-infused water, which was turned into sensationalist tweets claiming “Could Coffee Be the Answer in the Fight Against Zika Mosquitoes?”, but which still linked to the original correct information.

Thus, we show the varied nature of the rumors in the Zika stream. Those which were accompanied with mainstream coverage quickly decreased (R4-6), but even those which originated from the websites of various advocacy groups and were not met with official response were also short-lived (R1, R3). The longer-lived one is the one which concerned the daily occurrences (having a flu R2 or, possibly, coffee R6) which propagates in the Twitter lore.

3.4.4 Rumor Classification

Next, we turn to the supervised methods which have been proposed in previous work on news in social media that seek to establish the level of credibility of information automatically by observing specific features extracted from the social media. For instance, Castillo et al. (2011) and Qazvinian et al. (2011) suggested that the best features to assess the credibility of news topics are those that look into the user, message and topic features. Inspired by these works, we build a set of features in order to assess their power in automatically distinguishing rumors from non-rumors.

[8] http://www.march-against-monsanto.com/1991-government-document-confirms-tdap-vaccine-causes-microcephaly/

[9] https://twitter.com/WHO/status/708317001366806528

Gathering all the relevant tweets for the topics in Table 3.2, results in a total of 56,985 tweets. Later, we filter tweets that are exact duplicates (tweets sharing exact similar information including text, URLs, hashtags, and mentions) as the presence of the duplicates might influence the precision and recall values, resulting in a total of 26,728 tweets with human-assigned labels. We group the labels used in Table 3.2 such that we consider a rumor as the tweet that has been labeled by Crowdflower users as “rumor” (32% - 8,488 tweets) and a non-rumor as the tweet that has either been labeled as “clarification” or “other” (68% - 18,240 tweets). Note that we cannot consider “clarification” class alone, as it is vastly under-represented in our data (in part due to our focus on retrieving rumors in the previous steps).

The feature set is listed in Table 3.5 and consists of 48 features grouped into five categories. The first three categories (Twitter, sentiment, and linguistic features) have been previously implemented in news credibility detection (Gayo-Avello et al., 2013), whereas the last two (readability and medical features) are new to the work proposed in this chapter:

Twitter features As Castillo et al. (2011) use Twitter features to define credibility in news topics, we build 18 similar features including the number of retweets, number of user followers and following, the presence of hashtags and mentions, the user's number of tweets, etc.

Sentiment features We consider five measures of emotional state in our dataset: count of positive/negative words, count of positive/negative smileys, and sentiment score representing the strength of sentiment (Lowe et al., 2011).

Linguistic features We also introduce measures to characterize different linguistic styles in Twitter text (Castillo et al., 2011). We compute 17 different linguistic styles e.g.: counts of adjectives, adverbs, pronouns, sentences, upper and lower case characters.

Readability features Feng et al. (2010b) defined the readability score as a measure of how easy it is to understand a piece of text. We introduce a set of tweet text readability measures with the intuition that more readable information is more credible. We implemented the predefined readability scores by Feng et al. (2010b) (Flech, automated, Flesch_kincaid, Gunning, and SMOG scores) in addition to computing the number of complex words and average number of syllables per word. Moreover, we counted the number of words not in word2vec news vocabulary which may signal slang language (Mikolov et al., 2013).

Medical/Domain features We define specialized features in the medical domain by focusing on the medical lexicon of tweets and the reliability of sources shared using

URLs. First, we build a medical lexicon[10] which signals how many medical terms are used in the tweet. Prior studies showed that Wikipedia is a reliable knowledge base for medical data extraction tasks (Friedlin and McDonald, 2010). Additionally, as a source for lexical and contextual features, Wikipedia was used in the past to improve medical text relation extraction (Rink et al., 2011). Guided by prior work, we build a specialized lexicon by crawling a total of 113 Wikipedia pages under the category of "Infectious disease", resulting in 22,123 words representing corpus M. Then, we download the same number (22,123) of the most frequent words on all of Wikipedia, representing a general corpus W. These can then be used to compute a probability of every word in specialized corpus M as: $mp_w = count_w / \sum_w M$, as well as the probability of every word in general corpus W as $wp_w = count_w / \sum_w W$. Next, for every word in every corpus, we compute $p_w = mp_w - wp_w$. Intuitively, the differences in probabilities p_w provide the most descriptive words related to the "infectious disease" topic which are *not* as prevalent in the general Wikipedia. Ranking the terms by p_w, we only keep the top 13,300 meaningful words, as illustrated in Table 3.6 (note the topmost words are more specific, while those further down in the ranking are more general).

Additionally, Wikipedia references are considered trusted citations as Wikipedia increasingly includes references with high-impact factor medical journals such as the *New England Journal of Medicine*, *The Lancet*, the *Journal of the American Medical Association*, and the *British Medical Journal* among the 10 most frequently cited science journals in Wikipedia in 2007 (Heilman et al., 2011). As Wikipedia pages are usually among the top results of search engine queries (Laurent and Vickers, 2009; Heilman et al., 2011), we expect people to use Wikipedia pages and references as a major source of online health information. From the same Wikipedia pages used to collect the medical lexicon, we collect a total of 2,979 referenced URLs from 441 different domains,[11] including medical literature databases and news agencies. As most Twitter URLs are shortened, we expanded the URLs to detect the original domain. Finally, we manually classify tweet URL domains as *advocacy* group (advocating specific actions or policies, or claiming to be the best in providing the related information without official ties), *social_media* (YouTube, Facebook, and social media helper websites that forward and aggregate content), *news* (news sources CNN, Reuters, etc.), *informational* (reliable resources providing medical information: medical companies, government sites, Snopes, etc.) or *non-informative* (URLs having no specific domain type). Doing this, we have a total of four different domain-type features where every

[10] Available at http://bit.ly/2m56t0w

[11] Available at http://bit.ly/2m59wpm

feature is a count of the number of URLs belonging to a domain class in the tweet.

In order to pick the best features for the classification task, we employ two different automatic feature selection techniques. First, Information Gain (IG) which is a popular filtering approach that aims at removing irrelevant features after computing the gain value (amount of information a feature brings to the training set) (Cord and Cunningham, 2008). Second, we use Greedy backward elimination technique (GBE) that starts with a model having all features, and removes features one at a time until reaching a certain performance threshold (Cord and Cunningham, 2008).

Table 3.7 shows the top features each technique produced. Here, we list the top ten features by information gain value and GBE results selecting the best ten features. Based on both techniques, the most significant features correspond to the medical features (advocacy domains count, Wikipedia domains count) followed by the syntax of the tweet text (question marks, exclamation marks, etc.) and the sentiment features (sentiment score, count positive/negative words) and some Twitter features.

Note that advocacy feature domain type is the strongest feature with high IG value (Table 3.7). It is understandable that this feature would be useful, given that it requires expert annotation. Further, we find that out of the URLs cited in rumor tweets, 35.0% were from advocacy websites, 0.1% from social media, 39.1% were from news, and 25.9% were from informative domains, compared to 3.1% from advocacy and 0.6% from social media, 32.3% from news, and 64.0% from informational in non-rumors, making the presence of advocacy groups and informational sources the distinctive features, and, interestingly, not the news media. Wikipedia domains features is also among the top-selected features in both techniques and this features is automatically computed and can be used more broadly.

Finally, we train a supervised classifier to predict which tweets contain rumor and which do not. We build a classifier separately for the top 10 features of the IG and GBE techniques. We experiment with three different learning algorithms. First, Naïve-Bayes algorithm (Zhang and Su, 2004), a probabilistic based on Bayes' theorem with strong ("naïve") independence assumptions between the features. Second, Random Forest (Ho, 1995), which is a collection of classifiers where every classifier votes for one class and every instance is classified based on the majority class. Third, Random Decision Tree (Du and Zhan, 2002), a classifier that recursively builds a tree by splitting the training data based on a criterion until all partitions have the same class label. We find the best results using the Random Tree classifier using the top 10 GBE features. For training/validation process, we perform 10-fold cross validation, in which 10 experiments are performed on a different tenth of the data held out for testing, such that we take advantage of the whole dataset for both training and testing. A summary of the best classifier (Random Tree) with top 10 GBE features results are shown in Table 3.8. As it shows, the classifier achieves a precision of 0.946 with recall 0.944 which is significantly better than a random predictor. The F-value (a harmonic mean of precision and recall) is high, indicating a good balance between precision and recall values. The final row presents the average values from across both classes. Note that these results are overfitted, given the limited amount of data available, feature selection on the test set, and also that the method relies on manually labeled tweets, with the addition that the dataset is already topically specialized.

As we find having training data within the topic to be extremely helpful in building accurate classifiers, we explore a more challenging scenario wherein the classifier is trained on five topics and tested on the sixth. The results are shown in Table 3.9. Every row of the table shows which topic is excluded in training the classifier, and then is used for testing. We find the performance is not uniform, with topics 1 and 5 having the worst precision, while topics 3 and 4 having recall under 0.500. Once again, this points to the importance of expert labeled data that is topically matched to the one in question. In Chapter 4, we introduce a proposed solution to identify questionable rumor topics by detecting and tracking who might be propagating rumors in the future.

3.5 Discussion

Key Findings.

Communication on Twitter around a major public health crisis is an essential component in the public health response. In our search of rumors in the stream of Zika-related tweets, we find automatic topic discovery tools such as LDA to be too coarse-grained to tease out the rumors WHO and Snopes have cited as most concerning. Thus, we incorporate the expert knowledge to compose high-precision queries to retrieve the relevant tweets. We also show that further steps are needed; as after a closer examination, we find only roughly half of the captured tweets are actual rumors. This insight shows the perils of using keyword or hashtag-based topic definition, as is done, for example by “Truthy” (Ratkiewicz et al., 2011) where a topic is defined by a single hashtag, or even in the work of Castillo et al. (2011) who use Twitter Monitor algorithm to formulate keyword-based queries.

Further, within the small sample of topics we examined, we discovered a variety in terms of longevity. Topics relating to everyday activities, such as seasonal flu or coffee, can be a subject to hyperbole and humor which may propagate the misinformation. However, rumors originated from known advocacy websites such as http://www.march-against-monsanto.com/ may display a spike which quickly dissipates without correction. These websites adjust their stances to the new trending topics like Zika while maintaining their core message.

Interestingly, mainstream news websites were cited at roughly the same rate in rumor tweets (39.1%) as in others (32.3%), including the clarifications. This emphasizes the importance of authoritative sources outside mainstream news media in setting the record straight. Further, Towers et al. (2015) find that mainstream news media may help spread fear and misinformation, such as in the case of Ebola in 2014, "with each Ebola-related news video inspiring tens of thousands of Ebola-related tweets and Internet searches", effectively spreading unsubstantiated panic in the United States.

Public Health Relevance.

Detecting health rumors in a timely fashion can help public health officials tackle them before they spread. However, over-reacting to a rumor might in fact increase its damage by advertising the harmful misconceptions. In the case of the Ebola outbreak, some of the rumors circulated on the Internet, such as that drinking salty water was an effective protective measure, led to several deaths (Oyeyemi et al., 2014a; Jin et al., 2014). Rumors around a vaccination trial for a new Ebola vaccine sparked fears for a regular *Measles* vaccine, which was being used to tackle a *Measles* outbreak at the same time (The Vaccine Confidence Project, 2015). Public health decisions in one country can spark rumors and mistrust in another, such as when the HPV (*Human Papilloma* Virus) vaccine campaign discontinuation in Japan sparked concerns and rumors about its safety worldwide (Larson et al., 2014).

Further, the case of Zika is highly complex, as much uncertainty surrounded important information. For instance, the pathogenesis of microcephaly took months to be established. Previous works have highlighted the difficulty of early detection of rumors (i.e., the work of Zhao et al. (2015)) in public health cases – to assess the veracity of a rumor can take months of public health investigation. However, due to the unprecedented scale of the crisis, health authorities started to act before a clear link between the Zika outbreak and microcephaly was established. This was especially challenging, since it can happen that apparent rumors are in fact truth. As an example, reports on narcolepsy as a side effect of a flu vaccine in the Nordic countries were first depicted as rumors, but later, few cases were confirmed and that took years of research and still it is contested (Sturkenboom, 2015). Although correlation was found in epidemiological data, some scholars argue that an increase in awareness due the hype of the "vaccination crisis" might have caused the increase in cases. Public health authorities are continuously working in a complex crisis communication dilemma, since they have to act on some level of uncertainty. In this study, we chose the rumors which have been identified by authoritative sources as certain. However, a different approach may be called for the detection of possible health rumors, which is an exciting future research direction.

In this context, we believe more work is needed in the integration of rumor monitoring with public health officials, and especially the workflow of communication departments of public health authorities. A pipeline such as AIDR (which provided our collection), described by Imran et al. (2014) wherein volunteers provide labeled social media during a disaster to train automated methods, may also be useful for ongoing health emergencies.

Limitations.

One of the main challenges of this study is that we cannot be sure about the representativeness of the social media users compared to the general population. The demographics of social media users tend to be young and female (Duggan and Brenner, 2013), which may be important, as some have called women "gatekeepers" of their families' health (Calabretta, 2002; Warner and Procaccino, 2004). In addition, we need to consider that particular segments of the population are more at-risk (e.g. pregnant women) and it may be difficult to identify such users online (however, tracking this particular group of users would enlighten the effect Zika has on child-baring women). Further, the limited resources of this study were applied to only a handful of rumors – those especially brought up by WHO and Snopes – and a closer collaboration with health communication experts may provide further insight into the variety of misinformation both online and its interaction with mainstream media. Finally, Zika affected many countries, and our original dataset has covered several languages. The peculiarities of rumors in each language (and by proxy, in perhaps different cultures), could illuminate differences in the perception of medical information on social media.

Outside the scope of this study, it is important to consider how approaches such as the one we used could be applied to the day to day practice of health communications experts. Follow-up studies with health communications experts should address how to integrate our research into routine practice. However, we based our approach in the analysis of WHO official documentation (e.g., misinformation list).

3.6 Summary

This chapter presents a tool pipeline incorporating expert knowledge, crowdsourcing, and machine learning for health-related rumor discovery in a social media stream. Each step of the analysis was rigorously tested by manual evaluation, providing qualitative and quantitative insight into a process needed to collect data relevant to the health communication professionals.

In particular, our study shows that tracking health misinformation in social media is not trivial, and requires some expert supervision. This can then be augmented by "crowd" workers in order to provide additional annotation of the captured rumor-related tweets. We show the bursty and varied nature of the Zika rumors, some provoked by known advocacy groups, others propagated due to their propensity towards humor or light banter. We find traditional media sources not to be prominent in clarifying rumors, but instead show the importance of authoritative informational sources. We hope that the work presented in this chapter will encourage a collaboration between health professionals and data researchers in order to quickly understand and mitigate health misinformation on social media.

Continued work will address the multilingual nature of the dataset, and expand the efforts to cross-language analysis of rumors and their potential international spread. More studies on health rumors may provide richer test beds for building automatic classifiers not just for rumors, but for the detection of informational campaigns. Finally, a user-friendly interface similar to the work of Kostkova et al. (2016), which may involve expert input, like AIDR (Imran et al., 2014), would smooth the interaction between data scientists and health communication professionals.

The proposed work in this chapter suggests a tool to successfully detect health-related rumors about Zika on Twitter. However, the tool fails to identify rumor tweets when the rumor topic of interest is new as Table 3.9 shows. One potential approach to tackle this problem is to identify users responsible for spreading rumors. Detecting and tracking such users early on during health crisis would help predict potential future topics that might be a source of rumor. We establish this idea by proposing the study explained in the next Chapter (Chapter 4) where we look at identifying potential cancer rumor users on Twitter.

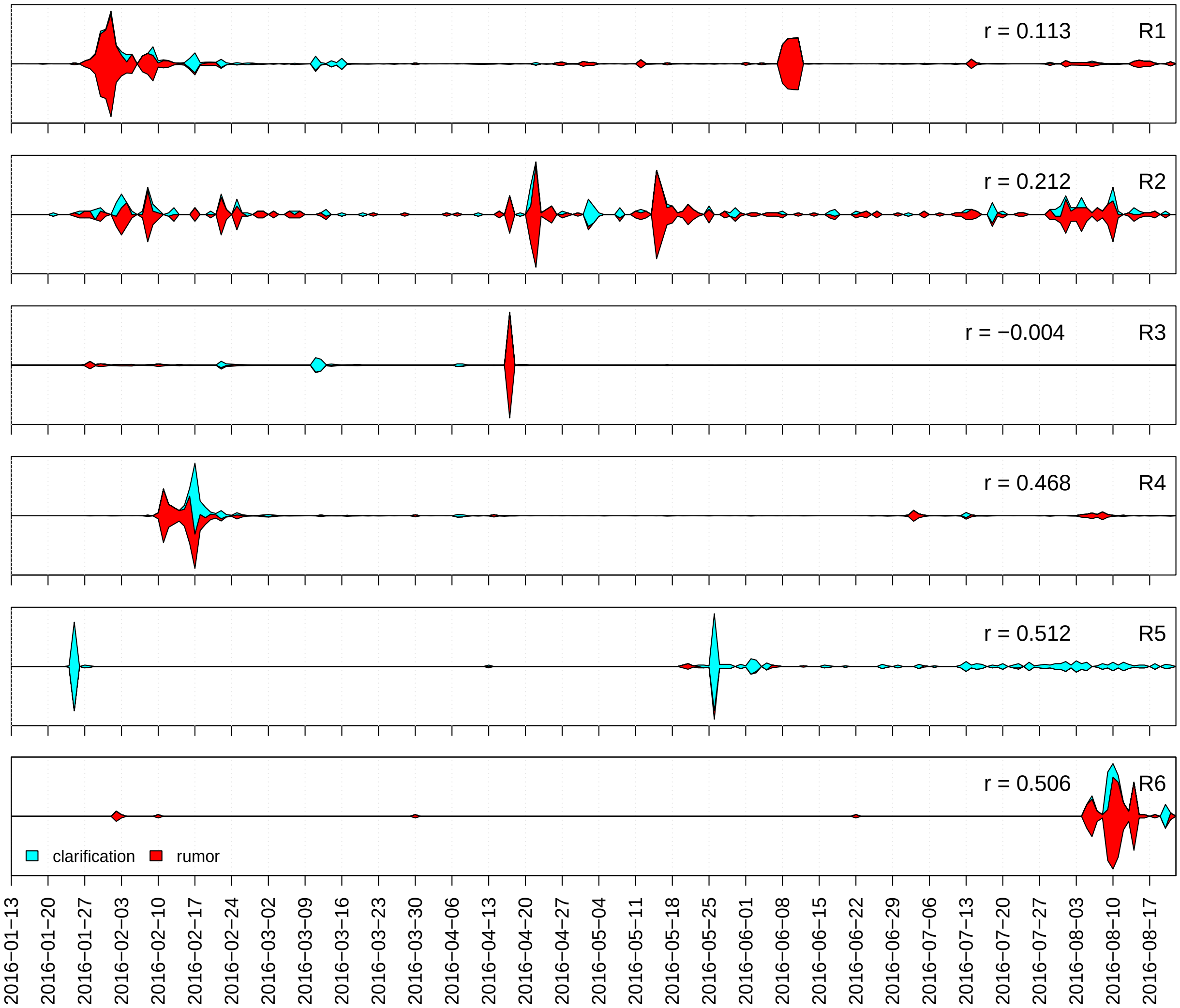


Figure 3.2: Volume of the six rumors and their clarifications, along with the Pearson product-moment correlation r between the rumor and clarification volumes.

Table 3.5: Automatically extracted features of tweets potentially belonging to a rumor.

Scope	**Feature**	**Description**
Twitter	IS RETWEET	Is a retweet; contains RT
	FOLLOWING	The number of people the user is following
	FOLLOWERS	The number of people following the user
	STATUS_COUNT	The number of tweets at posting time
	AGE	The time passed since the author registered his/her account, in days
	HAS MENTIONS	Mentions a user, eg: @CNN
	HAS HASH TAG	Contains hash_tags
	COUNT HASH TAG	Count total number of hash_tags
	DAY WEEKDAY	The day of the week in which the tweet was written
	COUNT URLS	Count total number of URLs in text
	COUNT RT	Count total number of Retweets
	COUNTRY	The country the tweet was originated from
Sentiment	SENTIMENT SCORE	sentiment score value (Lowe et al., 2011)
	POSITIVE/NEGATIVE WORDS	The number of positive/negative words in text
	EMOTICONS POS/NEG	Count total number of positive and negative emoticons in text
Linguistic	QUESTION MARK	Contains question mark ’?’
	EXCLAMATION MARK	Contains exclamation mark ’!’
	WORDS COUNT	Count total number of words in text
	COUNT SENTENCES	Count number of sentences
	CHAR COUNT	Count total number of characters in text
	UPPER COUNT	Count total number of upper case letters
	PERCENTAGE UPPER	The percentage of upper case characters
	PERCENTAGE UPPER/LOWER	The percentage of upper and lower case characters
	MULTIPLE QUES/EXCL	Contains multiple questions or exclamation marks
	COUNT NOUN	Count total number of nouns in text
	COUNT ADVERB	Count total number of adverbs in text
	COUNT ADJECTIVE	Count total number of adjectives in text
	COUNT VERB	Count total number of verbs in text
	COUNT PRONOUN	Count total number of pronouns in text
	HAS PRONOUN 1	Contains a personal pronoun in 1th person
	HAS PRONOUN 2	Contains a personal pronoun in 2nd person
	HAS PRONOUN 3	Contains a personal pronoun in 3rd person
Readability	COMPLEX WORDS	Count total number of complex words in text
	READABILITY SCORES	Flesch, Automated, Flesch_Kincaid, Gunning, and SMOG (Feng et al., 2010b)
	COUNT NOT WORD2VEC	Count total number of words not in “word2vec” Google News vocabulary
	AVG SYLLABLES	The Average number of syllables per word in text
Medical	MEDICAL_LEXICON	Count number of words in the medical lexicon
	WIKIPEDIA DOMAIN	Count number of URL domains mentioned in the wikipedia web pages
	ADVOCACY	Count number of URLs belonging to advocacy domains
	NEWS	Count number of URLs belonging to news domains
	SOCIAL	Count number of URLs belonging to social media domains
	INFORMATIVE	Count number of URLs belonging to informative/trusted domains

Table 3.6: Selected “infectious disease” Wikipedia medical lexicon words.

Word ($\boldsymbol{w}$)	$\boldsymbol{mp_w}$	$\boldsymbol{wp_w}$•	$\boldsymbol{p_w}$	**Rank**
*syphilis**	0.01	-	0.01	4
*bronchitis**	0.002	-	0.002	81
tetanus *	0.001	-	0.001	236
diarrhea	0.006	0.128	-0.121	13682
epidemiology	0.009	0.147	-0.138	15284
treatment	0.019	4.652	-4.633	33869
life	0.003	34.61	-34.608	35074

* Among the chosen top 13,300 words with highest $\boldsymbol{p_w}$

• - : is when $\boldsymbol{w}$ is not in the W corpus

Table 3.7: The features selected using information gain and greedy backward elimination.

Feature	**min, max**	μ (σ)	**IG**⋆	**GBE**•
(T) AGE	61, 281	188 (71)	9	✓
(T) HAS MENT	0, 1	0.177 (0.381)	10	✓
(T) COUNT RT	1, 2457	394 (713)	6	✓
(S) SENTIMENT	-2.2, 1.6	-0.332 (0.71)	8	✓
(S) NEG COUNT	0, 13	0.639 (0.871)	-	✓
(L) HAS QUEST	0, 1	0.193 (0.395)	4	✓
(L) HAS EXCL	0, 1	0.023 (0.161)	5	-
(L) VERB CNT	0, 38	0.673 (0.716)	-	✓
(L) ADVB CNT	0, 102	0.682(0.936)	3	-
(L) MULT. '?/!'	0, 1	0.014 (0.12)	2	✓
(M) ADVOCACY CNT	0, 2	0.045 (0.21)	1	✓
(M) WIKI CNT	0, 1	0.253 (0.435)	7	✓

⋆ Features are ranked desc according to information gain values.

• ✓: is in GBE best 10 feature subset, otherwise not.

Table 3.8: Classification performance on the rumor vs. non-rumor task, using random tree classifier with GBE features.

Class	**Precision**	**Recall**	**F-measure**
rumor	0.929	0.921	0.925
non-rumor	0.963	0.967	0.965
weighted average	0.946	0.944	0.945

Table 3.9: Classification performance of detecting rumor tweets in individual topics, using Random Tree classifier with GBE features.

Topic	**% Rumor**	**Precision**	**Recall**	**F-Measure**
R1. Zika linked to GMO	61%	0.296	0.869	0.440
R2. Flu symptoms similar to Zika	31%	0.746	0.504	0.602
R3. Vaccines cause microcephaly	71%	0.683	0.490	0.571
R4. Insecticide cause microcephaly	32%	0.594	0.432	0.500
R5. Americans are immune to Zika	32%	0.101	0.523	0.170
R6. Coffee as mosquito repellent	31%	0.688	0.688	0.688

Chapter 4

Fake Cures: User-centric Modeling of Health Misinformation in Social Media

Social media's unfettered access has made it an important venue for health discussion and a resource for patients and their loved ones. However, the quality of the information available, as well as the motivations of its posters, has been questioned.

In this study we turn to the individuals sharing questionable medical information on Twitter, in particular cancer treatments which have been medically proven to be ineffective. Having around 336 million monthly active users in the first quarter of 2018[1], Twitter is one of the largest social media websites expressly dedicated to the sharing of information, including that on cancer. Compiling hundreds of thousands of tweets on 139 queries spanning acupuncture, cinnamon, reflexology, and vitamin C, we apply strict selective criteria employing human/organization classification (McCorriston et al., 2015), name dictionaries, usage thresholds, and crowdsourced relevance refinement resulting in 4,212 users, which we then compare to those mentioning cancer in general from a previous study (Paul and Dredze, 2014). Employing previous research on rumor detection, we characterize these users in multi-faceted feature spaces, encompassing user attributes, linguistic style, sentiment, and post timing. We find users who have a more sophisticated language, who are interested in cancer, but who are not personally involved with the illness. We build a logistic regression model which, out of Twitter users mentioning cancer, is able to identify those who will eventually post a piece of misinformation with a high level of accuracy.

[1] https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Misinformation on social media is an urgent issue, and even more so in the health field. This study is one of the first to look into the characteristics of users propagating unverified "cures" of cancer on Twitter as a case study of tracking health misinformation the outside crisis communication management domain. The identification of potential sources of such misinformation would allow public health officials to monitor social media discourse, characterize the deficiencies in current communication strategy, and detect new misinformation before it causes serious harm.

In this Chapter, we first list our set of research questions, then we describe the data collection process including rumor selection, user selection, and tweets relevance refinement. Later, we present the results listing the collected users for every rumor topic as well as modeling the rumormongering behavior. Finally, we discuss the main findings and summarize the Chapter.

4.1 Research Questions

Looking at unproven cancer treatments in Twitter, we aim to answer the following research questions:

- R1. What are the best features to differentiate between rumor users and users interested in cancer in general?
- R2. Can we predict who propagates rumors looking at the users historical timeline of tweets?
- R3. Are there any linguistic differences in the historical tweets between rumor users and users interested in cancer?

4.2 Data Collection

The dataset used in this chapter consists of tweets belonging to two groups of users: (1) a "rumor" group who have posted content promoting one of 139 cancer "treatments" which have been proven ineffective as per medical expert sources, and (2) a "control" group who posted generally about cancer, but not on any of the above rumor topics. The initial data gathering, and multiple steps of user selection and relevance refinement are described below.

4.2.1 Health Rumor and Control Data Collection

As the focus of this study is the behavior of users who post on social media health content of a questionable nature, we begin by compiling a set of purported cancer "cures" which have been shown by experimentation and medical professionals to be ineffective. Four of such dubious cancer treatments come from White and Hassan (2014), where authors judged and reached a consensus about the medical treatments' efficacy by reading the corresponding Cochrane Review (Cipriani et al., 2011; Higgins et al., 2008). Next, we collect nine rumor topics from the David Colquhoun (Professor of Pharmacology at University College London) blog[2]. Professor Colquhoun's blog focuses particularly on alternative medicine such as homoeopathy, traditional Chinese medicine, and herbal medicine. Finally, we collect 126 unproven cancer treatments listed in the "List of unproven and disproven cancer treatments"[3] Wikipedia page which was refereed by Cancer Research UK[4]. The selection of these unproven cancer treatments is then supervised by a trained oncologist (Dr. Jeremie Arash Rafii Tabrizi, Professor of Genetic Medicine in Obstetrics and Gynecology, Weill Cornell Medical College - Qatar) in order to validate the ground truth of the treatments' efficacy, making sure all collected "treatments" are indeed ineffective. This process results in a total of 139 cancer treatment-related topics. Examples from the list of the collected unproven cancer treatments are: antioxidants and urine from David Colquhoun's blog, dance therapy from Cochrane Review, and acupuncture and ginger from the Wikipedia website. The full list of the collected unproven cancer treatments in this Chapter, as well as the source and keywords list, is available to the research community for developing further studies[5]. We will call these topics cancer treatment rumors, or simply rumor topics. Note that some of the above treatments may be effective in alleviating some of the symptoms of cancer, but do not actually affect the underlying progression of cancer (see Discussion for more).

Considering Twitter users posting about the above topics as the "rumor" group, we turn to existing research on health discussions for the "control" group. These would be people talking in general about cancer, such as cancer causes, prevention, symptoms, and awareness or sharing personal experiences with the medical condition. For this purpose, we use Paul and Dredze (2014)'s public health topics dataset, which consists of 144 million tweets that are related to a selection of health topics gathered during the period of 01

[2] http://www.dcscience.net/

[3] https://en.wikipedia.org/wiki/List_of_unproven_and_disproven_cancer_treatments#Ineffective_treatments

[4] http://scienceblog.cancerresearchuk.org/2014/03/24/dont-believe-the-hype-10-persistent-cancer-myths-debunked/#superfoods

[5] The topics, along with the keyword queries are available in https://tinyurl.com/y78mkg6s

August 2011 - 28 February 2013. As the focus of this study is cancer, we focus on the 676,236 users who have posted 969,259 tweets in this dataset (for a summary of our user selection process, see Figure 4.1).

Next, we turn back to the Rumor group and collect tweets on rumor topics that span the same time period as the control. For every rumor topic, we hand craft a query and expand it using general domain tools such as Google search and Google keyword planner[6], as well as medical domain tools including Mayo clinic[7], Merriam-Webster dictionary[8], and SNOMED CT BioPortal, which is a repository of biomedical ontologies (Whetzel et al., 2011). For instance, below is an expanded query for the topic *shark cartilage*, which has been shown to have no effect on the survival rate or quality of life for cancer patients (Loprinzi et al., 2005):

```
‘‘Shark cartilage’’ OR ‘‘AE-941’’ OR ‘‘Marine Collagen’’ OR ‘‘Marine Liquid Cartilage’’ OR ‘‘MSI-1256F’’
OR ‘‘Neovastat’’ OR ‘‘Sphyrna lewini’’ OR ‘‘Squalus:acanthias’’) AND cancer
```

It includes a typical way to refer to the topic, as well as more technical version of the treatment, and related products such as Neovastat, a shark cartilage extract[9]. Once again, the extended queries were verified by an oncologist for correctness and completeness. Using the Twitter Streaming Application Program Interface (API), we collected a total of 215,109 tweets about these rumor topics (see Figure 4.1), spanning 2011-2013 and 39,675 users.

4.2.2 User Selection

For both rumor and control tweets, we aimed at eliminating users that were not human such as bots, organizations, or whose tweets do not refer to the actual topics of interest (but were picked up due to faulty or ambiguous keyword matching). We perform several steps to raise the likelihood the selected users meet the above criteria.

- We apply the *Humanizr* tool (McCorriston et al., 2015) to the tweets, which was shown to have an accuracy of 94.1% at predicting whether a Twitter user is an organizational account. In this step, we remove 161 accounts from the rumor user set and 615 from the control set.

[6] https://adwords.google.com/ko/KeywordPlanner
[7] http://www.mayoclinic.org/
[8] https://www.merriam-webster.com/
[9] https://www.cancer.gov/publications/dictionaries/cancer-drug?cdrid=42021

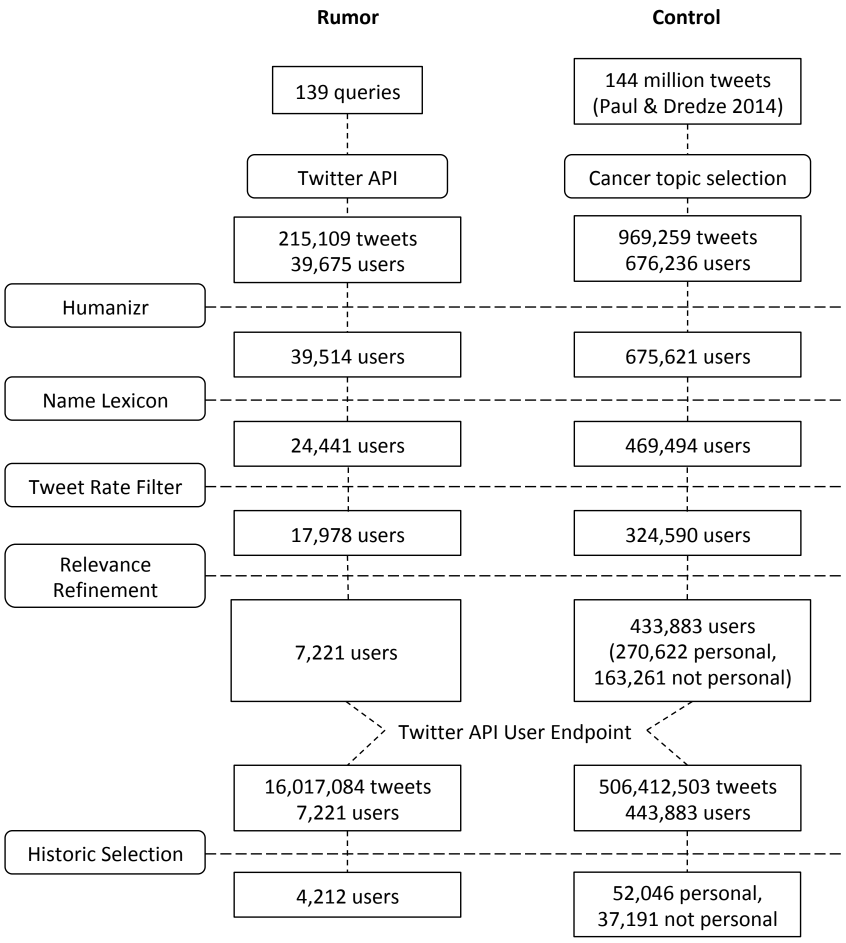

Rumor
Control
139 queries
144 million tweets (Paul & Dredze 2014)
Twitter API
Cancer topic selection
215,109 tweets 39,675 users
969,259 tweets 676,236 users
Humanizr
39,514 users
675,621 users
Name Lexicon
24,441 users
469,494 users
Tweet Rate Filter
17,978 users
324,590 users
Relevance Refinement
7,221 users
433,883 users (270,622 personal, 163,261 not personal)
Twitter API User Endpoint
16,017,084 tweets 7,221 users
506,412,503 tweets 443,883 users
Historic Selection
4,212 users
52,046 personal, 37,191 not personal

Figure 4.1: Data collection and refinement process.

Table 4.1: Average statistics of users whose names were found in name dictionary versus those not found.

Name match?	Followers	Followees	Tweets	Verified
Control				
yes	3,566	841	28,459	1.24%
no	5,594	1,011	20,012	0.88%
Rumor				
yes	5,306	1,559	25,347	1.17%
no	10,163	1,761	35,850	0.80%

- Next, we compile a (human) name dictionary with associated genders by combining names extracted from a large collection of Google+ accounts (Magno and Weber, 2014), with baby names published by National Records of Scotland[10] and United States National Security[11], resulting in a dictionary containing 106,683 names. After matching this dictionary to user names, as well as applying heuristics (such as having a "Mrs." or "Mr."), we keep only users with a matching name or identifier, excluding 15,164 (38.2%) users from the rumor and 207,394 (30.6%) from the control sets. As illustrated in Table 4.1, name matched accounts are more often verified accounts, have fewer overall tweets, followers, and following users than the non-gendered users, indicating they are less active than those not matching a name in our dictionary.

- Finally, we compute the average tweeting rate for every user as the ratio of total number of lifetime tweets over the number of days since the account was created. To exclude what are likely to be automated accounts in both the rumor and control datasets, we retain users with an average tweeting rate of less than or equal to 24 tweets per day (following posting activity thresholds such the work conducted by Olteanu et al. (2017) & Han Veiga and Eickhoff (2016)). Applying this criteria, we discard 6,463 (26%) users from the rumor and 144,904 (31%) users from the control sets.

For the remaining user accounts in both sets, we use the Twitter API user endpoint to collect the most recent 3,200 tweets, synchronizing the time spans for the two datasets to span Paul and Dredze (2014)'s timeline in 2012-2013.

[10] https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/vital-events/names/babies-first-names/full-lists-of-babies-first-names-2010-to-2014

[11] https://www.ssa.gov/oact/babynames/limits.html

4.2.3 Relevance Refinement

Human Labeling

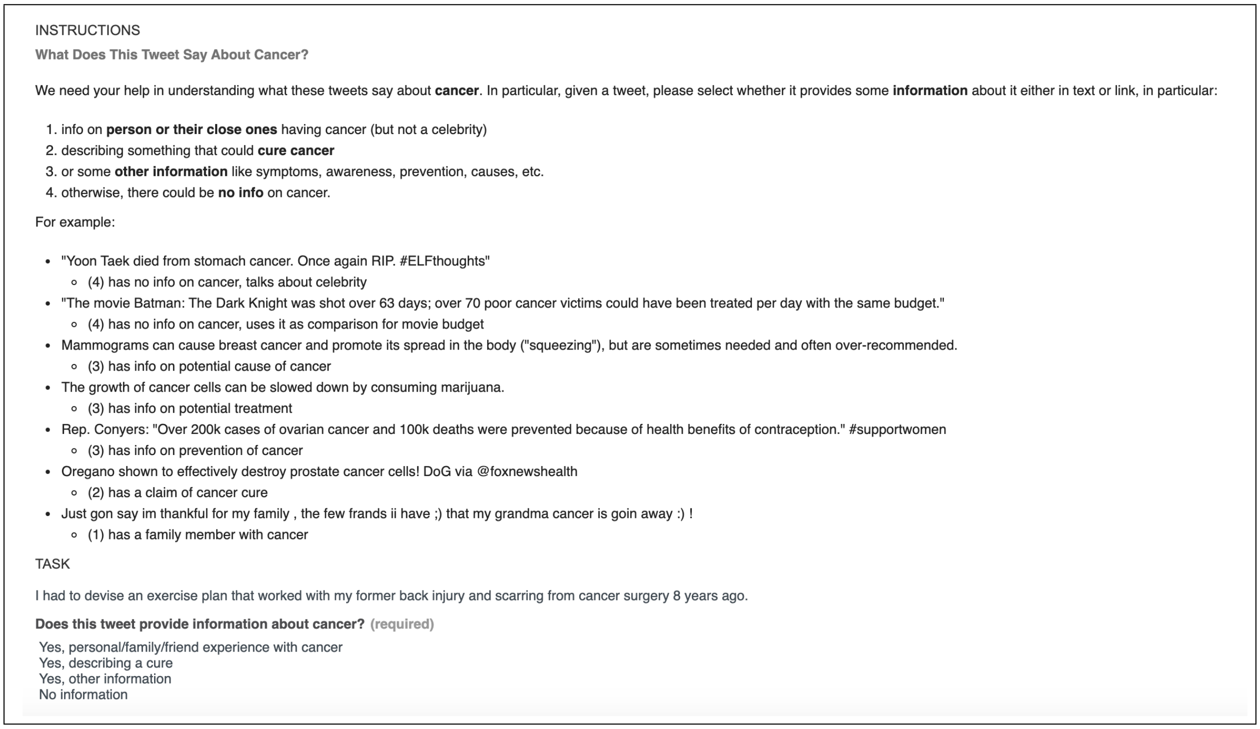
Because thus far the data has been gathered using keyword matching, we refine the tweet (and thus, the user) inclusion criteria by employing crowdsourced labeling and machine learning. In particular, we take this opportunity to make sure our data is on topic using the crowdFlower[12] crowdsourcing platform to label a subset of the data, which we then use to build topical classifiers to determine the labels for the rest. Note that instead of assessing the veracity of the claims, we are now interested in making sure the text of the tweets indeed contains the cure claims, requiring basic lay language understanding, as is reflected in the task description below.

We begin by sampling the datasets. To ensure representativeness, we stratify the sample of the rumor dataset such that at least 10 tweets from each topic are present, and the rest of the larger topics are sampled until a maximum of 100 tweets. This results in 4,152 tweets (which were de-duplicated by cleaned text). Similarly, we sample 4,000 tweets from the control set for labeling.

To ensure high quality of annotations, for each subset, 30 tweets were labeled and used as a "gold-standard". Using these tweets with known labels, the annotators are first given a quiz, and thereafter tested in each task (wherein the gold standards are hidden among other tweets). The annotator must pass the quiz and maintain at least 70% accuracy throughout the labeling process for their work to be accepted. A minimum of three independent labels were collected for each tweet to achieve a majority decision, and trial tasks of 100 tweets each were first run. A total of 184 annotators were selected by CrowdFlower, contributing a minimum of 60 annotations each. Due to this large number of participants we report the percentage of agreement instead of Fleiss' kappa.

The tasks themselves differed slightly between the data sources. For the control, we ask the workers to label each tweet on (i) whether it is about cancer, and if so, (ii) whether there is a personal (or friend/family) experience, (iii) whether there is a claim that something cures cancer, or (iv) whether some other cancer-related information is present. Figure 4.2 shows an example of a control worker labeling task. For rumor (recall these tweets also mention some treatment or remedy) we ask whether the tweet (i) is about some cancer

[12] http://crowdflower.com/



INSTRUCTIONS

What Does This Tweet Say About Cancer?

We need your help in understanding what these tweets say about **cancer**. In particular, given a tweet, please select whether it provides some **information** about it either in text or link, in particular:

1. info on **person or their close ones** having cancer (but not a celebrity)
2. describing something that could **cure cancer**
3. or some **other information** like symptoms, awareness, prevention, causes, etc.
4. otherwise, there could be **no info** on cancer.

For example:

- "Yoon Taek died from stomach cancer. Once again RIP. #ELFthoughts"
  - (4) has no info on cancer, talks about celebrity
- "The movie Batman: The Dark Knight was shot over 63 days; over 70 poor cancer victims could have been treated per day with the same budget."
  - (4) has no info on cancer, uses it as comparison for movie budget
- Mammograms can cause breast cancer and promote its spread in the body ("squeezing"), but are sometimes needed and often over-recommended.
  - (3) has info on potential cause of cancer
- The growth of cancer cells can be slowed down by consuming marijuana.
  - (3) has info on potential treatment
- Rep. Conyers: "Over 200k cases of ovarian cancer and 100k deaths were prevented because of health benefits of contraception." #supportwomen
  - (3) has info on prevention of cancer
- Oregano shown to effectively destroy prostate cancer cells! DoG via @foxnewshealth
  - (2) has a claim of cancer cure
- Just gon say im thankful for my family , the few frands ii have ;) that my grandma cancer is goin away :) !
  - (1) has a family member with cancer

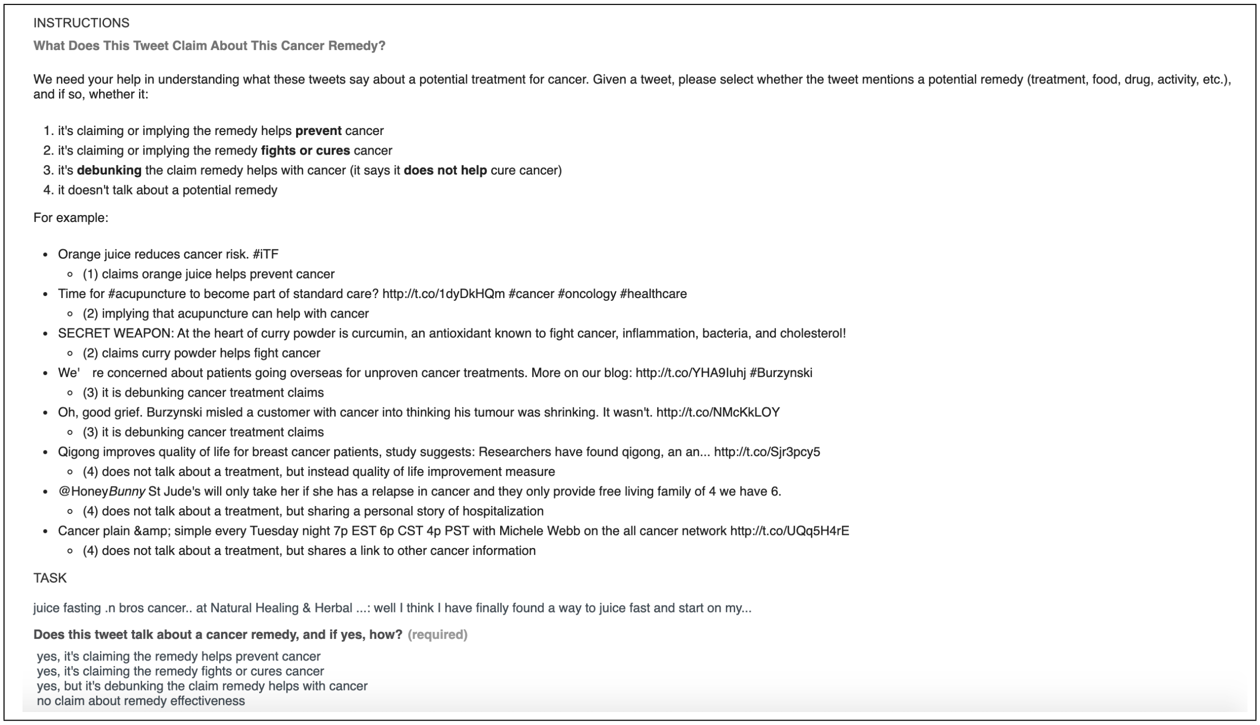
TASK

I had to devise an exercise plan that worked with my former back injury and scarring from cancer surgery 8 years ago.

Does this tweet provide information about cancer? (required)

Yes, personal/family/friend experience with cancer
Yes, describing a cure
Yes, other information
No information

Figure 4.2: The CrowdFlower control human labeling task example.



INSTRUCTIONS

What Does This Tweet Claim About This Cancer Remedy?

We need your help in understanding what these tweets say about a potential treatment for cancer. Given a tweet, please select whether the tweet mentions a potential remedy (treatment, food, drug, activity, etc.), and if so, whether it:

1. it's claiming or implying the remedy helps **prevent** cancer
2. it's claiming or implying the remedy **fights or cures** cancer
3. it's **debunking** the claim remedy helps with cancer (it says it **does not help** cure cancer)
4. it doesn't talk about a potential remedy

For example:

- Orange juice reduces cancer risk. #iTF
  - (1) claims orange juice helps prevent cancer
- Time for #acupuncture to become part of standard care? http://t.co/1dyDkHQm #cancer #oncology #healthcare
  - (2) implying that acupuncture can help with cancer
- SECRET WEAPON: At the heart of curry powder is curcumin, an antioxidant known to fight cancer, inflammation, bacteria, and cholesterol!
  - (2) claims curry powder helps fight cancer
- We'   re concerned about patients going overseas for unproven cancer treatments. More on our blog: http://t.co/YHA9Iuhj #Burzynski
  - (3) it is debunking cancer treatment claims
- Oh, good grief. Burzynski misled a customer with cancer into thinking his tumour was shrinking. It wasn't. http://t.co/NMcKkLOY
  - (3) it is debunking cancer treatment claims
- Qigong improves quality of life for breast cancer patients, study suggests: Researchers have found qigong, an an... http://t.co/Sjr3pcy5
  - (4) does not talk about a treatment, but instead quality of life improvement measure
- @Honey*Bunny* St Jude's will only take her if she has a relapse in cancer and they only provide free living family of 4 we have 6.
  - (4) does not talk about a treatment, but sharing a personal story of hospitalization
- Cancer plain &amp; simple every Tuesday night 7p EST 6p CST 4p PST with Michele Webb on the all cancer network http://t.co/UQq5H4rE
  - (4) does not talk about a treatment, but shares a link to other cancer information

TASK

juice fasting .n bros cancer.. at Natural Healing & Herbal ...: well I think I have finally found a way to juice fast and start on my...

Does this tweet talk about a cancer remedy, and if yes, how? (required)

yes, it's claiming the remedy helps prevent cancer
yes, it's claiming the remedy fights or cures cancer
yes, but it's debunking the claim remedy helps with cancer
no claim about remedy effectiveness

Figure 4.3: The CrowdFlower rumor human labeling task example.

remedy, and if so, whether there is (ii) a claim it helps with treating or curing cancer, (iii) prevents cancer, or (iv) debunks such a claim. Figure 4.3 shows an example of a rumor worker labeling task. Given the tasks had multiple selections, the agreement was relatively high at 78.7% for the control and 82.0% for the rumor. Note, as discussed earlier, in the instructions to the labelers, we emphasized looking for a claim that the remedy *treats or cures cancer*, not just a symptom, with several illustrative examples for clarity.

The results of the labeling task for the control tweets were as follows: 2,890 were labeled as having information about cancer whereas 1,110 tweets were labeled as non-related to the cancer topic. From the 2,890 cancer related tweets, 1,632 (40%) were about personal experience, 98 (2%) were about cancer cure, and 1,1160 (29%) were about other cancer-related information (symptoms, awareness, prevention, causes, etc.). The results of the labeling task for rumor tweets were as follows: 2,564 tweets were about a cancer cure and 1,587 were not about a cancer cure. From the 2,564 tweets about a remedy, 1,791 (43%) tweets claimed that the suggested treatment helped to cure cancer (claimed a rumor), 564 (13%) tweets were about prevention and 209 (5%) tweets were debunking the claim.

Classification

Next, we train several logistic regression classifiers on the labeled tweets using 1, 2, and 3-grams as features. We train the classifiers on the labeled tweets, which we then apply to the rest to characterize each user's behavior. Summaries of selection for the two datasets are below:

- Rumor: (1) is the tweet about a cancer cure? yes: 12,685, no: 7,872. Out of cancer cure tweets, (2) what kind of information does it have? claiming a cure: 9,549, prevention: 2,850, debunking claims of cure: 285. We define Rumor users as *users who claim a cure is helpful for curing or treating cancer and **not** users who talk about prevention or debunking*, resulting in 12,046 tweets for 7,221 users.

- Control: (1) is the tweet about cancer? yes: 339,047, no: 50,670. Out of cancer tweets, (2) which include a personal experience? yes: 197,608, no: 141,439. Further, (3) is the tweet is suggesting a cure? (Applying *Synthetic Minority Oversampling Technique* (Chawla et al., 2002) to balance classes) cure: 2,252, not 336,794. We define Control users as *users who post at least once about cancer, but **not** about a cancer cure*, resulting in 341,157 tweets for 270,622 users without and 199,343 tweets for 163,261 users with personal experience with cancer.

Table 4.2: Top rumor topics by number of unique users contributing tweets matching the expanded query.

#	Topic	Users	Tweets	Expanded Query
70	Juicing	6,656	13,083	juice OR juicing OR “juice diet” OR “juice plus” OR “juice +” OR “fruit vegetable juice”
11	Apitherapy	3,330	7,905	apitherapy OR honey OR pollen OR “bee bread” OR “propolis” OR “royal jelly” OR “bee venom” OR “bee sting”
52	Ginger	3,113	8,928	ginger
10	Antioxidants	2,908	5,671	antioxidant
121	Urine therapy	2,532	4,686	urine OR urinotherapy OR uropathy OR “auto-urine therapy” OR shivambu
9	Antineoplaston therapy	2,365	7,889	antineoplaston OR burzynski
81	Magnetic therapy	2,327	30,789	magnetic OR magnet OR magnets OR magnotherapy
124	Walnuts	2,013	5,474	walnut OR walnuts OR “Juglans regia” OR akhrot OR “wall nut”
4	Acupuncture	1,817	5,359	acupuncture OR accupuncture
103	Poly-MVA	1,705	7,252	“lipoic acid mineral complex” OR “poly-mva” OR “poly mva” OR “minerals vitamins and amino acids” OR vitalzym OR curcumin OR ahcc OR essiac

The overall process of user selection is summarized in Figure 4.1, with resulting 16M tweets for 7,221 users in the Rumor and 506M tweets for 443,883 users in the Control datasets. Note that we do not make the distinction in the Rumor set between personal and non-personal experiences, as in a separate crowdsourced evaluation we find only 4% to be about personal experiences.

4.3 Results

4.3.1 Rumor Topics

Table 4.2 shows the “treatments” (or “rumors”) which have the most user membership, along with the expanded queries which were used to collect the tweets. The most popular is *juicing*, followed by similar widely available remedies, *honey* and *ginger*, as well as the *antioxidant* keyword (which is often applied to a range of foods). We find a wide variety of claims surround foods and drinks. Some make bold claims outright: *“[...] University show that the soursop fruit kills cancer cells effectively, particularly prostate cancer cells, pancreas and lung”*, others speculate *“Can ginger help cure ovarian cancer? Since 2007, the University of [...] has been studying GINGER...< URL >”*, yet others invoke religious backing: *“RT @< user >: Islamic backed #cancer cure: Prophetic medicine cures woman of cancer using [...]: < URL >”*. More unusual topics include *Antineoplaston therapy* available in Dr. Burzynski clinic (for more, see Discussion), and *urine therapy.* Note

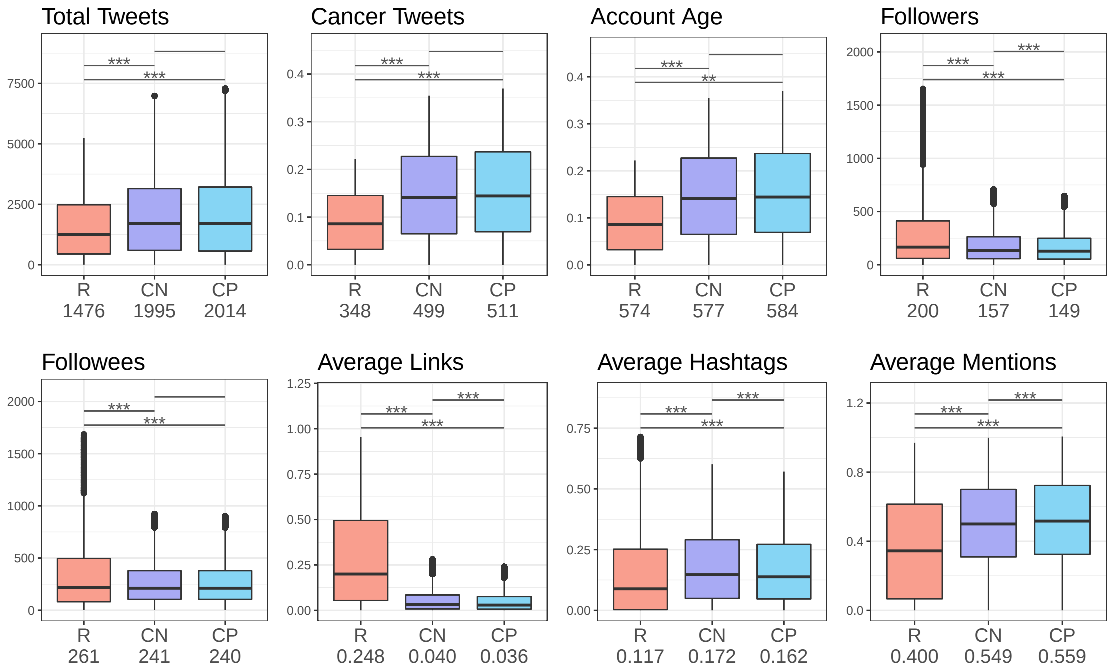

Total Tweets
Cancer Tweets
Account Age
Followers
Followees
Average Links
Average Hashtags
Average Mentions

Figure 4.4: Summary of characteristics of Rumor, Control Non-personal and Control Personal user groups. For each characteristic a box plot (excluding outliers outside 90th percentile) is shown with median values under the title. Differences in medians are tested using Mann-Whitney U test, for which p-values, Bonferroni adjusted for multiple hypothesis testing, are shown on the corresponding lines spanning the two variables being compared: $p < 0.0001$ ***, $p < 0.001$ **, $p < 0.01$ *.

that the keyword queries returned both outrageous claims as well as debunking, such as *"RT Dr. Burzynski He has the cure for cancer, the FDA want to shut him down* $< URL >$*"* and on the other side *"Burzynski Clinic libel threat to silence critics of fake cancer treatment* $< URL >$ @$< user >$*"*[13]. In the Data section, we describe how we apply supervised machine learning to sort out actual claims of purported cures (from unrelated content or rumor debunking, for instance), and use them to identify the users engaged in misinformation. Overall, of the 139 topics collected, the median number of tweets collected was 269.5, with a minimum of 11. These topics exemplify the breadth of the subjects covered in this dataset, as well as indicate the alternative cancer treatments popular on social media.

4.3.2 Modeling Rumormongering

We begin by comparing the users who have posted on these and other topics. Figures 4.4 show box plots of behavioral statistics for the three kinds of users identified above (with outliers beyond the 90th percentile excluded for clarity), such that the median is shown graphically as the bold line in each box, and also shown numerically under the label. The datasets are compared using Mann-Whitney U test – a non-parametric test that is more appropriate for highly skewed data for which normality cannot be assumed – in the bars above the plots, with p-value level indicated symbolically. We find the Rumor user set to be quite different from the other two sets of users, having fewer total account lifetime tweets (1,476 compared to around 2,000 for the Control), as well as cancer-related tweets in our dataset (more than 100 fewer), more followers and followees (some users being vastly more popular, note the long tails), and sharing more links (however fewer hashtags and mentions). Interestingly, in some behaviors there is a significant difference between personal and non-personal control tweets, with users having personal interactions with cancer having fewer followers, sharing fewer links and hashtags, but posting more mentions than non-personal control.

To examine the user behavior more deeply, we characterize the content that may be predictive of rumormongering behavior. In particular, we are interested in examining the tweets *before* a user started posting about a rumor, not necessarily the claims themselves. Thus, for the Rumor users, we select the tweets before the first rumor post, and for the Control, we sample such a date from a normal distribution having mean and variance of first rumor posts of the Rumor data. This way we aim to avoid biasing the selection to

[13] Tweets have been slightly re-phrased to preserve user's privacy.

different time periods which may trivially differentiate users. This selection allowed for the analysis of at least 100 posts for 4,212 Rumor users.

Building on our work presented in Chapter 3, and using multifaceted behavior and content features, which in the literature have been linked to credibility assessment of social media content, we generate a list of features, shown in Table 4.3, spanning user-specific statistics, as well as aggregated (via averaging) tweet-specific metrics. User features encompass proxies of popularity (number of followers and followees), as well as productivity (number of posts up to date). Tweet features can be grouped into surface and linguistic forms of the tweet and well as semantically enriched ones, including sentiment extracted from words and special characters, readability indices, and number of domains known to come from medical organization (computed in the work presented in Chapter 3). We also include a measure of entropy of the intervals between posts, which has been used to measure the predictability of retweeting patterns (Ghosh et al., 2011). Finally, we include the psycholinguistic resource LIWC[14], which has been shown to relate to user mindset (De Choudhury et al., 2013).

We then turn to examining the relationship between these variables and the tendency of the user to post about a cancer treatment rumor. To mitigate class imbalance, we undersample the Control group by randomly sampling users to achieve a one to one balance with the Rumor set. We then apply logistic regression with LASSO regularization, as the predicted class is binary and LASSO performs variable regularization and selection. However, as the data has a large number of potentially collinear features, we also use a forward feature selection method which employs Akaike Information Criterion (AIC) to select features contributing the most to the performance of the model (Venables and Ripley, 2013) (note the significant features remain largely the same, but the selection process assists in ranking most prominent ones). The resulting model is shown in Table 4.4, such that the features selected first are at the top. The McFadden R^2, the alternative to the R^2 of linear regression, is 0.925, indicating a good fit to the data. We also perform a matched experiment wherein we match Rumor to Control users on the number of followers, such that for each Rumor user the closest match in the Control is picked, resulting in McFadden R^2 of 0.906. The matched experiment shows that the classifier can be built to distinguish between rumor and control users after controlling for co-founding factors such as the number of followers. The matched experiment result suggests that when taking into account some of the behavioral peculiarities, the model fits the data less well. Examining the features, we can observe (Table 4.4):

- We find **readability** to be of importance, with the average number of syllables per

[14] https://liwc.wpengine.com/

word and SMOG readability score at the top, as well as other style-related features.

- Whether or not the account is **verified** is also important, however due to sparsity it is not statistically significant, indicating that such policing by the social media website may be of limited value.
- The top LIWC category is "ingest", one dealing with **eating and drinking**, echoing the user's interest in topics potentially related to some of the most popular remedies we found (juices, superfoods, supplements, etc.).
- These users are also more **prolific in writing about cancer**, with the number of cancer tweets being positively associated with posting a rumor (however each individual tweet counts little towards the overall probability, with coefficient at 0.001).
- They are also more likely to use **tentative language** (LIWC category 37), possibly speculating about topics other than the rumors captured in this data.
- Besides other LIWC categories pointing to speaking less positively, being male, and using more adverbs and numbers (as well as sharing more URLs), we find a weak negative relationship between using **first person pronouns** ("I","we"), indicating those engaging in posting about these rumors are not likely to be personally involved (remember also that we did not find many personal statements in the Rumor set during labeling as well).
- The positive relationship of posting **interval entropy** (Ghosh et al., 2011) means the higher inter-tweeting time entropy – and the less regular (not bot-like) is the posting behavior – the more likely the user is to post about a rumor, pointing to a largely "human" cohort.

Thus, we find (likely non-bot) users who have a more sophisticated language, who are interested in cancer, and whose language already contains speculations (besides the rumor), but who are not personally involved with the illness.

To examine the language of these groups of users in more detail, in Table 4.5 we summarize the top 20 words, with stop-words removed, in all historical tweets by control users (left), all historical tweets of rumor users (center), and only rumor tweets (right). The frequency list on the right shows some of the main trends in the tweets explicitly endorsing a "treatment". Again, we find juices and antioxidants to be popular, and prominent mentions of "help", "cure", and "treatment" (with "cure" being the more popular keyword than "treatment"). The center and the left lists show words in non-rumor tweets of the

Rumor users (center) and Control (left). Note that although both groups of users are in our dataset because at some point they have mentioned cancer, Rumor users are more focused on health, even when they are not explicitly talking about rumors, with these top 20 words containing five health-related words for Rumor users, and none for the Control. Thus, we find an encouraging sign that propensity for posting cancer treatment misinformation can be modeled and predicted automatically. Next, we discuss ramifications of this observation.

4.4 Discussion

This study expands the misinformation research prominent in Social Computing, which has been largely focused on the political domain (Shu et al., 2017; Chen et al., 2015; Gupta et al., 2013; Ma et al., 2015; Shao et al., 2016), to healthcare – where erroneous beliefs and actions may cause serious bodily harm. Complementing HCI literature on human computer use and its sociocultural implications, this work extends current work on monitoring social media during crises and pandemics (Gui et al., 2017), as well as on the tracking of specific behaviors within a community of interest (such as in the work proposed by Mejova et al. (2016), Almeida et al. (2016) and Mejova et al. (2017)). Below, we elaborate on the ecosystem of health communication and monitoring, possible application of our model, its theoretical contributions, and limitations.

Context and Case Studies. The Internet has long contributed to the ongoing "de-professionalization" of medical practice. As Michael S. Goldstein writes in *Persistence and Resurgence of Medical Pluralism*, "Health information on the Internet enhances the autonomy of those who are ill, demystifies the knowledge and practices of doctors, and increases overall skepticism about medicine" (Goldstein, 2004). In the larger history of antipathy toward professionals and their monopoly on knowledge, social media presents a new venue for patients, consumers, and concerned citizens to network and share their experiences and knowledge. It brings many of the remedies traditionally associated with home and family to a social domain. In such a networked setting, knowledge may spread and evolve. For instance, Lau et al. (2011) found that people tended to change their beliefs about a health topic when it did not concur with a majority of others. These findings emphasize the potential power of social and word-of-mouth (WOM) marketing, which could benefit both from positive messages and from controversy around the product (Kozinets et al., 2010).

Such marketing of lay health information may be especially effective in vulnerable populations, those having difficulty accessing medical care, or having poor medical literacy. For instance, a drug called "Laetrile", also known as "Amygdalin" or "Vitamin B17", is popularly promoted in India as an anti-cancer remedy (Bhatnagar et al., 2017). Despite a ban on its marketing as such by Food and Drug Administration (FDA) in the United States in 1979, it remains popular in India (Helen, 2008), and in our data we found 2,417 mentions of it in the context of cancer. Currently, Laetrile is promoted on YouTube[15] and other social media – media to which the general public has much more access than to scientific literature – allowing for an international audience to be reached. An exciting future research direction lies in enriching our dataset with geo-location in order to track the supporters of these treatments across the world.

However, an opposite reaction can also be possible. A clinic purporting to cure its patients of cancer by its founder Stanislaw R. Burzynski, MD has for several decades been a subject of scientific renunciation (Green, 1992). Recently, bloggers and activists which have been criticizing Burzynski for "disturbing business and research practices" have been attacked personally on social media, which encouraged these "skeptics" to organize and educate the public about the unproven nature of Burzynski's treatments (Blaskiewicz, 2016). In 2017, Dr. Burzynski was placed on probation for five years by the Texas Medical Board and ordered to pay a total of $60,000 in fines and restitution for not adequately informing patients about the treatments that they were receiving (Chang, 2017). We find social media and Twitter specifically to be a new battleground for the health claims of different parties, some of which may be businesses set to lose profit if their message is contested. In our dataset, we found 7,889 tweets mentioning Burzynski or "Antineoplaston therapy". Case studies of such two-sided interactions provide a window into the consequences of increased plurality in voices aiming to spread health-related information.

Applications. As beliefs are strongly linked to behavior, honing Internet-enabled communication with patients and the public at large is important in improving interventions in the health behaviors.

First, multi-faceted features proposed in this Chapter provide a foundation for examining both behavioral characteristics and interests of those prone to rumormongering. Such interest lists can be expanded beyond LIWC to specialized topical lexicons, such as those on vaccination hesitancy (Dunn et al., 2015), eating disorders (Yom-Tov et al., 2012; Ghaznavi and Taylor, 2015), antibiotics (Scanfeld et al., 2010), etc.

Second, further monitoring of suspect accounts will allow for timely identification of new potentially questionable content before it has a chance to propagate through the network, alerting public health officials of new waves of content or public interest. This content can then can be automatically pre-assessed for credibility using approaches such

[15] A YouTube search for "laetrile" on April 15, 2018 has resulted in a top video titled "Learn How Laetrile Kills Cancer Cells!".

the work proposed by Castillo et al. (2011), notifying officials if the content passes a certain threshold.

Note that automated tools will not be able to replace expert knowledge, but instead contribute to a fruitful human-expert-in-the-loop paradigm which has been proposed for research and machine learning processes (Girardi et al., 2016; Holzinger, 2016). In particular, we describe a pipeline for training the model for the tracking of discussions around "complementary and alternative medicines" for cancer, and we show that it achieves a high McFadden R^2 in fitting the data, however the pipeline can be applied to any other healthcare topic. Further, in order to remain relevant to the changing discourse, it must be periodically re-trained with fresh data in order to ameliorate "concept drift" (for which streaming solutions are also being developed such as the work proposed by Ghazikhani et al. (2014)).

Lessons learned from the empirical studies of psychological drives in rumormongering are essential in building more effective policies on communicating scientific information and managing public opinion on issues of medicine and related policies. Unlike in political domain, where bots can hijack the conversation (Shao et al., 2017; Davis et al., 2016), we find that posting interval entropy (measuring irregularity of post timings) was positively related with a rumormongering behavior, pointing to a more "human" trait. This finding emphasizes the importance of public education and communication campaigns as preventive measures targeting the public beyond social media. For instance, after a public outcry in Italy to legalize a stem cell-based treatment for neurological diseases unsupported by published evidence, researchers and public health officials called for an improvement in guidelines for its media on communicating scientific information to the public (Kamenova and Caulfield, 2015). Special care needs to be taken to promote clarifications and retractions, as it has been shown by Eysenbach and Kummervold (2005) that these are not as popular as, for example, the original incorrect news stories.

Limitations. Studying health misinformation on social media has several important limitations. Social media adoption and use differ widely between population segments: for instance, close to half (45%) of 18- to 24-year-olds in U.S. use Twitter, compared to 24% of all adults, as reported by Pew Research Center (2018). Detecting legitimate personal accounts (as opposed to bots or organizational accounts) remains a challenge, which we attempted to address using existing tools like Humanizr and baby name dictionaries, which undoubtedly introduce their own biases potentially excluding certain minorities. In particular, the Social Security name database includes all names registered at least 5 times in a year, dating back to 1880, capturing a large majority of names used. However, more resources could have been used to include the names of minorities, such as the Register of

Liberated Africans[16]. The results of this study should be taken in the light of this limitation, as we may have failed to detect misinformation in some communities. Improvement in the detection of real humans (versus bot or organizational accounts) will allow for a more accurate account selection for studying individuals.

Accessibility issues also bias the view of the populations having, for instance, visual impairment (Wu and Adamic, 2014) or other constraints to using the medium. Further, some health conditions and personal topics are associated with a social stigma which limit their discussion on social forums. For example, De Choudhury et al. (2014) found some illnesses to be searched more often than discussed in social media – a bias which may affect our selection of certain cancers. Finally, attitudes toward self-expression and trust in publicly available information may differ wildly between cultural subgroups, such as in case of Hong Kong youths, who were found to be significantly more likely to disclose personal health issues with peers online compared to their U.S. counterparts (Lin et al., 2016). Hong Kong youths also held the highest level of trust towards health-related information on social media, again pointing to the need for a personalized approach to health communications sensitive to the culture of the participants. Finally, observations in this study concern exclusively treatment claims of cancer, and may not generalize to other illnesses, especially if they have different societal stigmas. The model proposed in this chapter inherits the above limitations, thus any integration of such automated tracking must be closely monitored for bias and topic drift, and regularly updated in order to capture the latest developments in social media norms.

Finally, as any technology, the proposed analytical pipeline may be misused when applied within faulty policies. The expert definition of misinformation must be subjected to ethical constraints of medical and public health standards. How the information that is flagged by the system is handled must also avoid discouraging public discourse and information seeking.

Privacy. This study used Twitter posts which were publicly available at the time of data collection, with no private messages or messages deleted by the time of the collection included. Also, accounts which have been deleted since Paul & Dredze's collection have not been included in the study. Furthermore, the sharing of this dataset will be done according to Twitter's Terms of Use.

[16] http://liberatedafricans.org/

4.5 Summary

In this chapter, we present a case study of health misinformation on social media by examining Twitter users involved in propagating alternative treatments claiming to treat or cure cancer. The dataset collected for this study presents a highly-curated resource for the research community's future studies on the topic of health misinformation. Further, through a multi-stage process including machine learning, crowdsourcing, and heuristics, we select users who are likely to be real people, and who post on one of 139 such topics. We find that these users are likely to use more sophisticated language, and circulate in health domains prior to posting a rumor, but are not likely to be personally involved in the illness. Our findings suggest that cancer treatment misinformation may be spread not by patients, but by other actors. More research needs to be done to ascertain motivations, tactics, and the impact of such accounts. Understanding the impact of accounts spreading cancer misinformation is important as people can potentially make critical health-related decisions based what they see in their social media feeds.

Other than relying on social media, another very frequent method of collecting cancer treatment information is using online search. In the next Chapter, we move our focus from social media data to online search and broaden our focus from cancer treatments to different (less critical) medical conditions. We look into the influence of search results' bias on peoples' ability to correctly answer questions related to the effectiveness of different health treatments.

Table 4.3: User and aggregated tweet features.

Scope	**Feature**	**Description**
User	FOLLOWING	The number of people the user is following
	FOLLOWERS	The number of people following the user
	STATUS_COUNT	The number of tweets at posting time
	ACCOUNT AGE	The time passed since the author registered his/her account, in days
	VERIFIED	Whether account has been verified by Twitter
Sentiment	SENTIMENT SCORE	Sentiment score value (Lowe et al., 2011)
	POSITIVE/NEGATIVE WORDS	The number of positive/negative words in text
	EMOTICONS POS/NEG	Count total number of positive and negative emoticons in text
Linguistic	IS RETWEET	Is a retweet; contains RT
	HAS MENTIONS	Mentions a user, eg: @CNN
	HAS HASHTAG	Contains hash_tags
	URLS COUNT	Count total number of URLs in text
	HASHTAG COUNT	Count total number of hashtags
	MENTION COUNT	Count total number of mentions
	WORD COUNT	Count total number of words in text
	CHAR COUNT	Count total number of characters in text
	UPPER COUNT	Count total number of upper case letters
	COUNT SENTENCES	Count number of sentences
	QUESTION MARK	Contains question mark '?'
	EXCLAMATION MARK	Contains exclamation mark '!'
	PERCENTAGE UPPER/LOWER	The percentage of upper and lower case characters
	MULTIPLE QUES/EXCL	Contains multiple questions or exclamation marks
	COUNT NOUN	Count total number of nouns in text
	COUNT ADVERB	Count total number of adverbs in text
	COUNT ADJECTIVE	Count total number of adjectives in text
	COUNT VERB	Count total number of verbs in text
	COUNT PRONOUN	Count total number of pronouns in text
	HAS PRONOUN 1	Contains a personal pronoun in 1th person
	HAS PRONOUN 2	Contains a personal pronoun in 2nd person
	HAS PRONOUN 3	Contains a personal pronoun in 3rd person
	LIWC	73 categories from psycholinguistic resource LIWC
Readability	COMPLEX WORDS	Count total number of complex words in text
	READABILITY SCORES	Automated, Flesch_Kincaid, Gunning, and SMOG (Feng et al., 2010b)
	COUNT NOT WORD2VEC	Count total number of words not in "word2vec" Google News vocabulary
	AVG SYLLABLES	The Average number of syllables per word in text
Medical	MEDICAL_DOMAINS	Refers to URL from a known medical organization (Ghenai and Mejova, 2017)
Timing	INTERVAL ENTROPY	Entropy of hour intervals between tweets (Ghosh et al., 2011)

Table 4.4: Logistic regression with LASSO regularization model, predicting whether a user posts about a rumor, with forward feature selection. For each feature, coefficient (unstandardized), standard error, and accompanying p-value are shown. Significance levels: $p < 0.0001$ ***, $p < 0.001$ **, $p < 0.01$ *, $p < 0.05$.

variable	**coefficient**	**std. error**	**p-value**
(Intercept)	-6.160	1.405	***
Avg syllables per word	17.120	0.660	***
Is verified	-40.310	42310	
Percentage uppercase / lowercase	-0.201	0.018	***
Word count	1.491	0.131	***
SMOG readability score	-0.753	0.123	***
Percentage uppercase	0.191	0.019	***
Character count	-0.163	0.024	***
Number of cancer tweets	0.001	1.9E-04	***
LIWC48: ingest	1.839	0.722	*
Negative word count	-1.460	0.262	***
URL count	3.364	0.505	***
Is retweet	4.947	0.790	***
word2vec count	-0.634	0.165	***
LIWC55: focuspast	-1.636	0.567	**
LIWC37: tentat	2.531	0.859	**
Number of sentences	-0.610	0.205	**
LIWC32: male	-1.820	1.000	
Interval entropy	0.508	0.105	***
Account age	-0.001	2.7E-04	***
LIWC23: posemo	-0.490	0.384	
LIWC61: time	-1.431	0.378	***
LIWC13: adverb	1.758	0.536	**
LIWC20: number	2.936	1.317	*
Statuses count	7.1E-05	2.6E-05	**
LIWC42: hear	-4.742	1.799	**
Has 1st person pronoun	-1.504	0.662	*
LIWC62: work	1.591	0.665	*
LIWC40: percept	1.217	0.754	

Table 4.5: Word frequency tables summarizing the top 20 most popular terms, excluding stop-words, in all historical tweets by control users (left), all historical tweets of rumor users (center), and only rumor tweets (right).

Control History				Rumor History				Rumor Misinformation			
love	1.95%	night	0.66%	good	1.01%	video	0.54%	cancer	1.43%	cells	0.50%
good	1.55%	life	0.63%	health	1.00%	food	0.54%	juice	0.81%	out	0.48%
day	1.34%	happy	0.60%	day	0.96%	back	0.50%	RT	0.77%	healthy	0.45%
time	1.22%	ill	0.59%	love	0.85%	free	0.46%	breast	0.73%	diabetes	0.44%
people	1.00%	hope	0.58%	time	0.78%	work	0.45%	risk	0.61%	prostate	0.44%
lol	0.99%	feel	0.55%	great	0.73%	diet	0.44%	help	0.58%	antioxidant	0.42%
today	0.96%	haha	0.51%	people	0.71%	healthy	0.40%	health	0.55%	pain	0.40%
back	0.94%	follow	0.51%	today	0.68%	post	0.38%	helps	0.54%	chronic	0.37%
great	0.73%	home	0.49%	news	0.62%	weight	0.38%	cure	0.54%	patients	0.37%
work	0.70%	man	0.47%	life	0.57%	blog	0.36%	treatment	0.53%	study	0.36%

Chapter 5

The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments

This chapter aims at studying the extent to which search results influence people's health-related decision making. As search engines, nowadays, provide a mixture of correct and incorrect information, we aim at measuring the influence of incorrect information present in search results on people's decisions about the effectiveness of a health treatment. We conduct a controlled laboratory study with 60 participants where we bias the search results towards being correct or towards being incorrect for 10 different medical treatments. We also control the topmost rank of a correct result to investigate the effect of rank. During the study, we asked participants to pretend that they had a question about the effectiveness of a medical treatment and that they had decided to use a search engine to help them answer the question. For each of the ten treatments, we either present the participants with a search results page or a control condition where they had to directly answer the question without any search results at all.

We next list our research questions, then cover the details of the study and present the study results. Following the results, we discuss implications and conclude the chapter.

5.1 Research Questions

Designing the user study, we aim at answering a set of research questions. Here we list them as follows:

- R1. How does the search results bias influence participants' decision about the efficacy of medical treatments?
- R2. How does the rank of the search results influence participants' decision about the efficacy of medical treatments?
- R3. Does prior knowledge of the medical treatment and health issue effect participants' decision about the efficacy of medical treatments?
- R4. Does the participants' confidence effect their decision about the efficacy of medical treatments?
- R5. What potential user behaviors are correlated with participants' accuracy when deciding about the efficacy of medical treatments?

5.2 Study Methods and Materials

To measure the effect of search results on people's ability to evaluate the effectiveness of a medical treatment, we design a controlled within-subject laboratory study. We have two independent variables: the correctness of search results as well as the rank of the correct document being at either rank 1 or 3. Further, we either provide participants with search results or we don't show any results (control condition). We ended up with the total of five different experimental conditions. The medical treatment might either be *helpful* or *unhelpful*. The participants had to experiment the study conditions with five *helpful* treatments and five *unhelpful* treatments. In total, the participant evaluated the efficacy of ten medical treatments. In order to measure the performance of participants, we measure two dependent variables: first, the fraction of correct answers, and second the fraction of harmful answers.

We explain the details of the study including the medical treatments, the search results, the experimental conditions and the performance measures in the following section.

5.2.1 Medical Treatments

The medical treatments used for our study are received from White and Hassan (2014)'s work. White and Hassan evaluated the effectiveness of the medical treatments from the Cochrane Review which is an internationally recognized evidence- based health care resources. Each medical treatment can either be *helpful* (has a direct positive effect on people's health), *unhelpful* (has no effect or has a direct negative effect on people's health) or *inconclusive* (the effectiveness of the medical treatment is not defined yet). The definitions of these categories are provided to participants throughout the whole study.

The medical treatments and associated medical conditions are all formulated as "Does X help Y?" where X is the medical treatment and Y is the medical condition. An example of an *unhelpful* treatment would be: *Do insoles help back pain?*

Among the 249 medical treatments provided by White and Hassan (2014), we select five *helpful* treatments and five *unhelpful* treatments. Table 5.1 lists the ten medical treatments with their corresponding Cochrane ID. Further, table 5.1 lists the population knowledge about the ten medical treatments without search results' aid. It is interesting to note that participants have different knowledge about the medical treatments. While almost all participants did not believe caffeine is helpful for asthma, Most of them believe that traction is helpful for low back pain. Further, the majority of participants believe that surgery is helpful for obesity.

5.2.2 Experimental Conditions

In order to measure the ability of participants to answer the *helpful* and *unhelpful* medical treatments efficacy, we either provide them we search results or we ask them to evaluate the treatment efficacy without search results. The details of these two different experimental conditions are explained as follows:

Control Condition

During this experimental condition, no search results are provided to participants during the search task. This condition allows us to measure the population knowledge about the medical treatments without interacting with search results. Participants had to answer two medical treatments with the control condition.

			Fraction of Decisions Correct		
			Control (no search results)	Search Results Bias	
T	Medical Treatment (Cochrane ID Suffix)	Efficacy		Incorrect	Correct
T1	Do antioxidants help female subfertility? (7807.pub2)	Unhelpful	0.58 ± 0.15	0.08 ± 0.06	0.71 ± 0.09
T2	Do benzodiazepines help alcohol withdrawal? (5063.pub3)	Helpful	0.33 ± 0.14	0.29 ± 0.09	0.63 ± 0.10
T3	Do insoles help back pain? (5275.pub2)	Unhelpful	0.33 ± 0.14	0.17 ± 0.08	0.50 ± 0.10
T4	Do probiotics help treat eczema? (6135.pub2)	Unhelpful	0.33 ± 0.14	0.17 ± 0.08	0.75 ± 0.09
T5	Do sealants prevent dental decay in the permanent teeth? (1830.pub4)	Helpful	0.67 ± 0.14	0.46 ± 0.10	0.83 ± 0.08
T6	Does caffeine help asthma? (1112.pub2)	Helpful	0.08 ± 0.08	0.25 ± 0.09	0.79 ± 0.08
T7	Does cinnamon help diabetes? (7170.pub2)	Unhelpful	0.50 ± 0.15	0.00 ± 0.00	0.38 ± 0.10
T8	Does melatonin help treat and prevent jet lag? (1520)	Helpful	0.67 ± 0.14	0.38 ± 0.10	0.79 ± 0.08
T9	Does surgery help obesity? (3641.pub3)	Helpful	0.67 ± 0.14	0.46 ± 0.10	0.63 ± 0.10
T10	Does traction help low back pain? (3010.pub5)	Unhelpful	0.17 ± 0.11	0.08 ± 0.06	0.46 ± 0.10
-	Overall	-	0.43 ± 0.05	0.23 ± 0.03	0.65 ± 0.03

Table 5.1: The medical treatments (T1 - T10) with their corresponding efficacy and suffix to their Cochrane (Higgins et al., 2008) source ID.

Search Engine Result Page (SERP) Condition

We had a total of ten medical treatments to be evaluated by participants. Two among them were control conditions and the remaining eight medical treatments needed to be evaluated with the help of search engine result pages (SERP). In order to achieve that, we ask participants to pretend that they had a questions about the efficacy of the medical treatment and that they decided to use search engine to answer that question. When participants were faced with SERP condition, ten search results about the efficacy of the medical treatment were displayed. Every search result is either *correct* (agrees with Cochrane review information) or *incorrect* (contradicts with Cochrane review information).

In order to bias the search results towards correct information, we show eight correct search results and two incorrect ones. On the other hand, to bias the search results towards incorrect information, we show eight incorrect search results and two correct ones. The eight to two ratio reflects the actual search engine ratio. White and Hassan (2014) found that, out of the top ten ranked results, 80.69% were reporting that treatments are *helpful*, 12.29% were *inconclusive* and 7.01% were reporting that the treatment was *unhelpful*.

In addition to controlling the correctness, we control the rank of topmost correct search result. Specifically, we place a correct document at either rank 1 or at rank 3. We consider only the top three ranks because eye-tracking studies shows that not only the first and second results are more viewed, but also the attention from rank 1 to 3 drops by 50% (Pan et al., 2007).

We collected a pool of 8 to 10 *correct* documents and 8 to 10 *incorrect* documents for every medical treatment. To generate a SERP biased towards *correct* condition, we randomly select eight correct documents and two *incorrect* ones from the corresponding pools. In order to generate a SERP condition based towards *incorrect* condition, we randomly select eight *incorrect* documents and two *correct* documents from their corresponding pools. Further, the topmost correct document was randomly selected from the pool and assigned to rank 1 or 3. The remaining correct and incorrect documents were selected randomly from the pool and placed in the lower remaining ranks.

5.2.3 Documents and Snippets

To build the search result pages, we manually collected 158 documents about the efficacy of the medical treatments from Bing, Yahoo and Google search engines. We label the pages as either being *correct* or *incorrect* based on the truth from the Cochrane reviews. The collection and labeling process was performed by Amira Ghenai as well as Frances

A. Pogacar. For every document, we manually constructed a snippet which was displayed during the study with a title as well as a URL for the document. While Frances A. Pogacar selected the first two sentences as the document snippet for topics T1-T8 (Table 5.1), Amira Ghenai considered the snippet to be the most important descriptive sentence for topics T9 and T10. We did not realize that different techniques were employed until after the experiment was concluded. Given that we did not see significantly different click behavior across the different medical treatments, we do not believe that the different selection of snippets affects the results.

Sometimes, it was difficult to collect documents stating that the medical treatment was not helpful. In these cases, we pick documents that describe negative side effects or possible harm that the treatment might cause. We publicly share the generated documents and the snippets[1] in order for the research community to be able to replicate the experiments or do further research.

5.2.4 Performance Measures

To measure the performance of participants during the study, we define two dependent variables. The first measure is the fraction of *correct* decisions where a *correct* decision is the answer that matches the truth. In this case, if the participant answers *inconclusive*, this is considered an incorrect decision because all the treatments are either *helpful* or *unhelpful*.

The second measure is the fraction of *harmful* decisions where a *harmful* decision is the answer that is opposite to the truth. Note that if the participant answers *inconclusive*, this is not considered a *harmful* decision because this means that the participant needs more time to evaluate the effectiveness of the treatment.

To determine the statistical significance of the study results, we use a generalized linear (logistic) mixed effects model implemented in R (R Core Team, 2014) and in lme4 (Bates et al., 2015) package. Because our dependent variables (correct and harmful decisions) are binary, we use logistic regression. We model the medical treatments and the participants as random effects and the independent variables and explanatory variables as fixed effects. In order to test the significance of an independent or explanatory variable on a dependent variable, we build two models. Specifically, the first model is the complete model that includes the dependent variable, the applicable independent variables, and the random effects. The second model (null model) includes everything in the first model except the

[1] Available here: https://cs.uwaterloo.ca/~aghenai/user_study_pages.html

variable of interest. Then, using the complete and the null model, we compute a likelihood ratio test that reports a Chi-Square test statistic and p-value.

Note that the topmost correct rank is not included as a fixed effect in the model because the control condition has no search results i.e the rank is not applicable. The majority of the analysis is done using the four experimental conditions excluding the control conditions. For these analyses, we include both independent variables of Search Results Bias and Topmost Correct Rank in our models.

5.3 Study Design

The study starts by asking participants to sign a consent form. Next, they are required to fill a questionnaire about demographic information as well as information about their usage of search engines for health-related purposes. After that, we show a set of instructions followed by a quiz in order to make sure participants read carefully the study instructions. Further, the participants have the chance to go through the practice task where they have to determine the efficacy of two medical treatments. One of the medical treatments was displayed with no search results and the other was with search results. Then, participants start the study where they are asked to evaluate the effectiveness of ten medical treatments. For each search task, there is a pre-task and post-task question. Before the task, we asked participants about their knowledge of the health issue and treatment. After the task we asked the participants about their confidence in their answer. Different from White's work (White, 2014; White and Horvitz, 2015), we don't ask participants about their prior belief before the search task because we believe that participants' behavior will be biased i.e participants would resist changing the declared belief as they don't want to admit that they were wrong. After the search task, we ask participants about their confidence in the answer. At the end of the study, participants were debriefed and provided with the truth about each of the medical treatments.

Participants We obtained ethics approval from our university and then recruited participants via posters and email announcements to different graduate student email lists at the university. All participants gave their informed consent. Following their participation, we debriefed all participants and provided them with the correct answers regarding the efficacy of the medical treatments. We paid participants $15. Participants were 60 students (27 male, 33 female) from different majors (36 from engineering and mathematics, 20 from arts and sciences and 4 from other majors) with an age between 18 and 36 years old (22% less than 20, 50% between 20 and 25 and 28% greater than 25, with an average age of 23).

During the course of the study, four participants had to be replaced because of failure to successfully complete the study due to technical or other issues. After a careful examination of the study data from the 60 participants, we did not find any irregularities and thus did not clean or modify the data before analysis.

User Interface We build the study as a web application. For 8 of the 10 medical treatments, participants interact with a search engine results page. For the other two medical treatments, the participants receive the control condition, with no search results. We model the search results page after the traditional style of web search engines. At the top of the page, we display the medical treatment question that the user is asked to answer followed by a short boxed paragraph showing definitions of the health issue and treatment. We obtained the definitions from either Merriam-Websters[2] or the Mayo Clinics[3] medical dictionaries. We show the definitions to avoid confusion and to make sure participants had a basic understanding of what was meant by the health issue and medical treatment. The medical treatment question and definitions remain visible throughout the entire task.

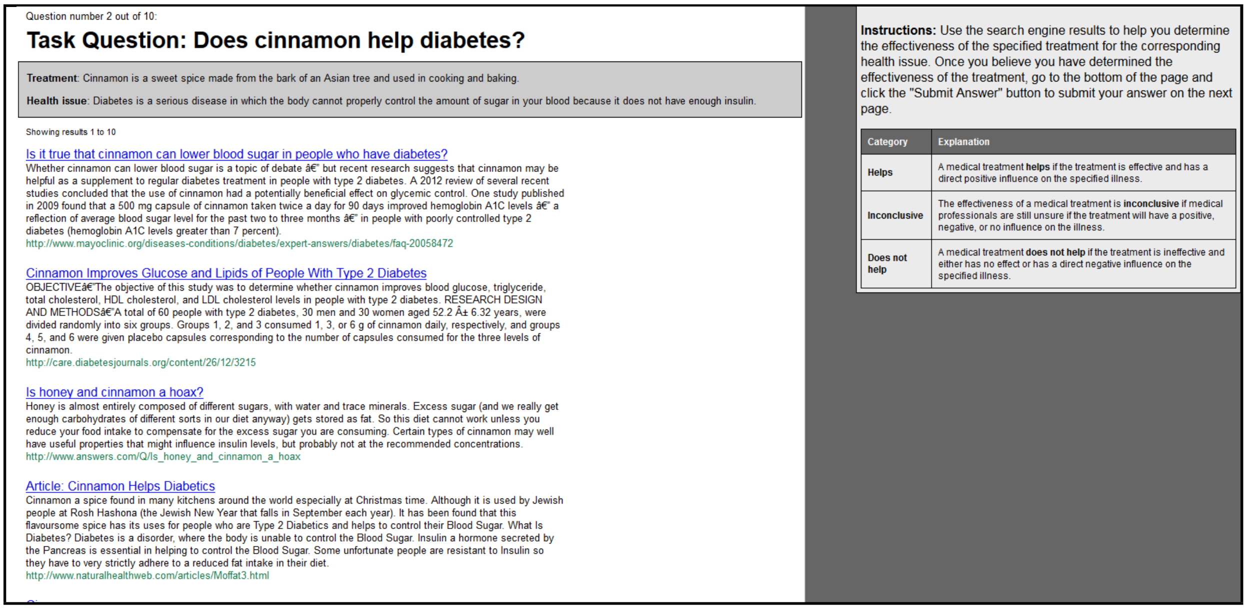
The search results page allows participants to click on the search results, but they can not issue additional queries or obtain additional results. On the right side of the search results page, we display a reminder of the definitions of the different categories of medical treatment efficacies: helps, does not help and inconclusive. Figure 5.1 shows the design of the SERP page during the search task.

For every document summary, we first show the document title followed by a snippet and a link to the actual page. When a participant clicks on a search result, we take them to a screen-shot of the web page rather than to the actual web page. We do this because we want to make sure that the participant is not able to click on any links and view any pages outside the scope of the study. In addition, this approach allows us to be certain that each participant is exposed to the same version of the web page, and that we do not have to fear the loss of pages during the study. We place a button at the bottom of the search results page that, when pressed, takes the user to a page to submit the decision regarding the efficacy of the medical treatment.

Balanced Design We use a 10 × 10 Graeco-Latin square to fully balance and randomize the medical treatments and the experimental conditions. First we generate a Latin square for the five experimental conditions, a Latin square for the five *helpful* medical treatments and a Latin square for the five *unhelpful* medical treatments. Then, we overlay the experimental conditions Latin over the *helpful* treatments Latin and over the *unhelpful* treatments Latin creating two new Graeco-Latin squares. The two new Graeco-Latin

[2] https://www.merriam-webster.com

[3] http://www.mayoclinic.org



Question number 2 out of 10:

Task Question: Does cinnamon help diabetes?

Treatment: Cinnamon is a sweet spice made from the bark of an Asian tree and used in cooking and baking.

Health issue: Diabetes is a serious disease in which the body cannot properly control the amount of sugar in your blood because it does not have enough insulin.

Showing results 1 to 10

Is it true that cinnamon can lower blood sugar in people who have diabetes?
Whether cinnamon can lower blood sugar is a topic of debate â€" but recent research suggests that cinnamon may be helpful as a supplement to regular diabetes treatment in people with type 2 diabetes. A 2012 review of several recent studies concluded that the use of cinnamon had a potentially beneficial effect on glycemic control. One study published in 2009 found that a 500 mg capsule of cinnamon taken twice a day for 90 days improved hemoglobin A1C levels â€" a reflection of average blood sugar level for the past two to three months â€" in people with poorly controlled type 2 diabetes (hemoglobin A1C levels greater than 7 percent).
http://www.mayoclinic.org/diseases-conditions/diabetes/expert-answers/diabetes/faq-20058472

Cinnamon Improves Glucose and Lipids of People With Type 2 Diabetes
OBJECTIVEâ€"The objective of this study was to determine whether cinnamon improves blood glucose, triglyceride, total cholesterol, HDL cholesterol, and LDL cholesterol levels in people with type 2 diabetes. RESEARCH DESIGN AND METHODSâ€"A total of 60 people with type 2 diabetes, 30 men and 30 women aged 52.2 Â± 6.32 years, were divided randomly into six groups. Groups 1, 2, and 3 consumed 1, 3, or 6 g of cinnamon daily, respectively, and groups 4, 5, and 6 were given placebo capsules corresponding to the number of capsules consumed for the three levels of cinnamon.
http://care.diabetesjournals.org/content/26/12/3215

Is honey and cinnamon a hoax?
Honey is almost entirely composed of different sugars, with water and trace minerals. Excess sugar (and we really get enough carbohydrates of different sorts in our diet anyway) gets stored as fat. So this diet cannot work unless you reduce your food intake to compensate for the excess sugar you are consuming. Certain types of cinnamon may well have useful properties that might influence insulin levels, but probably not at the recommended concentrations.
http://www.answers.com/Q/Is_honey_and_cinnamon_a_hoax

Article: Cinnamon Helps Diabetics
Cinnamon a spice found in many kitchens around the world especially at Christmas time. Although it is used by Jewish people at Rosh Hashona (the Jewish New Year that falls in September each year). It has been found that this flavoursome spice has its uses for people who are Type 2 Diabetics and helps to control their Blood Sugar. What Is Diabetes? Diabetes is a disorder, where the body is unable to control the Blood Sugar. Insulin a hormone secreted by the Pancreas is essential in helping to control the Blood Sugar. Some unfortunate people are resistant to Insulin so they have to very strictly adhere to a reduced fat intake in their diet.
http://www.naturalhealthweb.com/articles/Moffat3.html

Instructions: Use the search engine results to help you determine the effectiveness of the specified treatment for the corresponding health issue. Once you believe you have determined the effectiveness of the treatment, go to the bottom of the page and click the "Submit Answer" button to submit your answer on the next page.

Category	Explanation
Helps	A medical treatment **helps** if the treatment is effective and has a direct positive influence on the specified illness.
Inconclusive	The effectiveness of a medical treatment is **inconclusive** if medical professionals are still unsure if the treatment will have a positive, negative, or no influence on the illness.
Does not help	A medical treatment **does not help** if the treatment is ineffective and either has no effect or has a direct negative influence on the specified illness.

Figure 5.1: User interface during the search task.

squares ensure that the experimental conditions are equally balanced over the *helpful* and *unhelpful* treatments. Finally, we randomize the columns of rows of these two Graeco-Latin squares. Repeating the previous steps would result in two new Graeco-Latin squares for helpful and unhelpful treatments. With these four Latin squares, we can generate a 10×10 Graeco-Latin square and we randomize its columns and rows and assign every row to a participant.

5.4 Results

In this section, we summarize, first, the main study results, then the confidence and knowledge findings and, finally, the click behavior during the search task.

Main Results Looking at the two independent variables in our user study, we can measure the effect of the controlled search results bias (correctness and rank) on the participants' ability to determine the effectiveness of the medical treatment. Table 5.2 shows the fraction of correct and harmful decisions for the 60 participants corresponding to the search result bias and the topmost correct rank. The table shows that when the topmost correct document is at rank 1 and there was a bias towards *correct*, the accuracy increased to 70% and the harm reduced to 6% from 20% in the control condition. On the contrary, when results are biased towards *incorrect*, the accuracy reduced to 23% from 43% in the control condition while the harm was doubled.

The statistical significance of these results is evaluated and reported in table 5.3. The table shows that search result bias has a statistical significance on both the accuracy and harm. However, the topmost correct rank is not statistically significant on accuracy but is statistically significant on harm with p=0.06. Even though the topmost correct rank did not show a strong effect on the results, we believe that the rank has an effect on the search behavior as the accuracy was 70% when topmost correct was at rank 1 compared to 59% accuracy when topmost correct was at rank 3. With larger study, we might be able to see a stronger effect of the topmost correct rank independent variable.

The results explained above show that search results have a strong effect on people's ability to determine the efficacy of medical treatments. When people are biased towards correct information, they perform better then when biased towards incorrect information. Further, we noticed that when exposed to incorrect information, participants perform worse than when no search results are provided.

Looking at the performance of participants for every medical treatment, table 5.1 shows that for nine out of ten medical treatments, the incorrect bias results in reduced accuracy compared to the control condition. The medical treatment T6 Does caffeine help asthma? (truth = helpful) does not behave as expected: the accuracy improves when there is a bias towards incorrect information. Further, when exposed to correct information bias, the accuracy increases compared to the control condition, except for the cases of T7 Does cinnamon help diabetes? (truth = unhelpful) and T9 Does surgery help obesity? We believe that a follow up study needs to be designed in order to speculate and analyze participants behavior for these specific medical treatments.

White and Hassan (2014)'s work suggests that participants as well as search engines have a strong bias towards positive information. This result also aligns with the work conducted by Kazai et al. (2019) where they showed that people are more likely to click on emotionally charged results than emotionless results. We investigate this trend by splitting our data by the medical treatment type of helpful, inconclusive, unhelpful to investigate the trends and behaviors of our participants. We observe in table 5.4 that, similar to the work conducted by White and Hassan (2014) and Kazai et al. (2019), there is an overall bias towards saying that the treatments are *helpful* in both the control and the experimental conditions. Specifically, in the control condition, participants answer the truly *unhelpful* medical treatments correctly as often as answering *inconclusive*. For the experimental conditions, participants answer the truly *unhelpful* medical treatments as *inconclusive* more

Independent Variables		Dependent Variables	
Results Bias	Topmost Correct Rank	Fraction of Decisions	
		Correct	Harmful
Incorrect	3	0.23 ± 0.04	0.41 ± 0.05
Incorrect	1	0.23 ± 0.04	0.35 ± 0.04
Control (No search results)		0.43 ± 0.05	0.20 ± 0.04
Correct	3	0.59 ± 0.05	0.13 ± 0.03
Correct	1	0.70 ± 0.04	0.06 ± 0.02

Table 5.2: Main user study results.

Independent Variable	Dependent Variable	Pr(>Chisq)
Search Results Bias	Correct Decision	$\ll 0.001$
Search Results Bias	Harmful Decision	$\ll 0.001$
Topmost Correct Rank	Correct Decision	0.16
Topmost Correct Rank	Harmful Decision	0.06

Table 5.3: Statistical significance of independent variables.

often than answering correctly the search task. One possible explanation would be that participants are looking for positive information and rather answer *inconclusive* than believing that the medical treatments are not helpful. When a treatment is truly *unhelpful*, searchers can be heavily influenced by search results with incorrect information, claiming that the treatment is *helpful*.

Knowledge and Confidence Before participants had to answer the search task, they were asked to rate their knowledge of the health treatment and health issue on a 5 point scale. Figure 5.2 shows the knowledge count of the health issue. Knowledge did not have a statistical significance on the dependent variables. However, we noticed that more knowledge resulted in higher fraction of correct results when biased towards incorrect information. Investigating more the knowledge, we decided to group the two highest levels of knowledge into one group *high* and the three lowest levels of knowledge into one group *low*.

Considering only the experimental condition when results are biased towards incorrect information, the fraction of correct decisions for *low health issue* knowledge was 0.19 ± 0.03 compared to 0.28 ± 0.04 for *high health issue* knowledge which was not statistically

Control Condition (No Search Results)

	Participant Decision			
Truth	Unhelpful	Helpful	Inconclusive	Total
Unhelpful	23	16	21	60
Helpful	8	29	23	60
Total	31	45	44	120

Experimental Conditions (Interact with Search Results)

	Participant Decision			
Truth	Unhelpful	Helpful	Inconclusive	Total
Unhelpful	79	64	97	240
Helpful	50	132	58	240
Total	129	196	155	480

Table 5.4: Confusion matrices.

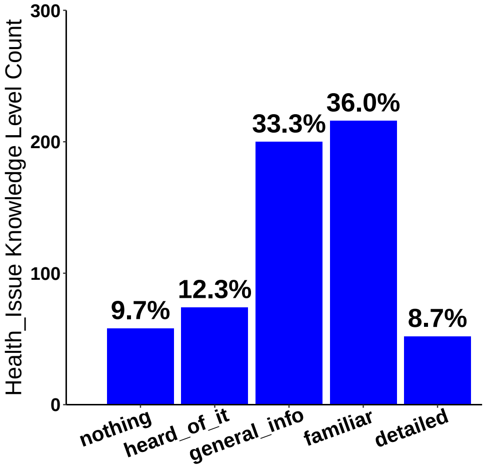


Figure 5.2: Health issue knowledge count.

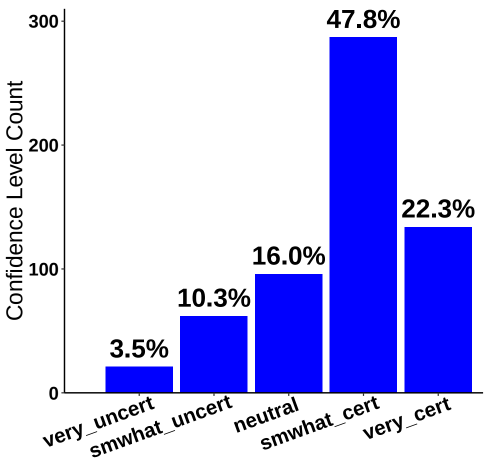


Figure 5.3: Confidence count.

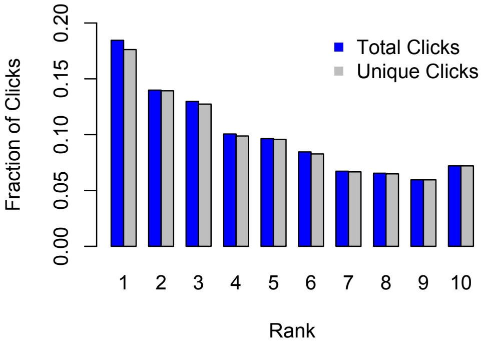


Figure 5.4: The fraction of total clicks and unique clicks for each of the 10 search result ranks.

significant (p=0.14). The fraction of correct decisions for *low medical treatment* knowledge was 0.20 ± 0.03 compared to 0.33 ± 0.06 for *high medical treatment* knowledge which was statistically significant (p=0.04). Knowledge of the medical treatment results in better performance when the results are biased towards incorrect information. Comparing the control condition with *high medical treatment* knowledge using two-sided t-test, we fail to reject the null hypothesis that they are the same (p=0.21). We conclude that knowledge of the medical treatment does not increase the accuracy above no exposure to search results.

After participants finished the search task, they were asked to evaluated the confidence of the answers on 5 point scale from 1="very uncertain" to 5="very certain". Figure 5.3 shows the count distribution of confidence after the search task. Results showed that less confident participants tend to answer *inconclusive* more frequently than participants who have high confidence (who tend to answer *helpful* or *unhelpful*).

Click Behavior To measure the level of interaction between participants and search results, we tracked the click behavior. Figure 5.4 shows the distribution of clicks over the search result ranks. First, we see that the difference between unique clicks and total clicks is bigger at results at rank 1 which means that participants find the document at rank 1 to be important and they click on it multiple times. Second, the overall distribution of unique clicks is similar to what is seen in real search results which proves that participants are interacting with the search task in a realistic fashion. Further, table 5.5 shows the average number of clicks for the correct decisions and the harmful decisions fractions for the four SERP experimental conditions. The average number of clicks was higher for the correct decisions 3.73 ± 0.2 compared to the incorrect decisions 3.32 ± 0.2. Further, the average number of clicks was lower for harmed decisions 3.03 ± 0.3 compared to unharmed decision

Dependent Variables	Average Number of Clicks
Harmed Decisions	3.02 ± 0.3
Unharmed Decisions	3.65 ± 0.3
Correct Decisions	3.73 ± 0.2
Incorrect Decisions	3.32 ± 0.2

Table 5.5: Average number of clicks for each dependent variable.

3.65 ± 0.2. The difference in the number of clicks was statistically significant and it shows that when participants interact more with the search results, they are less likely to make incorrect decisions.

5.5 Discussion

The key findings of this chapter can be summed up as follows:

Search result pages can significantly effect how people search for health questions online in a positive as well as a negative manner. When search results are biased towards *correct* information, the accuracy improved from 43% to an average of 65% (Table 5.2). On the other hand, when search results are biased towards *incorrect* information, the accuracy dropped to 23%(Table 5.2). More importantly, we noticed that people perform worse when exposed to *incorrect* information than when no search result pages where provided. As search results are a mixture of correct and incorrect information, this might potentially cause harm especially if people believe what they read online without further investigation.

We found that people have a uncontrolled bias towards believing positive information. In the health domain, this behavior is a result of people having hopes for finding cures to different diseases. Sharot et al. (2007) defined this behavior as the "optimism bias" and they concluded that, in general, people expect positive events in the future even when there is no evidence to support such expectations. This outcome is crucial in the health domain as people might believe in the Hoxsey therapies available in search results and might stop their medically proven treatments or they might harm their health by taking unproven medical treatments.

These findings suggest that information retrieval researchers need to improve the documents retrieval and, in addition to relevance, incorporate a notion of *correctness* in evaluating the documents. Achieving this, search results would contain less *incorrect* content and more high quality and trustworthy documents.

Current information retrieval techniques have the notion of non-relevant document as a document causing time loss. As a result, non-relevant documents are given a zero gain value. However, the notion of incorrect documents is not measured when it might cause harm. One possible novel solution to incorporate correctness is introducing a notion of negative gain and assign it to incorrect information content.

In this Chapter, we showed that people are significantly influenced with the search results by measuring their accuracy in answering health-related questions about the effectiveness of medical treatments with the help of search results. However, the study did not highlight potential factors that lead to such influence. As an extension to the work conducted in this study, in the next Chapter (Chapter 6), we design another study with the aim of understanding the potential factors affecting people's decisions when doing online search. Understanding these potential factors helps build better systems to support people's decision making in the health domain.

Chapter 6

A Think-Aloud Study about Medical Misinformation in Search Results

The majority of US internet users rely on web search to look for information about a health issue or a medical treatment (Fox and Duggan, 2013). However, there is an increased concern over the lack of accountability and dubious quality of this online content. Prior research (White and Hassan, 2014) has shown that search engines can be biased towards stating that medical treatments are helpful, regardless of the truth. Given the substantial impact of search engines on people's decision making, if results are biased towards incorrect information, people's accuracy reduce and there is a potential harm, especially if people believe what they read online (as Chapter 5 showed).

In Chapter 5, we showed that people are heavily influenced with the search results bias. Even when the correct answer is always at either rank 1 or 3, participants fail to find the correct information. To better understand the possible reasons behind this huge influence of incorrect results on peoples' decisions, we implement a new study using the think-aloud method. Collecting and analyzing think aloud protocols has been used in literature to build models of cognitive processes during a problem solving task (Van Someren et al., 1994). Applying the think-aloud method, we aim to gain some insight on the strategies used by participants while using search engines to answer health-related queries. The insights will be helpful to improve and build search engines that better support people's decision making.

In the think-aloud study, we ask participants to determine the effectiveness of four medical treatments. We provide participants with search result pages that help them answer the questions about the treatment's efficacy. While doing the task, we ask participants to say out loud what goes through their head by stating directly what they think. Later, we ask participants about their decisions during the task and about using search engines for health-related purposes.

In this Chapter, we first list our research questions, then we describe the study design. We later cover the details of the study material and methods and present the study's results, along with our conclusions.

6.1 Research Questions

The goal of the study presented in this Chapter is to understand how people use online search to answer health-related questions. Considering a question about the efficacy of a medical treatment, for example, we aim to answer the following specific research questions:

- R1. While thinking out loud, does the search results bias influence participants' decisions about the efficacy of medical treatments?
- R2. While thinking out loud, does the search results rank influence participants' decisions about the efficacy of medical treatments?
- R3. What are the factors that influenced participants' decisions about the efficacy of medical treatments when presented biased search results?

6.2 Study Design

First, we calibrate the eye tracking device to measure the participants eye movement. Next, participants sign the consent forms then fill out a questionnaire providing demographic information. Following the questionnaire, they read detailed instructions about the participation before proceeding with the study. After that, with the help of search results, participants have the chance to practice determining the effectiveness of a medical treatment. While doing the practice task, participants are asked to articulate and say their thoughts out loud. Later, we start video and audio recording of the participations. Then, participants begin the main study where they have to determine the effectiveness of four medical treatments while thinking out loud (Concurrent think-aloud).

While participants are doing this search task, we write down notes about the verbal and non-verbal interactions. After finishing the search task, we show participants the video recording of the participation, with their eye movements to help them remember their thoughts, and ask them questions about their decisions (Retrospective think-aloud). Finally, we ask participants general questions about their usage of search engines for health-related purposes (Questionnaire).

The study is designed as a web application and the search results are modeled as a traditional style of web search engine. In this study, we recreate the interface explained in the work presented in Chapter 5 (refer to Section 5.3 to get more details about the user interface design). We detail each step of the study as follows:

6.2.1 Think-aloud Protocol

During the think-aloud task, we ask subjects to articulate their thinking and decision making process while doing the search task (also called *concurrent think-aloud* - CTA). We choose to apply CTA as it is helpful in extracting immediate thoughts while doing the task (Kuusela and Pallab, 2000). This is helpful in order to reveal potential factors influencing the decision making process of people using online search to answer health-related questions.

We capture the think-aloud data by audio recording of the participation with the aid of a computer microphone. Further, we record the screen of the computer as the participant completes the search task using the Tobii Pro Studio software[1]. While participants articulate their thoughts, we note the non-verbal responses during the think-aloud in addition to the words said by the participant (such as pauses, smiles, misreading, periods of silence, pace of speech, body movements, tone variations and volume changes).

One known challenge of the concurrent think-aloud method is that participants might find it difficult to simultaneously articulate their thoughts while doing the search task (Kelly et al., 2009). In order to address this limitation, we implement a number of points. First, we restrict our recruitment process to only accept subjects with English as their first language. If English is the mother tongue, we believe that it is easier to express thoughts while performing the study tasks. Second, we design a practice task where participants complete a short training before starting the actual study. During the training, participants get a chance to think-out loud while determining the effectiveness of a medical treatment. Third, we believe that the study is suitable to apply the think-aloud protocol as the tasks are of intermediate level of difficulty (Charters, 2003). Finally, during the think-aloud task, there is no interaction between the participant and the searcher in order to not annoy or

[1] https://www.tobiipro.com/product-listing/tobii-pro-studio/

distract the participant. Instead, a "KEEP TALKING" sign is used to remind participants to talk and encourage the thinking-aloud.

6.2.2 Stimulated Retrospective Think-Aloud

After the concurrent think-aloud part, a *stimulated retrospective think-aloud* - RTA is used where we ask participants about their thoughts after completing the search task (Salkind, 2010; Kelly et al., 2009). The RTA, where participants are asked questions after finishing the study tasks, is a more natural activity than the concurrent think-aloud process (Kelly et al., 2009). We implement the RTA method as it is helpful in the case where participants do not verbalize enough the ideas. It is also a chance to deeper thoughts and better interpret and validate the CTA (such as asking about pauses and facial expressions etc.) (Kuusela and Pallab, 2000) .

In the stimulated retrospective think-aloud study, there is a delay between the study task and the discussion afterwards. In order to help participants remember their thoughts, we use an eye tracking during the search task. During the RTA part, while playing back the video recording of the concurrent think-aloud data, we show participants the captured eye movements to help them recall their thoughts and ideas. We perform eye tracking using Tobii Pro X3-120 [2] device mounted on the monitor.

Later, the participants are asked further **ad-hoc** questions about their decisions and interactions with the search results as the CTA video recording is being played. When formulating the questions, we pay special attention to not introduce any bias and we avoid leading questions. For this reason, we always make sure to ask questions that start with "What", "When", "Where" and "How". Examples of the questions we ask participants in this part are:

- What was it that made you decide to click on this specific page?
- How did you make up your mind and decided that the treatment is *unhelpful*?
- What did you think of the content in this web page?

6.2.3 Post-task Questionnaire

After the CTA and RTA the think-aloud parts, we perform a post-task questionnaire where we ask participants general questions about using online search for health purposes.

[2] https://www.tobiipro.com/product-listing/tobii-pro-x3-120/

Table 6.1: This table shows the list of post-task questions along with the counts of responses for each question.

No	Question	Yes	No	Maybe
1	Do you believe that exposure (i.e. most results say the treatment helps/does not help) is important in determining the effectiveness of the medical treatment? And why?	13	2	1
2	Do you believe that rank (i.e. highly ranked results say the treatment helps/does not help) is important in determining the effectiveness of the medical treatment? And why?	9	6	1
3	Do you believe that quality is important in determining the effectiveness of the medical treatment? And please elaborate on what quality means to you?	15	0	1
4	Do you believe that the web page layout is important in determining the effectiveness of the medical treatment? And why?	12	2	2
5	Do you believe that social factors (i.e. experience of other people you know such as friends, family etc.) is important in determining the effectiveness of the medical treatment? And why?	9	5	2
6	Did you notice any manipulation of the search results? If yes, then can you guess what was it?	9	7	0
7	How do you describe your experience with the think-aloud process?	-		

Further, this questionnaire is a change to gather feedback about the think-aloud experience. In this part, we ask **open** questions where subjects have the ability to provide responses in the way they prefer (no restricted choices). This type of questions is helpful to gain additional varied insights about the decision making process of participants while doing the search talk (Kelly et al., 2009). The full list of questions asked in the post-task questionnaire are shown in Table 6.1. While designing the questions, we pay special attention to the wording and make sure that the questions are not biased or bouble-barreled (Kelly et al., 2009).

Table 6.2: This table shows the medical treatments with their corresponding efficacy.

T	Medical Treatment	Efficacy
T1	Do antioxidants help female subfertility?	Unhelpful
T2	Do benzodiazepines help alcohol withdrawal?	Helpful
T3	Do probiotics help treat eczema?	Unhelpful
T4	Does caffeine help asthma?	Helpful
T5	Does cinnamon help diabetes?	Unhelpful
T6	Does melatonin help treat and prevent jet lag?	Helpful
T7	Does surgery help obesity?	Helpful
T8	Does traction help low back pain?	Unhelpful

6.3 Materials and Methods

6.3.1 Study Material and Performance Measures

We use the study material from the publicly available dataset[3] from the work explained in Chapter 5. We control search result content in terms of two levels. First, the search result bias which is either *correct* or *incorrect*. Second, the topmost correct search result where we place the first correct result at either rank 1 or 3. Further, we measure participants' performance by keeping track of the fraction of correct decision and the fraction of harmful decisions. Participants had to determine the efficacy of medical treatments as either *helpful*, *unhelpful*, or *inconclusive*.

Medical Treatments

We use a list of 8 medical treatments from the study designed in Chapter 5. Each medical treatment can either be: *helps* (the medical treatment has a direct positive influence on a specific illness), *inconclusive* (medical professionals are not sure about the effectiveness of the medical treatment) or *does not help* (the medical treatment has either a direct negative influence or no influence on a specific illness). Out of the 8 medical treatments, four were *helpful* and four were *unhelpful*. Table 6.2 shows the list of the medical treatments in the think-aloud study with their corresponding effectiveness.

[3] https://cs.uwaterloo.ca/~aghenai/user_study_pages.html

Search Results

During the study, we ask participants to pretend they have a question about the effectiveness of a medical treatment and have decided to use a search engine to help them answer the question. We show participants a web page that has ten search results, with the general appearance of a standard search engine results page (SERP). The search results are either biased towards correct or incorrect information. When biased towards correct, we show eight correct search result pages and two incorrect ones. When biased towards incorrect, we show participants eight incorrect search result pages and two correct ones. We further control for the rank of the topmost correct result page to either be at rank 1 or 3. We randomly assign the search results to the corresponding ranks from a pool of 8-10 correct and 8-10 incorrect documents. The search results setup is similar to the one used in Chapter 5.

Documents and Snippets

To build the SERP pages, we use the same 158 documents used in Chapter 5. Every document is either correct (contains information about the treatment efficacy that agrees with the truth) or incorrect (contains information about the treatment efficacy that contradicts with the truth). For every search result, we show the document's title, URL, and snippet. We use the same titles, URLs, and snippets used in the study explained in Chapter 5.

Performance and Statistical Significance

We measure the participants' performance in the user study by computing two different measures: the fraction of correct decisions and the fraction of harmful decisions. A participant's decision is correct if it agrees with the truth. Note that, inconclusive is considered an incorrect decision as all medical treatments are either helpful or unhelpful. Further, a participants' decision is harmful if it is opposite to the truth where inconclusive is not considered a harmful decision.

The fractions of correct and harmful decisions are the dependent variables. The search result bias and the topmost correct result are the independent variables. In order to measure the statistical significance of the independent variables on the fractions of correct and harmful decisions, we use generalized linear mixed effects model using R. More details about the modeling method can be found in Chapter 5 (see Sections 5.2.4 and 5.3).

6.3.2 Transcription

We video record the concurrent and retrospective think-aloud process while participants interact with the search results and audio record the questionnaire part. We recruited an outsource transcription service (REV[4]) to transcribe all the parts of the collected data. Before sending the audio files to the transcription service, we manually checked for the quality of transcription of a sample audio file. The transcription service includes timestamps for the transcribed scripts without verbatim (filler words are removed from the transcripts). The results reported in Section 6.4 are based on the transcribed data.

6.3.3 Coding Scheme

After transcribing the think-aloud recordings, we start the coding process in which we generate tags in order to quantify the observations during the think-aloud. We use QSR International's NVivo 12 qualitative data analysis software (Ltd, Version 12, 2018) for the coding process.

We perform qualitative analysis by introducing a set of categories used to summarize responses for the think-aloud data. Specifically, we perform open-coding using a mixed methods research for both the bottom-up and the top-down approach (Gu, 2014a; Kelly et al., 2009). Some of the codes were inspired by existing research such as prior belief (White, 2014) and rank (Allam et al., 2014; Haas and Unkel, 2017) (top-down). While other codes have been added and modified as we explore the think-aloud transcribed data such as advertisements, statistics and studies (bottom-up). Applying the mixed method approach, we aim to discover the possible strategies participants apply when using search engine to answer a health-related question.

The initial coding process was performed by myself. Then, the transcripts were re-coded once again by me at a later date. In order to increase the reliability of the coding, new coding rules were defined. We perform non-mutually exclusive codes in order to allow more than one code per item. We compute the intra-coder reliability which is helpful to verify the consistency and agreement of the codings generated in different time periods (Given, 2008). To test the intra-rater reliability, we compute Cohen's kappa (McHugh, 2012) of the codings in the two different time periods. Cohen's kappa is the ratio of difference between observed agreement and probability of chance agreement over probability of chance disagreement. Cohen's kappa is known to be more robust than a simple agreement percentage as it takes into account the possibility of the agreement occurring by chance (McHugh, 2012).

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

The first set of codes were developed a priori based on the research questions and the literature review, and were structured around the post-task questionnaire questions such as the majority, authoritativeness and prior beliefs. The second list of codes were developed inductively during the analysis process of the think-aloud data such as: advertisements, date and statistics and studies .When coding, we keep track of each coding occurrence to compute the frequency counts. Table 6.3 shows the list of codes with the number of participants mentioning the code as well as the corresponding references counts (i.e number of times the code was mentioned over all participants). We use this quantitative method in order to identify which of the codes are more or less important for participants during the decision making process.

6.3.4 Participants

We obtained ethics approval from the Office of Research Ethics at our university. Next, we recruited participants using posters and email announcements to different graduate student email lists. As the user study involved an English language think-aloud process and, in order for participants to be able share their thoughts easier, one of the recruiting requirements was to have only native English speakers. All participants gave their informed consent. Following their participation, we debriefed all participants and provided them with the correct answers regarding the efficacy of the medical treatments. We paid participants $15. Participants were 16 students (7 male, 9 female) from different majors (7 from engineering and mathematics, 8 from arts and sciences and 1 from environment) with an age between 18 and 28 years old (37.5% less than 20, 56.25% between 20 and 25 and 6.25% greater than 25, with an average age of 21).

6.4 Results

This section explains the results of the user study including the main findings showing the participants' performance during the search task. Further, we describe the the think-aloud data as well as the retrospective think-aloud results. Finally, we show the results of the post-task questionnaire part.

6.4.1 Main Results

Table 6.4 reports the fraction of correct and harmful decisions of the 16 participants during the study. We see that, similar to Chapter 5, results with bias towards correct informa-

Table 6.3: This table shows the list of codes with their corresponding description. Each code is assigned a label C1-C16 that we use throughout the chapter to refer to specific codes. The table also shows the number of participants mentioning a particular code, and the total number of references assigned to the code.

No	**Name**	**Description**	**Participants**	**References**
C1	*Majority*	The majority of the search results stating that the treatment helps or that the treatment does_not_help or looking for a consensus of different search results.	14	36
C2	*Authoritativeness*	The trustworthiness and reliability in the content of the search results page.	13	153
C3	*Statistics & Studies*	The presence of statistics, numbers and detailed research studies in the search results page.	12	20
C4	*Advertisements*	The presence of messages to promote or sell a product, service or idea in a search results page.	7	16
C5	*Date*	The date and time the search results page was first published to the public or the dates mentioned in the page content reflecting how old the information is.	7	15
C6	*References*	Having a list of sources that have been cited to support the information in the search results page.	7	12
C7	*Negative information*	Mentioning negative information about the treatment in the search results page such as listing the side effects or explaining the dangers of using the treatment etc.	6	15
C8	*Information representation*	The information related to the style of the content presented in the search results page such as list versus grid representation, colors, the page layout, capital letters and special characters etc.	5	18
C9	*Prior belief*	Trusting the information that agrees with our prior knowledge and disregarding facts that contradict with it, regardless of the actual truth (White and Horvitz, 2015).	5	8
C10	*Readability*	The style of writing and the quality of content being easy to read (Feng et al., 2010a).	4	8
C11	*Relevance*	The relevance to the topic about the effectiveness of the medical treatment.	4	7
C12	*Past experience*	Having a prior experience with the topic (either the medical condition or the treatment) that may effect how much we trust the information in the search results page regardless of the factual correctness.	3	3
C13	*Text length*	The amount of text content in the search results page which might impact the reliability. For example, longer explanations might lead to higher levels of trust.	3	3
C14	*Images*	The presence of visuals in the search results page. The intuition behind this is that images might help better remember the information which may interfere with the decision making process.	2	6
C15	*Rank*	The order of search results in the SERP page that might effect the trustworthiness and reliability of the sources.	2	4
C16	*Social factor*	Relate the information about the topic to people we know. For example, whether a friend or a family member's opinion effects our preferences and decision making.	1	2
		Overall	16	326

Table 6.4: Main results. Based on the decisions the 16 participants made, we compute the fraction of correct and harmful decisions. Fractions are shown along with their standard errors.

Results Bias	Fraction of Decisions	
	Correct	Harmful
Correct	0.67 ± 0.08	0.06 ± 0.03
Incorrect	0.32 ± 0.06	0.28 ± 0.06

Table 6.5: Statistical significance of independent variables.

Independent Variable	Dependent Variable	Pr(>Chisq)
Search Results Bias	Correct Decision	$\ll 0.001$
Search Results Bias	Harmful Decisions	$\ll 0.01$
Topmost Correct Rank	Correct Decision	0.8
Topmost Correct Rank	Harmful Decisions	0.05

tion lead to an increased accuracy up to 67% while lowering harmful decisions to 6%. Conversely, results biased towards incorrect information reduce accuracy to 32% while increasing harmful decisions to 28%.

Table 6.5 reports the statistical significance of the search results bias and topmost correct rank on the correct and harmful decisions. Similar to the results in Chapter 5, we find that the search result bias has a statistically significant effect on the fraction of correct decisions and harmful decisions. Due to the smaller sample (16 participants in this study compared to 60 participants in Chapter 5), we find that the topmost correct rank has less of an effect on the correct and harmful decisions.

As verbalization makes people take longer time doing the search tasks (39 minutes average participation time in the Chapter 5 study compared to 65 minutes in the current think-aloud study), we expect people to be more conscious about their decisions and search results bias to have less or no effect on people's decisions. However, results demonstrated once again that search results have a potentially strong effect on people's decisions.

White and Hassan (2014) as well as the study presented in Chapter 5 demonstrated that participants have a strong bias towards believing that treatments are helpful ("optimism bias"). Looking at the current think-aloud data, we split the medical treatment types into "helpful", "unhelpful" and "inconclusive" treatments to further investigate this trend. Similar to White and Hassan (2014) as well as Chapter 5 findings, the results in Table 6.6 show that helpful is the most frequent option people tend to answer during the study.

Table 6.6: Confusion matrices. This table shows the decisions made by the study participants regarding the efficacy of the 2 helpful and 2 unhelpful medical treatments.

Truth	Participants			Total
	Unhelpful	Helpful	Inconclusive	
Unhelpful	13	6	13	32
Helpful	5	18	9	32
Total	18	24	22	64

Furthermore, participants are more likely to answer inconclusive more frequently than what we observed in Chapter 5 i.e., when thinking out loud, people tend to respond inconclusive more frequently than when not thinking out loud.

6.4.2 Think-aloud Method

The coding process shows some insights of the potential reasons why people are influenced with the search results even when the correct answer is always placed at either rank 1 or 3. The average participation time of the concurrent think-aloud part is 39 minutes with a maximum participation of 1 hour and 39 minutes and a minimum of 14 minutes. Table 6.3 shows the number of participants mentioning each code and the total number of references for that corresponding code. The codes are arranged in a descending ordered by the number of participants then references. In this table, we only report the coding performed during the first time period for two main reasons. First, the Cohen Kappa inter-rater ratio computed was computed between the coding of the two different time periods and the overall value was 0.7 (See Table 6.7) which is a substantial inter-rater ratio (McHugh, 2012). Second, we reached the same main results with both codings. The main coding results are described as follows:

First, from the transcribed data, 14 out of 16 participants mentioned *Majority* with a total number of 36 mentions. Majority means that participants try to find out what most websites state about the treatment effectiveness or try to look for an agreement between them. If participants are exposed to results geared towards a specific direction, they end up being influenced by what the majority of the search results state. This finding explains why search result bias (in both this study and the study presented in Chapter 5) has a significant effect on people's decisions. Here, we provide examples of the majority effect from the think-aloud transcript with the participant number in parentheses:

Table 6.7: This table shows the agreement percentages and the Cohen Kappa ratio for the coding done in different periods of times. A is the coding done first and B is the coding done afterwards.

No	**Name**	**Kappa**	**Agreement**	**A And B (%)**	**Not A And Not B (%)**	**Disagreement**	**A And Not B (%)**	**B And Not A (%)**
C1	Majority	0.89	99.71	1.23	98.48	0.29	0.06	0.23
C2	Authoritativeness	0.9	98.85	5.31	93.54	1.15	0.53	0.62
C3	Statistics & Studies	0.65	99.48	0.49	98.99	0.52	0.31	0.21
C4	Advertisements	0.89	99.9	0.39	99.51	0.1	0.06	0.04
C5	Date	0.9	99.88	0.54	99.34	0.12	0.08	0.04
C6	References	0.74	99.8	0.27	99.53	0.2	0.16	0.04
C7	Negative information	0.96	99.94	0.72	99.22	0.06	0.04	0.02
C8	Information representation	0.73	99.61	0.54	99.07	0.39	0.18	0.21
C9	Prior_belief	0.87	99.92	0.27	99.65	0.08	0.08	0
C10	Readability	0.71	99.82	0.21	99.61	0.18	0.14	0.04
C11	Relevance	0.69	99.85	0.18	99.67	0.16	0.04	0.12
C12	Past experience	1	100	0.08	99.92	0	0	0
C13	Text_length	0.82	99.95	0.14	99.81	0.06	0	0.06
C14	Images	0.92	99.96	0.21	99.75	0.04	0	0.04
C15	Rank	0.87	99.96	0.14	99.82	0.04	0.04	0
C16	Social factor	1	100	0.06	99.94	0	0	0
Overall Weighted Kappa		**0.7**						

> **(Participant 5)** *I'm going to say helps because a lot of people, like it was just, the vast number were in agreement.*
> **(Participant 6)** *So I'm seeing a lot of doctors recommending the melatonin pill. Yeah, I think this helps.*
> **(Participant 9)** *I think that's the common trend that we're seeing. So I'm going to submit and say that it does help.*

It is important to note that some people look at search results as individuals having opinions (Participants 5 and 6 in the above examples). They lean towards a specific direction because they believe that the majority of search results reflects the majority of opinions in real life which is a potentially dangerous misconception.

Further, we find that 45% of the total codes are about *authoritativeness* with 13 participants talking about it and a total of 153 references. Authoritativeness refers to the amount of reliability and trustworthiness towards a specific content. We observed that participants talk about authoritativeness in three different ways: 40% of the time people state that the content is not authoritative (negative authoritativeness), 34% of the mentions state that the content is trustworthy (positive authoritativeness) and the remaining 26% are about not being sure whether or not to trust the content (neutral authoritativeness). Bellow, we show some examples of each case from the think-aloud transcript:

> **(Participant 17)** *Health.com, I've seen it before, not really ... I don't really rely on it for information the first time I see it.*
> **(Participant 10)** *WebMD. It's a more trust worthy source, I think.*
> **(Participant 14)** *Okay. I don't really know what this website is. Medications for management of alcohol withdrawal.*

The high percentage of mentions about authoritativeness shows the importance of this factor to participants when evaluating the effectiveness of treatments. When we designed the study in Chapter 5, we did not control for the authoritativeness of search results i.e. correct answers might be in non-authoritative web pages. Doing this, we potentially harmed participants' performance especially with an incorrect search results bias. This might be another possible reason why people have been heavily influenced during the study.

Participants talk about many clues that define the quality of search results during the think-aloud. Concepts C3-6, C8, C10 and C13-14 in Table 6.3 are all about quality. For example, 12 participants mention 20 times the statistical analysis and detailed research studies during the think-aloud process (C3) in order to evaluate the quality of information in the search results. Examples of such beliefs can be found in these bellow transcribed participations:

> **(Participant 12)** *...so this is explaining a study. Who had been given cinnamon reduced their blood sugar by 18 to 29 percent. Well that seems like some good numbers. So that's interesting. I think, based on that, I'd probably say that it helps because it had really evidence from a study.*
> **(Participant 15)** *So this looks like a research study, so I think it's pretty reliable.*

We further note some notion about prior beliefs during the think-aloud (C9) where 5 participants mentioned this concept a total of 8 times. Bellow, we show some examples:

> **(Participant 16)** *And I was also taught from school that benezenes are harmful to health so though I might be bias I have this thought that benzene would not exactly help with certain health concerns.*
> **(Participant 3)** *So this Kurt Donsbach, PhD ... He will claim that it has no positive function at all, but I've heard different, so right away I'm not convinced by this page.*

We also coded the concept of rank where Table 6.3 shows that only 2 participants out of 16 mentioned rank a total of 6 times. We show an example from the think-aloud transcript bellow:

> **(Participant 19)** *I'll just go to the first link, even though it's wikiHow, it is the first link. I don't really know much about search engines, but I feel like the first link ... they're trying to give you the most helpful link. So I'll just open it, but still.*

Looking at the transcribed data, people rarely talk about rank when, in prior research (Allam et al., 2014; Haas and Unkel, 2017), authors showed that rank has a potential effect on people's decisions. A possible explanation is that people are unconsciously influenced with the higher ranked search results, however, they are not aware of its effect. Furthermore, this finding shows the limitation of the think-aloud method when used to explore cognitive biases during the decision making process. An experiment similar to the one proposed in Chapter 5 can be used to better study the rank with a larger number of participants.

Additionally, none of the participants mention anything about having preference towards material stating treatments are helpful. The optimism bias shown in Table 6.6 as well as in Chapter 5 is also another example of unconscious bias that such think-aloud study fails to reveal.

6.4.3 Retrospective Think-aloud

We audio record, transcribe and summarize the data gathered during the retrospective think-aloud. The average participation time of the retrospective think-aloud part is 25 minutes with a maximum participation of 37 minutes and a minimum of 17 minutes. Appendix B lists the summaries of all participations' retrospective think-aloud part. Looking at the summarized participations is helpful in giving insights of new strategies participants used during the study that might not be captured during the concurrent think-aloud part. Here is a list of strategies caught from the retrospective think-aloud summaries (between parentheses we specify the participant mentioning the strategy):

- Reading pages that state the medical treatment *does not help* in order to understand the opposite arguments. (Participants 3, 7 and 18)
- Finding reliable sources first, then quickly checking relevant less reliable websites (such as answers.com and Yahoo answers) in order to look for consistency. (Participant 4)
- Reading the search result page first to understand the causes of the health issue, before reading about the effectiveness of the medical treatment. (Participant 7)
- Using the first clicked on link as a base reference for all future websites the participant decides to look at. (Participant 8)
- In case no consistency exists between search results, the participant tries initiating a new search query with different keywords. (Participant 9)
- In case of no consistency between search results, the participant looks at the dates the information was published in order to check whether the non-agreement happens because of time difference. (Participants 9 and 17)
- When the same hostname/website appears more than once in the SERP page, the participant believes that this is a reliable source. (Participant 9)
- Deciding which websites to click on by looking at URL titles to check whether they contain the exact words as the search keywords (Participant 13).
- Whenever there is a website that does not give a clear answer about the treatment efficacy (for example a discussion on when the treatment helps and when it does not help), the participant opens the URL and reads more details about both sides of the story. (Participant 13)

- The participant trusts a non-credible website when there are other websites that state the same information as the non-credible website. (Participant 16)
- The participant only opens websites based on prior experience i.e. the participant opens websites that have been reliable and helpful in the past and does not trust or does not open websites that are not familiar. (Participant 17)

6.4.4 Post-task Questionnaire

Table 6.1 shows the list of questions as well as the participants answers. From the table, we can notice that 13 out of 16 participants believe that exposure is important when evaluating the search result pages where we define exposure as what the majority of the search results state. This answer aligns with what we observed in the think-aloud transcriptions as majority was mentioned by 14 participants in total. Participants are consciously aware of the influence of majority while evaluating the treatments effectiveness because they possibly believe that majority reflects real life opinions.

Further, when explicitly asked about the rank, only 57% of the participants believe that rank is important in evaluating the search result pages. Similarly, looking back at the think-aloud data, we observed that only 2 participants mention rank during the think-aloud task. Again, this shows that rank is a subconscious factor that effect people's decision making while doing online search.

Next, 15 out of 16 participants strongly believe that quality is important in doing online search. Eleven participants explained quality as a notion of authoritativeness, while two participants believe that quality has to do with readability and three participants stated that layout is the major factor to determine the quality of websites. When specifically asked about the layout, 12 out of 16 participants believe that the page design is important in evaluating the search results page.

When asked about the social factor (i.e the experience of other people we know such as friends and family etc.), only 9 out of 16 participants believe that it is important in evaluating search results. Social factor is a type of subconscious bias where people tend to believe that family and friends do not effect the decisions when they, subconsciously, do.

We, further, ask participants whether they noticed any manipulation of the search results during the study and, if they did, we ask whether they could guess the factor of manipulation. Seven out of sixteen participants could feel that there is a manipulation while seven participants could not notice any manipulation. Four participants guessed that rank was the manipulation factor while another four suggested authoritativeness as the manipulation factor. Two participants though that the URL was changed during the study design. One participant suggested that we introduced duplicate results and one participant felt the manipulation was about correctness (which was the only right guess among all participations). The participants' responses show that we successfully designed the deception part of the study which confirms that participants behaved normally (without any behavioral influences that would make the observations invalid).

Finally, when asked about the experience of participants with the think-aloud process, five participants found the study interesting and insightful, five participants found the study representing a good experience, two participants found the think-aloud part to be hard because of the thinking while talking process, and one participant found himself being more conscious about the decisions while stating them out loud.

6.5 Conclusion

Search result pages potentially contain a number of incorrect information when people perform online search about the effectiveness of medical treatments. With the effect of content bias, people are being influenced and potentially harmed. In order to build systems that better support people's decisions, we need to gain insights about strategies people use during this decision making process. Understanding the cognitive biases while using search engines to answer a health-related question is a complex phenomenon, mainly because there is a large number of biases and unconscious factors effecting the decision making process. In this study, we perform a think-aloud method where we ask participants to verbalize their thoughts while using search results to decide about the effectiveness of a medical treatment. Results revealed some strategies people use doing online search for health-related topics.

We expected that the think-aloud process would lessen the effect of the search result bias, for people were asked to be conscious of their decision making process and carefully perform the task in front of a researcher. However, people were still significantly influenced by the incorrect information, which shows how much search bias can affect people's decision making.

Additionally, biased content leads people to believing this reflects real life opinions. The implications are profound when, for example, searching for cancer treatments on today's popular web search engines that might return a mix of correct and incorrect results.

Finally, when people use search engines to answer health questions, there are many factors that, unconsciously, effect their decision making such as rank as per prior research. The think-aloud study fails to show such biases when studying the decision making process.

Chapter 7

Conclusion

In this section, we summarize the PhD research findings listing our contributions and the piratical implications of the research, then describe our plans for future work.

7.1 Summary

This PhD research involved two mixed-method approaches to address the health misinformation topic in online search and social media. First, We sought to understand how rumors affect real health-care behavior, using social media observational studies. Then, we measured the effect of misinformation in search results using a controlled laboratory study. We later extended the work using a think-aloud method to gain insights into people's cognitive biases when using online search to determine the effectiveness of a medical treatment.

In Chapters 3 and 4, we contributed to the rumor detection in social media domain as follows (1) extracting rumors in social media content is not trivial where hight-precision approaches (such as keyword search) captures only half of the actual rumor content; (2) rumors vary in terms of longevity and severity and different techniques should be built to address the different types of rumors; (3) we show that we can successfully build an automatic rumor classifier; (4)we can further build a tool to detect users prone to propagating health rumor-mongering; (5) we present a foundation for examining behavioral characteristics and interests of people susceptible to rumors; (6) the dataset collected in Chapter 4 presents a highly curated resource for the research community's future studies.

In Chapters 5 and 6, we contributed to the health misinformation in online search research area as follows (1) people can be significantly influenced by the information present in search result pages. When biased towards correct information, the accuracy of answering health-related questions increases from 43% to 65%. On the other hand, when there is a bias towards incorrect information, the accuracy decreases from 43% to 23%; (2) the rank of the topmost correct result has some effect on people's accuracy; (3) people have an uncontrolled bias towards believing that treatments are helpful, regardless of the ground-truth; (4) even with the think-aloud, participants were heavily influenced by a search result bias; (5) people pay a large amount of attention to what the majority of the search results state in addition to looking for clues for the quality and authoritativeness.

Listing our research contributions, here we now discuss the broader implications of this research.

Applicability to Search Engines

As current search engines return relevant documents containing a mixture of correct and incorrect information, this work showed that incorrect documents might be harmful. First, given this finding, we need to address this problem by improving the retrieval techniques of the current algorithms to incorporate a notion to measure incorrect information. Incorrect documents might be given a negative gain value in the retrieval process to push such documents down in the search results. This might protect searchers from incorrect information help them make more informed decisions.

Second, prior research has shown that there is a bias towards stating that treatments are helpful when using online search to look for the effectiveness of medical treatments. Further, the proposed work showed that people tend to agree with what the majority of the search results state. This trend is dangerous when search results state incorrect information about the effectiveness of an unproven cancer treatment. Current search engines need to introduce a measure to the current search results' helpfulness bias. Search result pages need to be presented in a neutral way to prevent influencing searchers towards incorrect information.

Educational Efforts

Gathering verbal protocols when people perform health-related search was insightful in understanding how people perform online search. The findings suggested that people potentially have some misconceptions of how search engines work. For example, a number of participants during the conducted study stated that they believe the majority of what search results state reflects the real-world opinion majority - which is not the case. We need to build efforts to educate the public on how search engines work. Starting from the early stages of the educational system, we need to incorporate subjects that teach people how to use search engines properly and about the potential existing search results biases . These efforts might help people better use online search and make the correct decisions.

Public Health Monitoring

We propose a tool to detect and track health-related rumors on Twitter. This work is important for improving health crisis management, as it can enable health authorities detect and track health rumors in a timely fashion. Thus, it is imperative that the impact of health-awareness campaigns is monitored in real time, as well as internationally. The tools described here can help public health practitioners in tackling the large scale of social media streams.

Use of Persuasive Technologies

In Chapter 4, we presented a tool to automatically identify users spreading toxic communication and detect new misinformation long before it causes serious harm. In the age of personalization, this model can be employed to target individuals potentially susceptible to follow questionable accounts, consume poor quality health information, and propagate it. Early identification of rumor-prone individual accounts allows for refining traditional broad-spectrum information campaigns via personalized employment of *persuasive technologies* which offer a way to tailor content to the individual and track individual progress (Berkovsky et al., 2012). In particular, such technologies attempt to nudge the user to change his or her attitude or behavior through persuasion or social influence (but not misinformation or coercion), and have already been applied to health coaching and communication (Cugelman, 2013; Yardley et al., 2015)

7.2 Future Work

A number of open questions have not been answered in this thesis. Some are related to online search, while others are related to social media sites.

7.2.1 Authoritativeness in Online Health Search

When designing the study in Chapters 5 and 6, we did not control for the authoritativeness of the search results. This means that correct results might have been present on non-authoritative pages which might have harmed the participants' performance. Participants pay a huge amount of attention to authoritativeness, as the results of the study in Chapter 6 showed, so it would be interesting to see how authoritativeness effects the decision making process. If we make sure that authoritative sources contain the correct information, will this protect people from the incorrect information present in non-authoritative search results? This question warrants further investigation.

7.2.2 Rank Factor in Online Health Search

In the work described in Chapter 5, we noted that rank has a potential influence on people's accuracy when determining the effectiveness of a treatment. The study showed that rank is important, but then results were not statistically significant. We believe that other experimental conditions might reveal the importance of rank in online health search. In this context, possible directions to explore may involve experimenting with more rank conditions rather than manipulating only the first three ranks. To have a higher controlled bias, we could craft the first four search results. Specifically, to bias towards correct information, we might place three correct results and one incorrect result at ranks 1 to 4 and place the remaining results randomly. To bias towards incorrect information, we could place three incorrect results and one correct result ranked from 1 to 4 and place the remaining results randomly. With this setup, we might obtain stronger results suggesting that rank does influence people's decisions about the efficacy of medical treatments in online search.

7.2.3 Population Differences and Query Format

When recruiting participants for the study in Chapters 5 and 6, we considered a medium-sized population of limited variety (students). Future work might involve replicating the study with a larger and more varied population to check whether different groups of people are influenced differently, as Epstein and Robertson (2015) demonstrated. Further, in relying on White and Hassan (2014)'s medical treatments set, we concentrated only on a specific query format. It would be worth examining the influence of different question types on people's decision making.

7.2.4 Analyzing Verbal Utterances

Another possible future direction might address the unexpected behavior of a few topics, like T6 and T7, as discussed in the results section of Chapter 5. Additional information could be incorporated into the study, such as analyzing verbal utterances, would record previously suggested by Kammerer et al. (2013) in which a post-study task would record participants' opinions about those specific topics.

7.2.5 Theories of Rumor-mongering

The psychological underpinnings of believing and propagating rumors have been studied in the context of larger societal impact. In post WWII Europe, Allport and Postman (1947) studied rumors out of concern for their potential to damage morale and national safety. In the ample literature following, rumors have been defined as public communications that are infused with private hypotheses about how the world works, in particular to help us cope with our anxieties and uncertainties (Rosnow, 1991). Hypothesizing about the nature of rumor, Allport and Postman postulated that the strength of a rumor (R) will vary with the importance of the subject to the individual concerned (i) multiplied by the ambiguity of the evidence pertaining to the topic at hand (a), or $R \approx i \times a$ (Rosnow and Foster, 2005). Contributing to this theoretical work, our empirical analysis of social media health misinformation in Chapter 4 shows that those engaged in spreading unproven cancer treatments are for the most part not personally involved in the matter. Thus, we propose to extend the definition of "importance" i to other motivating factors beyond the personal, which may be at play in the public sphere of Twitter. Further analysis is required to reveal motivating factors in such health misinformation spread, as cancer fraud has been acknowledged in oncology literature (Vogel, 2011).

7.2.6 False Advertisements Spread in Social Media

Existing channels for disseminating information about unproven cancer cures currently include statements by the US Food and Drug Administration, which can ban products and companies for what the agency determines to be false advertising (Urciouli, 2016). Extending the work proposed in Chapter 4, it would be interesting to identify who is spreading such information and how it is reaching the public.

References

S. Abbar, Y. Mejova, and I. Weber. You tweet what you eat: Studying food consumption through twitter. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3197–3206. ACM, 2015.

A. Abbasi, T. Fu, D. Zeng, and D. Adjeroh. Crawling credible online medical sentiments for social intelligence. In *2013 International Conference on Social Computing*, pages 254–263. IEEE, 2013.

A. Abkowitz. China issues new internet search rules following baidu probe; regulator mandates 'objective, fair and authoritative results'. *Wall Street Journal (Online)*, Jun 26 2016.

A. Abkowitz and J. Chin. China orders baidu to revamp advertising results in online searches; action comes a week after government probe of company practices, May 10 2016.

V. Agarwal, L. Zhang, J. Zhu, S. Fang, T. Cheng, C. Hong, and N. H. Shah. Impact of predicting health care utilization via web search behavior: a data-driven analysis. *Journal of medical Internet research*, 18(9):e251, 2016.

A. Aker, H. Gravenkamp, S. J. Mayer, M. Hamacher, A. Smets, A. Nti, J. Erdmann, J. Serong, A. Welpinghus, and F. Marchi. Corpus of News Articles Annotated with Article Level Subjectivity. In *Workshop on Reducing Online Misinformation Exposure - ROME*, 2019.

A. Allam, P. J. Schulz, and K. Nakamoto. The impact of search engine selection and sorting criteria on vaccination beliefs and attitudes: two experiments manipulating google output. *Journal of medical Internet research*, 16(4):e100, 2014.

G. W. Allport and L. Postman. *The psychology of rumor.* Oxford, England: Henry Holt, 1947.

T. Almeida, R. Comber, and M. Balaam. Hci and intimate care as an agenda for change in women's health. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2599–2611. ACM, 2016.

S. Amir, G. Coppersmith, P. Carvalho, M. J. Silva, and B. C. Wallace. Quantifying mental health from social media with neural user embeddings. *arXiv preprint arXiv:1705.00335*, 2017.

S. Amir, M. Dredze, and J. W. Ayers. Mental health surveillance over social media with digital cohorts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 114–120, 2019.

M. L. Antheunis, K. Tates, and T. E. Nieboer. Patients' and health professionals' use of social media in health care: motives, barriers and expectations. *Patient education and counseling*, 92(3):426–431, 2013.

M. Araújo, Y. Mejova, I. Weber, and F. Benevenuto. Using facebook ads audiences for global lifestyle disease surveillance: Promises and limitations. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 253–257. ACM, 2017a.

M. L. D. Araújo, Y. Mejova, M. Aupetit, and I. Weber. Visualizing health awareness in the middle east. In *ICWSM*, pages 725–726, 2017b.

A. Arseniev-Koehler, H. Lee, T. McCormick, and M. A. Moreno. # proana: pro-eating disorder socialization on twitter. *Journal of Adolescent Health*, 58(6):659–664, 2016.

S. Bagroy, P. Kumaraguru, and M. De Choudhury. A social media based index of mental well-being in college campuses. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1634–1646. ACM, 2017.

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

J. Belbey. The FDA And Social Listening For Adverse Effects. *Forbes*, 2016. URL https://www.forbes.com/sites/joannabelbey/2016/07/15/the-fda-and-social-listening-for-adverse-effects/.

M. Benigeri and P. Pluye. Shortcomings of health information on the internet. *Health promotion international*, 18(4):381–386, 2003.

S. Berkovsky, J. Freyne, and H. Oinas-Kukkonen. Influencing individually: Fusing personalization and persuasion. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(2):9, 2012.

A. Bhatnagar, S. Mittal, and A. Garg. Laetrile: A wonder drug or farce? *International Journal of Applied Dental Sciences*, 2017.

R. Blaskiewicz. Skeptic activists fighting for burzynski's cancer patients. *Skeptical Inquirer*, 2016.

L. Bode and E. K. Vraga. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, 65(4):619–638, 2015.

L. Bode and E. K. Vraga. See something, say something: Correction of global health misinformation on social media. *Health communication*, 33(9):1131–1140, 2018.

A. Broom and P. Tovey. The role of the Internet in cancer patients' engagement with complementary and alternative treatments. *Health:*, 12(2):139–155, 2008.

N. Calabretta. Consumer-driven, patient-centered health care in the age of electronic information. *Bulletin. Medical Library Association*, 90(1):32–37, 2002.

M.-A. Cartright, R. W. White, and E. Horvitz. Intentions and attention in exploratory health search. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 65–74. ACM, 2011.

C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 675–684. ACM, 2011.

Centres for Disease Control and Prevention. Clinical guidance for healthcare providers for prevention of sexual transmission of zika virus. http://www.cdc.gov/zika/hc-providers/clinical-guidance/sexualtransmission.html, November 2016. Accessed: 2016-08-02.

S. Chancellor, J. A. Pater, T. Clear, E. Gilbert, and M. De Choudhury. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1201–1213. ACM, 2016.

J. Chang. Texas Medical Board sanctions controversial cancer doctor Burzynski, mar 2017. URL https://www.mystatesman.com/news/texas-medical-board-sanctions-controversial-cancer-doctor-burzynski/L9lDsfNTbBOuaWWqLEDBUI/.

E. Charters. The use of think-aloud methods in qualitative research an introduction to think-aloud methods. *Brock Education Journal OLD*, 12(2), 2003.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

X. Chen and L. L. Siu. Impact of the media and the internet on oncology: survey of cancer patients and oncologists in canada. *Journal of Clinical Oncology*, 19(23):4291–4297, 2001.

Y. Chen, N. J. Conroy, and V. L. Rubin. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, pages 15–19. ACM, 2015.

A. Cipriani, T. Furukawa, and C. Barbui. What is a cochrane review? *Epidemiology and psychiatric sciences*, 20(03):231–233, 2011.

D. Coady. Rumour has it. *International Journal of Applied Philosophy*, 20(1):41–53, 2006.

E. W. Coiera and V. Vickland. Is relevance relevant? user relevance ratings may not predict the impact of internet search on decision outcomes. *Journal of the American Medical Informatics Association*, 15(4):542–545, 2008.

M. Cord and P. Cunningham. *Machine learning techniques for multimedia: case studies on organization and retrieval.* Springer, 2008.

B. Cugelman. Gamification: what it is and why it matters to digital health behavior change developers. *JMIR Serious Games*, 1(1), 2013.

T. Cunha, I. Weber, and G. Pappa. A warm welcome matters!: The link between social feedback and weight loss in/r/loseit. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1063–1072. International World Wide Web Conferences Steering Committee, 2017.

B. David, N. Andrew, and J. Michael. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, Jan 2003.

C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 273–274. International World Wide Web Conferences Steering Committee, 2016.

M. De Choudhury and E. Kıcıman. The language of social support in social media and its effect on suicidal ideation risk. 2017.

M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting Depression via Social Media. *International Conference on Web and Social Media (ICWSM)*, 13:1–10, 2013.

M. De Choudhury, M. R. Morris, and R. W. White. Seeking and sharing health information online: comparing search engines and social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1365–1376. ACM, 2014.

M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *CHI*, pages 2098–2110, 2016.

S. de Lusignan and C. van Weel. The use of routinely collected computer data for research in primary care: opportunities and challenges. *Family Practice*, 23(2):253–263, 2006.

S. Dhoju, M. Main Uddin Rony, M. Ashad Kabir, and N. Hassan. Differences in health news from reliable and unreliable media. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 981–987. ACM, 2019.

G. Dong and H. Liu. *Feature engineering for machine learning and data analytics*. CRC Press, 2018.

M. Dredze, D. A. Broniatowski, and K. M. Hilyard. Zika vaccine misconceptions: A social media analysis. *Vaccine*, 34:3441–3442, 2016.

W. Du and Z. Zhan. Building decision tree classifier on private data. In *Proceedings of the IEEE international conference on Privacy, security and data mining-Volume 14*, pages 1–8. Australian Computer Society, Inc., 2002.

M. Duggan and J. Brenner. *The demographics of social media users, 2012*, volume 14. Pew Research Center's Internet & American Life Project Washington, DC, 2013.

A. G. Dunn, J. Leask, X. Zhou, K. D. Mandl, and E. Coiera. Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: an observational study. *Journal of medical Internet research*, 17(6), 2015.

C. J. Dy, S. A. Taylor, R. M. Patel, A. Kitay, T. R. Roberts, and A. Daluiski. The effect of search term on the quality and accuracy of online information regarding distal radius fractures. *Journal of Hand Surgery*, 37(9):1881–1887, 2012.

M.-D. Ebel, I. Rudolph, C. Keinki, A. Hoppe, R. Muecke, O. Micke, K. Muenstedt, and J. Huebner. Perception of cancer patients of their disease, self-efficacy and locus of control and usage of complementary and alternative medicine. *Journal of cancer research and clinical oncology*, 141(8):1449–1455, 2015.

D. Elsweiler and M. Kattenbeck. Understanding credibility judgements for web search snippets. *Aslib Journal of Information Management*, 2019.

R. Epstein and R. E. Robertson. The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521, 2015.

G. Eysenbach and P. E. Kummervold. ”is cybermedicine killing you?” -the story of a cochrane disaster. *Journal of Medical Internet Research*, 7(2), 2005.

L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd international conference on computational linguistics: Posters*, pages 276–284. Association for Computational Linguistics, 2010a.

L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284. Association for Computational Linguistics, 2010b.

L. Fernandez-Luque and T. Bau. Health and social media: perfect storm of information. *Healthcare informatics research*, 21(2):67–73, 2015. ISSN 2093-3681.

E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016.

S. Fox. *The social life of health information, 2011.* Pew Internet & American Life Project Washington, DC, 2011.

S. Fox and M. Duggan. Health online 2013. *Health*, 2013:1–55, 2013.

K. S. Freeman and J. H. Spyridakis. An examination of factors that affect the credibility of online health information. *Technical Communication*, 51(2):239–263, 2004.

L. C. Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1 (3):215–239, 1978.

J. Friedlin and C. J. McDonald. An evaluation of medical knowledge contained in wikipedia and its use in the loinc database. *Journal of the American Medical Informatics Association*, 17(3):283–287, 2010.

Y. Frish and D. Greenbaum. Is social media a cesspool of misinformation? clearing a path for patient-friendly safe spaces online. *The American Journal of Bioethics*, 17(3):19–21, 2017.

L. Y. Fu, K. Zook, Z. Spoehr-Labutta, P. Hu, and J. G. Joseph. Search engine ranking, quality, and content of web pages that are critical versus noncritical of human papillomavirus vaccine. *Journal of Adolescent Health*, 58(1):33–39, 2016.

N. Fuhr, A. Giachanou, G. Grefenstette, I. Gurevych, A. Hanselowski, K. Järvelin, R. Jones, Y. Liu, J. Mothe, W. Nejdl, et al. An information nutritional label for online documents. In *SIGIR Forum*, volume 51, pages 46–66, 2017.

M. Gahr, Z. Uzelac, R. Zeiss, B. J. Connemann, D. Lang, and C. Schönfeldt-Lecuona. Linking annual prescription volume of antidepressants to corresponding web search query data: a possible proxy for medical prescription behavior? *Journal of clinical psychopharmacology*, 35(6):681–685, 2015.

P. Galán-García, J. G. d. l. Puerta, C. L. Gómez, I. Santos, and P. G. Bringas. Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. *Logic Journal of the IGPL*, 24(1):42–53, 2016.

Q. Gao, Y. Tian, and M. Tu. Exploring factors influencing chinese users perceived credibility of health and safety information on weibo. *Computers in human behavior*, 45: 21–31, 2015.

D. Gayo-Avello, P. T. Metaxas, E. Mustafaraj, M. Strohmaier, H. Schoen, P. Gloor, C. Castillo, M. Mendoza, and B. Poblete. Predicting information credibility in time-sensitive social media. *Internet Research*, 2013.

A. Ghazikhani, R. Monsefi, and H. S. Yazdi. Online neural network model for non-stationary and imbalanced data stream classification. *International Journal of Machine Learning and Cybernetics*, 5(1):51–62, 2014.

J. Ghaznavi and L. D. Taylor. Bones, body parts, and sex appeal: An analysis of #thinspiration images on popular social media. *Body Image*, 14:54–61, 2015.

A. Ghenai and Y. Mejova. Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter. In *Proceedings - 2017 IEEE International Conference on Healthcare Informatics, ICHI 2017*, 2017. ISBN 9781509048816. doi: 10.1109/ICHI.2017.58.

R. Ghosh, T. Surachawala, and K. Lerman. Entropy-based classification of 'retweeting' activity on twitter. *Proceedings of KDD workshop on Social Network Analysis (SNA-KDD)*, 2011.

A. A. Ginart, S. Das, J. K. Harris, R. Wong, H. Yan, M. Krauss, and P. A. Cavazos-Rehg. Drugs or dancing? using real-time machine learning to classify streamed "dabbing" homograph tweets. In *Healthcare Informatics (ICHI), 2016 IEEE International Conference on*, pages 10–13. IEEE, 2016.

D. Girardi, J. Küng, R. Kleiser, M. Sonnberger, D. Csillag, J. Trenkler, and A. Holzinger. Interactive knowledge discovery with the doctor-in-the-loop: a practical example of cerebral aneurysms research. *Brain informatics*, 3(3):133, 2016.

L. M. Given. *The Sage encyclopedia of qualitative research methods*. Sage publications, 2008.

M. S. Goldstein. The persistence and resurgence of medical pluralism. *Journal of health politics, policy and law*, 29(4):925–945, 2004.

S. Green. 'antineoplastons': an unproved cancer therapy. *JAMA*, 267(21):2924–2928, 1992.

J. A. Greene, N. K. Choudhry, E. Kilabuk, and W. H. Shrank. Online social networking by patients with diabetes: a qualitative evaluation of communication with facebook. *Journal of General Internal Medicine*, 26(3):287–292, 2011.

Y. Gu. To code or not to code: Dilemmas in analysing think-aloud protocols in learning strategies research. *System*, 43:74 – 81, 2014a. ISSN 0346-251X. doi: https://doi.org/10.1016/j.system.2013.12.011. URL http://www.sciencedirect.com/science/article/pii/S0346251X13001875. Language Learning Strategy Research in the Twenty-First Century: Insights and Innovations.

Y. Gu. To code or not to code: Dilemmas in analysing think-aloud protocols in learning strategies research. *System*, 43:74–81, 2014b.

Z. Guan, S. Lee, E. Cuddihy, and J. Ramey. The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 1253–1262. ACM, 2006.

X. Gui, Y. Kou, K. H. Pine, and Y. Chen. Managing uncertainty: Using social media for risk assessment during a public health crisis. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 4520–4533. ACM, 2017.

A. Gupta and R. Kaushal. Improving spam detection in online social networks. In *2015 International conference on cognitive computing and information processing (CCIP)*, pages 1–6. IEEE, 2015.

A. Gupta and P. Kumaraguru. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st workshop on privacy and security in online social media*, page 2. Acm, 2012.

A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*, pages 729–736. ACM, 2013.

A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*, pages 228–243. Springer, 2014.

M. Gupta, P. Zhao, and J. Han. Evaluating event credibility on twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pages 153–164. SIAM, 2012.

A. Haas and J. Unkel. Ranking versus reputation: perception and effects of search result credibility. *Behaviour & Information Technology*, 36(12):1285–1298, 2017.

M. Han Veiga and C. Eickhoff. A cross-platform collection of social network profiles. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 665–668. ACM, 2016.

Q. He, E. Agu, D. Strong, B. Tulu, and P. Pedersen. Characterizing the performance and behaviors of runners using twitter. In *Healthcare Informatics (ICHI), 2013 IEEE International Conference on*, pages 406–414. IEEE, 2013.

J. M. Heilman, E. Kemmann, M. Bonert, A. Chatterjee, B. Ragar, G. M. Beards, D. J. Iberri, M. Harvey, B. Thomas, W. Stomp, et al. Wikipedia: a key tool for global public health promotion. *Journal of Medical Internet Research*, 13(1):e14, 2011.

E. S. Helen. Cancer, access to investigational drugs, and patient rights in the usa and india. *Indian Journal of Medical Ethics*, 5(4), 2008.

A. Hern. How social media filter bubbles and algorithms influence the election. *The Guardian*, 2017. URL https://www.theguardian.com/technology/2017/may/22/social-media-election-facebook-filter-bubbles.

J. P. Higgins, S. Green, et al. *Cochrane handbook for systematic reviews of interventions*. Wiley Online Library, 2008.

T. K. Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE, 1995.

I. Hochberg, D. Daoud, N. Shehadeh, and E. Yom-Tov. Can internet search engine queries be used to diagnose diabetes? analysis of archival search data. *Acta diabetologica*, pages 1–6, 2019.

A. Holzinger. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131, 2016.

J. Huebner, O. Micke, R. Muecke, J. Buentzel, F. J. Prott, U. Kleeberg, B. Senf, K. Muenstedt, et al. User rate of complementary and alternative medicine (cam) of patients visiting a counseling facility for cam of a german comprehensive cancer center. *Anticancer research*, 34(2):943–948, 2014.

M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg. Aidr: Artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 159–162. ACM, 2014.

X. Ji, S. A. Chun, and J. Geller. Monitoring public health concerns using twitter sentiment classifications. In *Healthcare Informatics (ICHI), 2013 IEEE International Conference on*, pages 335–344. IEEE, 2013.

F. Jin, W. Wang, L. Zhao, E. Dougherty, Y. Cao, C. T. Lu, and N. Ramakrishnan. Misinformation propagation in the age of Twitter. *Computer*, 47(12):90–94, 2014. ISSN 00189162. doi: 10.1109/MC.2014.361.

L. M. Jonathan. Spiders, lizards and frogs can help prevent dengue, zika outbreaks. http://www.businessmirror.com.ph/spiders-lizards-and-frogs-can-help-prevent-dengue-zika-outbreaks/, 2016. Accessed: 2017-01-02.

E. H. Jung, K. Walsh-Childers, and H.-S. Kim. Factors influencing the perceived credibility of diet-nutrition information web sites. *Computers in Human Behavior*, 58:37–47, 2016.

J. Kaicker, W. Dang, and T. Mondal. Assessing the quality and reliability of health information on ercp using the discern instrument. *Health Care: Current Reviews*, pages 1–4, 2013.

K. Kamenova and T. Caulfield. Stem cell hype: media portrayal of therapy translation. *Science Translational Medicine*, 7(278):278ps4–278ps4, 2015.

Y. Kammerer and P. Gerjets. How the interface design influences users' spontaneous trustworthiness evaluations of web search results: comparing a list and a grid interface. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 299–306. ACM, 2010.

Y. Kammerer, I. Bråten, P. Gerjets, and H. I. Strømsø. The role of internet-specific epistemic beliefs in laypersons source evaluations and decisions during web search on a medical issue. *Computers in Human Behavior*, 29(3):1193–1203, 2013.

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*, pages 179–188. ACM, 2012.

V. O. Kayhan. Seeking health information on the web: positive hypothesis testing. *International journal of medical informatics*, 82(4):268–275, 2013.

G. Kazai, P. Thomas, and N. Craswell. The Emotion Profile of Web Search. pages 1097–1100, 2019. ISBN 9781450361729.

D. Kelly et al. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval*, 3(1–2):1–224, 2009.

I. Khaldarova and M. Pantti. Fake news: The narrative battle over the ukrainian conflict. *Journalism Practice*, 10(7):891–901, 2016.

A. Kinsora, K. Barron, Q. Mei, and V. V. Vydiswaran. Creating a labeled dataset for medical misinformation in health forums. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 456–461. IEEE, 2017.

H. Korda and Z. Itani. Harnessing social media for health promotion and behavior change. *Health Promotion Practice*, 14(1):15–23, 2013.

P. Kostkova, V. Mano, H. J. Larson, and W. S. Schulz. Vac medi+ board: Analysing vaccine rumours in news and social media. In *Proceedings of the 6th International Conference on Digital Health Conference*, pages 163–164. ACM, 2016.

R. V. Kozinets, K. De Valck, A. C. Wojnicki, and S. J. Wilner. Networked narratives: Understanding word-of-mouth marketing in online communities. *Journal of Marketing*, 74(2):71–89, 2010.

J. Kulshrestha, M. Eslami, J. Messias, M. B. Zafar, S. Ghosh, I. Shibpur, I. K. P. Gummadi, and K. Karahalios. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proc. of CSCW*, 2017.

S. Kumar, R. West, and J. Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602. International World Wide Web Conferences Steering Committee, 2016.

A. E. Kuriyan, T. A. Albini, J. H. Townsend, M. Rodriguez, H. K. Pandya, R. E. Leonard, M. B. Parrott, P. J. Rosenfeld, H. W. Flynn Jr, and J. L. Goldberg. Vision loss after intravitreal injection of autologous stem cells for amd. *N Engl J Med*, 2017(376):1047–1053, 2017.

H. Kuusela and P. Pallab. A comparison of concurrent and retrospective verbal protocol analysis. *The American journal of psychology*, 113(3):387, 2000.

L. Laranjo, A. Arguel, A. L. Neves, A. M. Gallagher, R. Kaplan, N. Mortimer, G. A. Mendes, and A. Y. Lau. The influence of social networking sites on health behavior change: a systematic review and meta-analysis. *Journal of the American Medical Informatics Association*, 22(1):243–256, 2014.

L. Laranjo, A. Arguel, A. L. Neves, A. M. Gallagher, R. Kaplan, N. Mortimer, G. A. Mendes, and A. Y. S. Lau. The influence of social networking sites on health behavior change: a systematic review and meta-analysis. *Journal of the American Medical Informatics Association*, 22(1):243–256, 2015.

H. J. Larson, R. Wilson, S. Hanley, A. Parys, and P. Paterson. Tracking the global spread of vaccine sentiments: The global response to Japan's suspension of its HPV vaccine recommendation. *Human Vaccines and Immunotherapeutics*, 10(9):2543–2550, 2014. ISSN 2164554X. doi: 10.4161/21645515.2014.969618.

B. Latané. The psychology of social impact. *American psychologist*, 36(4):343, 1981.

A. Lau and E. Coiera. Impact of web searching and social feedback on consumer decision making: a prospective online experiment. *Journal of medical Internet research*, 10(1): e2, 2008.

A. Lau, T. Kwok, and E. Coiera. How online crowds influence the way individual consumers answer health questions. *Applied clinical informatics*, 2(02):177–189, 2011.

A. Y. Lau and E. W. Coiera. Do people experience cognitive biases while searching for information? *Journal of the American Medical Informatics Association*, 14(5):599–608, 2007.

A. Y. Lau and E. W. Coiera. Can cognitive biases during consumer health information searches be reduced to improve decision making? *Journal of the American Medical Informatics Association*, 16(1):54–65, 2009.

A. Y. Lau, T. M. Kwok, E. W. Coiera, et al. The influence of crowds on consumer health decisions: an online prospective study. In *MedInfo*, pages 33–37, 2010.

A. Y. Lau, E. Gabarron, L. Fernandez-Luque, and M. Armayones. Social media in health - what are the safety concerns for health consumers? *Health Information Management Journal*, 41(2):30–35, 2012.

M. R. Laurent and T. J. Vickers. Seeking health information online: does wikipedia matter? *Journal of the American Medical Informatics Association*, 16(4):471–479, 2009.

A. Leavitt and J. J. Robinson. The role of information visibility in network gatekeeping: Information aggregation on reddit during crisis events. In *CSCW*, pages 1246–1261, 2017.

R. Lederman, H. Fan, S. Smith, and S. Chang. Who can you trust? credibility assessment in online health forums. *Health Policy and Technology*, 3(1):13–25, 2014.

K. Lee, K. Hoti, J. D. Hughes, and L. Emmerton. Dr google and the consumer: a qualitative study exploring the navigational needs and online health information-seeking behaviors of consumers with chronic health conditions. *Journal of medical Internet research*, 16 (12):e262, 2014.

J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM, 2009.

W.-Y. Lin, X. Zhang, H. Song, and K. Omori. Health information seeking in the web 2.0 age: Trust in social media, uncertainty reduction, and self-disclosure. *Computers in Human Behavior*, 56:289–294, 2016.

X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1867–1870. ACM, 2015.

C. L. Loprinzi, R. Levitt, D. L. Barton, J. A. Sloan, P. J. Atherton, D. J. Smith, S. R. Dakhil, D. F. Moore, J. E. Krook, K. M. Rowland, et al. Evaluation of shark cartilage in patients with advanced cancer. *Cancer*, 104(1):176–182, 2005.

X. Lou, A. Flammini, and F. Menczer. Information pollution by social bots. *arXiv preprint arXiv:1907.06130*, 2019.

W. Lowe, K. Benoit, S. Mikhaylov, and M. Laver. Scaling policy preferences from coded political texts. *Legislative studies quarterly*, 36(1):123–155, 2011.

Q. I. P. Ltd. *NVivo qualitative data analysis software*. Version 12, 2018.

R. Ludolph, A. Allam, and P. J. Schulz. Manipulating googles knowledge graph box to counter biased information processing during an online search on vaccination: application of a technological debiasing strategy. *Journal of medical Internet research*, 18(6): e137, 2016.

J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1751–1754. ACM, 2015.

J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 3818–3824, 2016.

T. J. Ma and D. Atkin. User generated content and credibility evaluation of online health information: a meta analytic study. *Telematics and Informatics*, 34(5):472–486, 2017.

G. Magno and I. Weber. International gender differences and gaps in online social networks. In *International Conference on Social Informatics*, pages 121–138. Springer, 2014.

J. A. Maxwell. *Qualitative research design: An interactive approach*, volume 41. Sage publications, 2012.

J. McCorriston, D. Jurgens, and D. Ruths. Organizations are users too: Characterizing and detecting the presence of organizations on twitter. In *ICWSM*, pages 650–653, 2015.

F. McGregor, J. E. Somner, R. R. Bourne, C. Munn-Giddings, P. Shah, and V. Cross. Social media use by patients with glaucoma: what can we learn? *Ophthalmic and Physiological Optics*, 34(1):46–52, 2014.

M. L. McHugh. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282, 2012.

Y. Mejova, H. Haddadi, S. Abbar, A. Ghahghaei, and I. Weber. Dietary habits of an expat nation: Case of qatar. In *Healthcare Informatics (ICHI), 2015 International Conference on*, pages 57–62. IEEE, 2015a.

Y. Mejova, H. Haddadi, A. Noulas, and I. Weber. #foodporn: Obesity patterns in culinary interactions. In *DH'15: International Conference on Digital Health 2015*, pages 51–58. ACM, 2015b.

Y. Mejova, S. Abbar, and H. Haddadi. Fetishizing food in digital age:# foodporn around the world. In *ICWSM*, pages 250–258, 2016.

Y. Mejova, Y. Benkhedda, and Khairani. #halal culture on instagram. *Frontiers in Digital Humanities*, 4:21, 2017.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

T. Mitra, S. Counts, and J. W. Pennebaker. Understanding anti-vaccination attitudes in social media. In *ICWSM*, pages 269–278, 2016.

S. A. Moorhead, D. E. Hazlett, L. Harrison, J. K. Carroll, A. Irwin, and C. Hoving. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *Journal of medical Internet research*, 15(4), 2013.

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*, pages 441–450. ACM, 2012.

S. Mukherjee, G. Weikum, and C. Danescu-Niculescu-Mizil. People on drugs: credibility of user statements in health communities. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 65–74. ACM, 2014.

A. Nastasi, T. Bryant, J. K. Canner, M. Dredze, M. S. Camp, and N. Nagarajan. Breast cancer screening and social media: a content analysis of evidence use and guideline opinions on twitter. *Journal of Cancer Education*, pages 1–8, 2017.

M. W. Newman, D. Lauterbach, S. A. Munson, P. Resnick, and M. E. Morris. It's not that i don't have problems, i'm just not putting them on facebook: challenges and opportunities in using online social networks for health. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 341–350. ACM, 2011.

A. Nourbakhsh, X. Liu, S. Shah, R. Fang, M. M. Ghassemi, and Q. Li. Newsworthy rumor events: A case study of twitter. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 27–32. IEEE, 2015.

B. Nyhan and J. Reifler. Does correcting myths about the flu vaccine work? an experimental evaluation of the effects of corrective information. *Vaccine*, 33(3):459–464, 2015.

B. Nyhan, J. Reifler, S. Richey, and G. L. Freed. Effective messages in vaccine promotion: a randomized trial. *Pediatrics*, 133(4):e835–e842, 2014.

D. J. Odgers, R. Harpaz, A. Callahan, G. Stiglic, and N. H. Shah. Analyzing search behavior of healthcare professionals for drug safety surveillance. In *Pacific Symposium on Biocomputing Co-Chairs*, pages 306–317. World Scientific, 2014.

D. Ohashi, R. Cohen, and X. Fu. The current state of online social networking for the health community: Where trust modeling research may be of value. In *Proceedings of the 2017 International Conference on Digital Health*, pages 23–32. ACM, 2017.

A. Olteanu, S. Vieweg, and C. Castillo. What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 994–1009. ACM, 2015.

A. Olteanu, O. Varol, and E. K\ic\iman. Distilling the Outcomes of Personal Experiences: A Propensity-scored Analysis of Social Media. In *Proc. of The 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 2017.

W. H. Organization et al. Risk communication in the context of zika virus: interim guidance. Technical report, World Health Organization, March 2016.

Y. Ouyang. Student's death highlights gaps in china's health regulations. *Lancet Oncology*, 17(6):709, 2016.

S. O. Oyeyemi, E. Gabarron, and R. Wynn. Ebola, Twitter, and misinformation: a dangerous combination? *British Medical Journal*, 349(October):g6178, 2014a. ISSN 1756-1833. doi: 10.1136/bmj.g6178. URL http://www.bmj.com/cgi/doi/10.1136/bmj.g6178.

S. O. Oyeyemi, E. Gabarron, and R. Wynn. Ebola, twitter, and misinformation: a dangerous combination? *Bmj*, 349:g6178, 2014b.

P. Ozturk, H. Li, and Y. Sakamoto. Combating rumor spread on social media: The effectiveness of refutation and warning. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, pages 2406–2414. IEEE, 2015.

B. Pan, H. Hembrooke, T. Joachims, L. Lorigo, G. Gay, and L. Granka. In google we trust: Users decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication*, 12(3):801–823, 2007.

M. J. Paul and M. Dredze. Discovering health topics in social media using topic models. *PloS one*, 9(8):e103408, 2014.

S. A. Paul, L. Hong, and E. H. Chi. Is twitter a good place for asking questions? a characterization study. In *ICWSM*, 2011.

Pew Research Center. Social Media Use in 2018, 2018. URL http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/.

F. A. Pogacar, A. Ghenai, M. D. Smucker, and C. L. Clarke. The positive and negative influence of search results on people's decisions about the efficacy of medical treatments. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, pages 209–216. ACM, 2017.

K. Purcell, L. Rainie, A. Mitchell, T. Rosenstiel, and K. Olmstead. Understanding the participatory news consumer. 2010.

K. Purcell, J. Brenner, and L. Rainie. Search engine use 2012. 2012.

V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599. Association for Computational Linguistics, 2011.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/.

J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*, pages 249–252. ACM, 2011.

R. E. Rice. Influences, usage, and outcomes of internet health information searching: multivariate results from the pew surveys. *International journal of medical informatics*, 75(1):8–28, 2006.

B. Rink, S. Harabagiu, and K. Roberts. Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, 18 (5):594–600, 2011.

J. Roozenbeek and S. van der Linden. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, 5(1):12, 2019.

R. L. Rosnow. Inside rumor: A personal journey. *American Psychologist*, 46(5):484, 1991.

R. L. Rosnow and E. K. Foster. Rumor and gossip research. *Psychological Science Agenda*, 19(4):1–2, 2005.

V. L. Rubin, Y. Chen, and N. J. Conroy. Deception detection for news: three types of fakes. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, page 83. American Society for Information Science, 2015.

I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review*, 18(2):95–145, 2003.

A. Sadilek, H. A. Kautz, and V. Silenzio. Modeling spread of disease from social interactions. In *ICWSM*, pages 322–329, 2012.

M. Salathé and S. Khandelwal. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. *PLoS computational biology*, 7 (10):e1002199, 2011.

N. J. Salkind. *Encyclopedia of Research Design*, volume 1. SAGE, 2010.

H. Samuel and O. Zaıane. Medfact: Towards improving veracity of medical information in social media using applied machine learning. *Canadian Conference on Artificial Intelligence*, 2018.

D. Scanfeld, V. Scanfeld, and E. L. Larson. Dissemination of health information through social networks: Twitter and antibiotics. *American journal of infection control*, 38(3): 182–188, 2010.

K. Schmidt and E. Ernst. Assessing websites on complementary and alternative medicine for cancer. *Annals of Oncology*, 15(5):733–742, 2004.

I. Seaman and C. Giraud-Carrier. Prevalence and attitudes about illicit and prescription drugs on twitter. In *Healthcare Informatics (ICHI), 2016 IEEE International Conference on*, pages 14–17. IEEE, 2016.

E. Seo, P. Mohapatra, and T. Abdelzaher. Identifying rumors and their sources in social networks. In *Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR III*, volume 8389, page 83891I. International Society for Optics and Photonics, 2012.

C. Shao, G. L. Ciampaglia, A. Flammini, and F. Menczer. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 745–750. International World Wide Web Conferences Steering Committee, 2016.

C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini, and F. Menczer. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*, pages 96–104, 2017.

C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer. The spread of low-credibility content by social bots. *Nature communications*, 9(1):4787, 2018.

M. Sharma, K. Yadav, N. Yadav, and K. C. Ferdinand. Zika virus pandemicanalysis of facebook as a social media health information platform. *American journal of infection control*, 45(3):301–302, 2017.

T. Sharot, A. M. Riccardi, C. M. Raio, and E. A. Phelps. Neural mechanisms mediating optimism bias. *Nature*, 450(7166):102, 2007.

K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

S. Sikdar, B. Kang, J. ODonovan, T. Höllerer, and S. Adah. Understanding information credibility on twitter. In *2013 International Conference on Social Computing*, pages 19–24. IEEE, 2013.

E. Sillence, P. Briggs, L. Fishwick, and P. Harris. Trust and mistrust of online health sites. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 663–670. ACM, 2004.

E. Sillence, P. Briggs, P. Harris, and L. Fishwick. A framework for understanding trust factors in web-based health advice. *International Journal of Human-Computer Studies*, 64(8):697–713, 2006.

E. Sillence, P. Briggs, P. R. Harris, and L. Fishwick. How do patients evaluate and make use of online health information? *Social science & medicine*, 64(9):1853–1862, 2007.

L. Soldaini and E. Yom-Tov. Inferring individual attributes from search engine queries and auxiliary information. In *Proceedings of the 26th international conference on World Wide Web*, pages 293–301. International World Wide Web Conferences Steering Committee, 2017.

K. Starbird. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *ICWSM*, pages 230–239, 2017.

M. C. Sturkenboom. The narcolepsy-pandemic influenza story: Can the truth ever be unraveled? *Vaccine*, 33:B6–B13, 2015.

P. Suárez-Serrato, M. E. Roberts, C. Davis, and F. Menczer. On the influence of social bots in online protests. In *International Conference on Social Informatics*, pages 269–278. Springer, 2016.

S. Syed-Abdul, L. Fernandez-Luque, W.-S. Jian, Y.-C. Li, S. Crain, M.-H. Hsu, Y.-C. Wang, D. Khandregzen, E. Chuluunbaatar, P. A. Nguyen, et al. Misleading health-related information promoted through video-based social media: anorexia on youtube. *Journal of medical Internet research*, 15(2):e30, 2013.

T. T. Tang, N. Craswell, D. Hawking, K. Griffiths, and H. Christensen. Quality and relevance of domain-specific search: A case study in mental health. *Information Retrieval*, 9(2):207–225, 2006.

T. R. Tangherlini, V. Roychowdhury, B. Glenn, C. M. Crespi, R. Bandari, A. Wadia, M. Falahi, E. Ebrahimzadeh, and R. Bastani. mommy blogs and the vaccination exemption narrative: results from a machine-learning approach for story aggregation on parenting social media sites. *JMIR public health and surveillance*, 2(2):e166, 2016.

Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1): 24–54, 2010.

The Vaccine Confidence Project. Sierra Leone: Ebola fears hamper vaccination campaigns against Polio and Measles. http://www.vaccineconfidence.org/sierra-leone-ebola-fears-hamper-vaccination-campaigns-against-polio-and-measles/, jun 2015. Accessed: 2016-10-16.

Y.-L. Theng, L. Y. Q. Goh, M. O. Lwin, and S. F. Shou-Boon. Dispelling myths and misinformation using social media: A three-countries comparison using the case of tuberculosis. In *Healthcare Informatics (ICHI), 2013 IEEE International Conference on*, pages 147–152. IEEE, 2013.

T. L. Thompson. *Encyclopedia of health communication*. Sage Publications, 2014.

S. Towers, S. Afzal, G. Bernal, N. Bliss, S. Brown, B. Espinoza, J. Jackson, J. Judson-Garcia, M. Khan, M. Lin, et al. Mass media and the contagion of fear: the case of ebola in america. *PloS one*, 10(6):e0129179, 2015.

B. Urciouli. Navigating the world of fake news and phony cancer cures. *Cure Today*, 2016.

U.S. Census Bureau. International data base country rankings. http://www.census.gov/population/ international/data/idb/rank.php, September 2016. Accessed: 2016-09-30.

US Food and Drug Administration. The Safety Reporting Portal, 2018. URL https://www.safetyreporting.hhs.gov/.

M. Van Someren, Y. Barnard, and J. Sandberg. *The think aloud method: a practical approach to modelling cognitive*. Citeseer, 1994.

O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *Eleventh international AAAI conference on web and social media*, 2017.

W. N. Venables and B. D. Ripley. *Modern applied statistics with S-PLUS*. Springer Science & Business Media, 2013.

A. Venkatraman, D. Mukhija, N. Kumar, and S. J. S. Nagpal. Zika virus misinformation on the internet. *Travel medicine and infectious disease*, 14(4):421, 2016.

W. H. Vogel. Internet oncology: Cure seekers beware! *Journal of the Advanced Practitioner in Oncology*, 2(6):409–412, 2011.

E. K. Vraga and L. Bode. Using expert sources to correct health misinformation in social media. *Science Communication*, 39(5):621–645, 2017.

E. K. Vraga and L. Bode. I do not believe you: how providing a source corrects health misperceptions across social media platforms. *Information, Communication & Society*, 21(10):1337–1353, 2018.

H. S. Wald, C. E. Dube, and D. C. Anthony. Untangling the web - the impact of internet use on health care and the physician–patient relationship. *Patient education and counseling*, 68(3):218–224, 2007.

D. Warner and J. D. Procaccino. Toward wellness: Women seeking health information. *Journal of the American Society for Information Science and Technology*, 55(8):709–730, 2004.

P. M. Waszak, W. Kasprzycka-Waszak, and A. Kubanek. The spread of medical fake news in social media–the pilot quantitative study. *Health Policy and Technology*, 7(2):115–118, 2018.

H. Webb, P. Burnap, R. Procter, O. Rana, B. C. Stahl, M. Williams, W. Housley, A. Edwards, and M. Jirotka. Digital Wildfires: Propagation, Verification, Regulation, and Responsible Innovation. *ACM Transactions on Information Systems*, 34(3):1–23, 2016. ISSN 10468188. doi: 10.1145/2893478. URL http://dl.acm.org/citation.cfm?id=2915200.2893478{%}5Cnhttp://dl.acm.org/citation.cfm?doid=2915200.2893478.

L. Weitzel, J. P. M. de Oliveira, and P. Quaresma. Measuring the reputation in user-generated-content systems based on health information. *Procedia Computer Science*, 29: 364–378, 2014.

P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen. Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic acids research*, 39(suppl_2):W541–W545, 2011.

R. White. Beliefs and biases in web search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. ACM, 2013.

R. W. White. Belief dynamics in web search. *Journal of the Association for Information Science and Technology*, 65(11):2165–2178, 2014.

R. W. White and A. Hassan. Content bias in online health search. *ACM Transactions on the Web (TWEB)*, 8(4):25, 2014.

R. W. White and E. Horvitz. Experiences with web search on medical concerns and self diagnosis. In *AMIA*, 2009.

R. W. White and E. Horvitz. Studies of the onset and persistence of medical concerns in search logs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 265–274. ACM, 2012.

R. W. White and E. Horvitz. Belief dynamics and biases in web search. *ACM Transactions on Information Systems (TOIS)*, 33(4):18, 2015.

R. W. White, S. Dumais, and J. Teevan. How medical expertise influences web search interaction. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 791–792. ACM, 2008.

R. W. White, S. Wang, A. Pant, R. Harpaz, P. Shukla, W. Sun, W. DuMouchel, and E. Horvitz. Early identification of adverse drug reactions from search log data. *Journal of biomedical informatics*, 59:42–48, 2016.

WHO. Risk communication in the context of zika virus. http://www.who.int/risk-communication/zika-virus/risk-communication-presentation.pdf?ua=1, 2016. Accessed: 2017-01-03.

A. Wongkoblap, M. A. Vadillo, and V. Curcin. Classifying depressed users with multiple instance learning from social network data. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 436–436. IEEE, 2018.

World Health Organization. Mixed uptake of social media among public health specialists. *Bull World Health Organ*, 89:784–785, 2011. doi: 10.2471/BLT.11.021111. URL http://www.who.int/bulletin/volumes/89/11/11-031111.pdf?ua=1.

World Health Organization. The history of zika virus. http://www.who.int/emergencies/zika-virus/history/en/, 2016. Accessed: 2016-09-25.

World Health Organization. Dispelling rumours around zika and complications. http://www.who.int/emergencies/zika-virus/articles/ rumours/en/, September 2016. Accessed: 2016-07-20.

K. Wu, S. Yang, and K. Q. Zhu. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st International Conference on Data Engineering*, pages 651–662. IEEE, 2015.

S. Wu and L. A. Adamic. Visually impaired users on an online social network. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 3133–3142. ACM, 2014.

F. Yang, Y. Liu, X. Yu, and M. Yang. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, page 13. ACM, 2012.

H. Yang and C. C. Yang. Harnessing social media for drug-drug interactions detection. In *Healthcare Informatics (ICHI), 2013 IEEE International Conference on*, pages 22–29. IEEE, 2013.

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*, pages 575–586. ACM, 2013.

Q. Yang. Are social networking sites making health behavior change interventions more effective? a meta-analytic review. *Journal of health communication*, 22(3):223–233, 2017.

L. Yardley, L. Morrison, K. Bradbury, and I. Muller. The person-based approach to intervention development: application to digital health-related behavior change interventions. *Journal of medical Internet research*, 17(1), 2015.

E. Yom-Tov. Ebola data from the internet: An opportunity for syndromic surveillance or a news event? In *Proceedings of the 5th international conference on digital health 2015*, pages 115–119. ACM, 2015.

E. Yom-Tov. *Crowdsourced Health: How What You Do on the Internet Will Improve Medicine*. Mit Press, 2016.

E. Yom-Tov. Predicting drug recalls from internet search engine queries. *IEEE journal of translational engineering in health and medicine*, 5:1–6, 2017.

E. Yom-Tov and D. M. Boyd. On the link between media coverage of anorexia and pro-anorexic practices on the web. *International Journal of Eating Disorders*, 47(2):196–202, 2014.

E. Yom-Tov and E. Gabrilovich. Postmarket drug surveillance without trial costs: discovery of adverse drug reactions through large-scale analysis of web search queries. *Journal of medical Internet research*, 15(6):e124, 2013.

E. Yom-Tov and S. Lev-Ran. Adverse reactions associated with cannabis consumption as evident from search engine queries. *JMIR public health and surveillance*, 3(4):e77, 2017.

E. Yom-Tov, L. Fernandez-Luque, I. Weber, and S. P. Crain. Pro-anorexia and pro-recovery photo sharing: a tale of two warring tribes. *Journal of medical Internet research*, 14(6), 2012.

E. Yom-Tov, D. Borsa, I. J. Cox, and R. A. McKendry. Detecting disease outbreaks in mass gatherings using internet data. *Journal of medical Internet research*, 16(6):e154, 2014a.

E. Yom-Tov, R. W. White, and E. Horvitz. Seeking insights about cycling mood disorders via anonymized search logs. *Journal of medical Internet research*, 16(2):e65, 2014b.

E. Yom-Tov, D. Borsa, A. C. Hayward, R. A. McKendry, and I. J. Cox. Automatic identification of web-based risk markers for health events. *Journal of medical Internet research*, 17(1), 2015a.

E. Yom-Tov, I. J. Cox, and V. Lampos. Learning about health and medicine from internet data. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 417–418. ACM, 2015b.

B. Yu, M. Willis, P. Sun, and J. Wang. Crowdsourcing participatory evaluation of medical pictograms using amazon mechanical turk. *Journal of Medical Internet Research*, 15(6): e108, 2013.

R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. Defending against neural fake news. *arXiv preprint arXiv:1905.12616*, 2019.

H. Zhai, T. Lingren, L. Deleger, Q. Li, M. Kaiser, L. Stoutenborough, and I. Solti. Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. *Journal of Medical Internet Research*, 15(4):e73, 2013.

H. Zhang and J. Su. Naïve bayesian classifiers for ranking. In *Machine Learning: ECML 2004*, pages 501–512. Springer, 2004.

L. Zhao, T. Hua, C.-T. Lu, and R. Chen. A topic-focused trust model for twitter. *Computer Communications*, 76:1–11, 2016a.

M. Zhao, C. C. Yang, J. Thrul, and D. Ramo. Patterns of interaction within smoking cessation groups on facebook: A social network analysis. In *Healthcare Informatics (ICHI), 2016 IEEE International Conference on*, pages 305–305. IEEE, 2016b.

Z. Zhao, P. Resnick, and Q. Mei. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1395–1405, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3469-3.

APPENDICES

Appendix A

English Spanish Portuguese Map

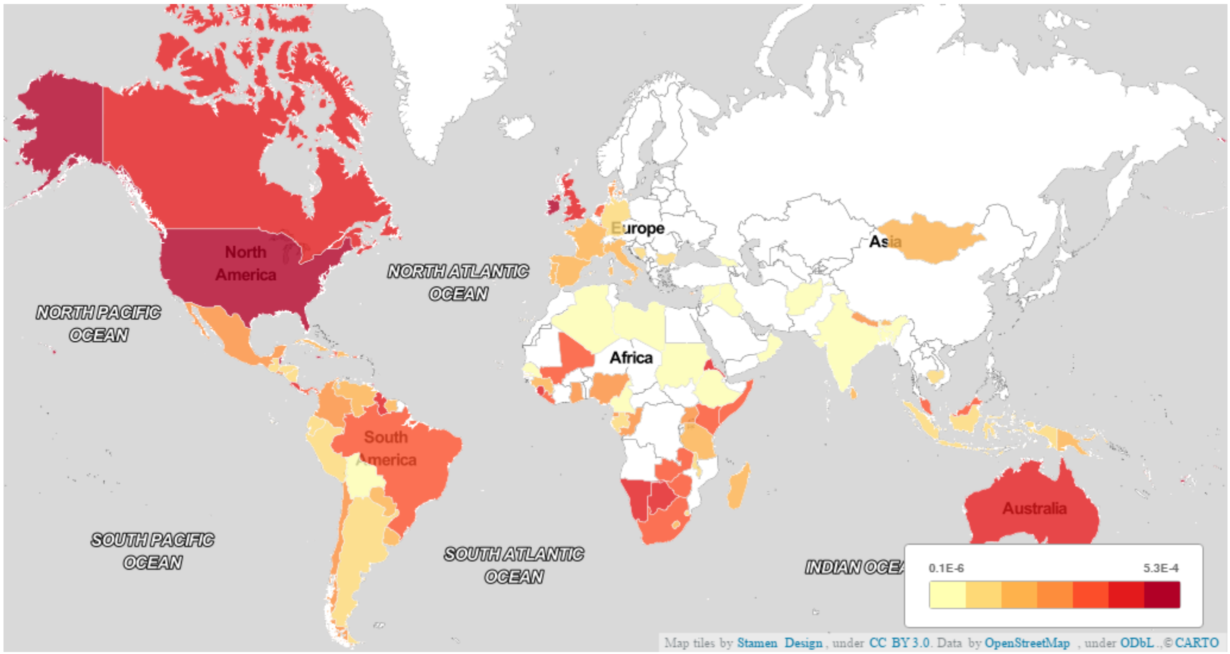


Figure A.1: Geographic distribution of English Zika-related tweets, normalized by the number of Internet users.

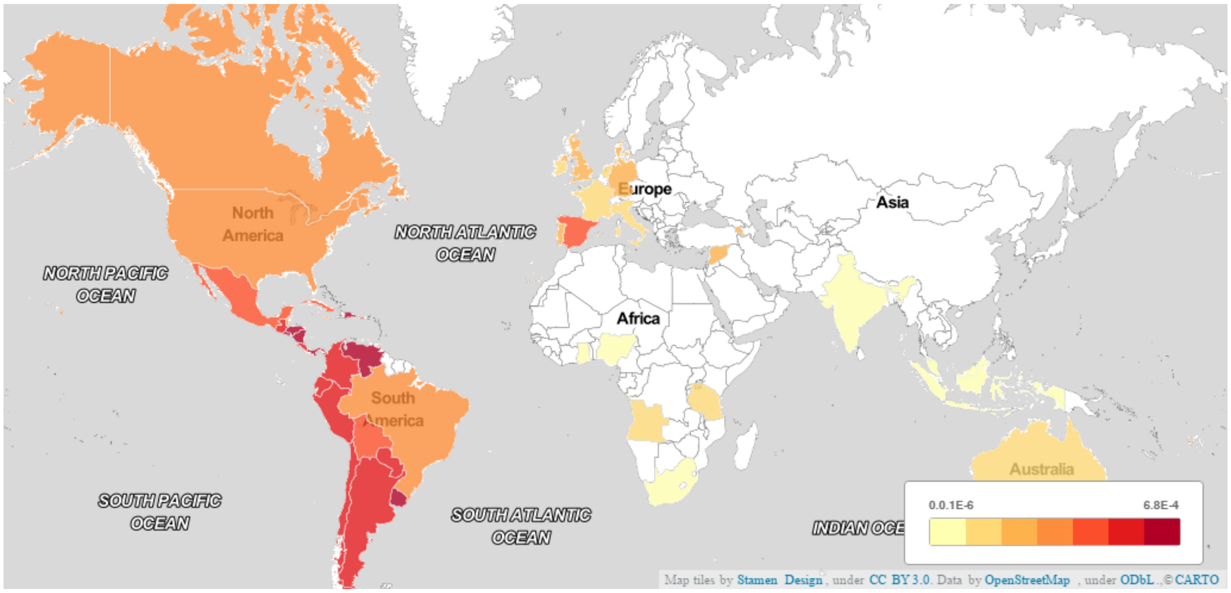


Figure A.2: Geographic distribution of Spanish Zika-related tweets, normalized by the number of Internet users.

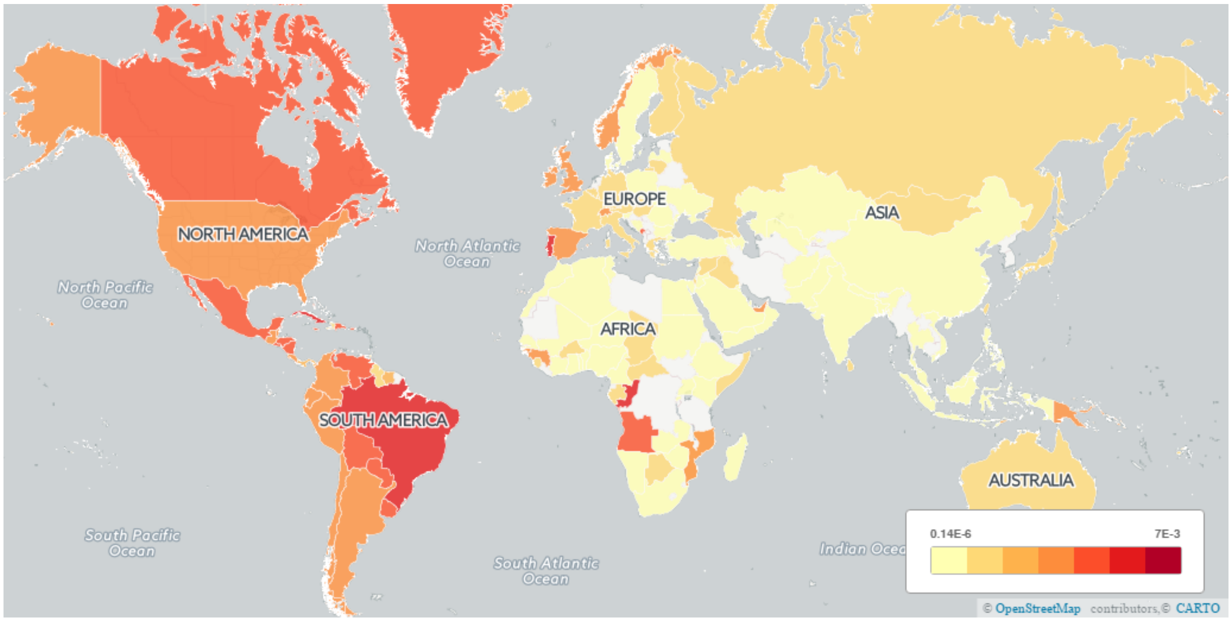


Figure A.3: Geographic distribution of Portuguese Zika-related tweets, normalized by the number of Internet users.

Appendix B

Retrospective Think-aloud Summary

B.1 Participant 3

The participant believes that websites which do not contain actual studies or are not research articles are not legitimate sources. Websites with real studies and statistical testing (such as WebMD), news sources, sources from known health organizations (such as the American Diabetes Association) or websites citing authoritative sources are the most trustworthy sources of medical information. Additionally, the participant does not rely on websites containing advertisement content and usually opens the first ranked links (described by the participant as the"top hits") even before reading the title and snippet out of habit. The participant compares information from different sources and looks for consensus to make the final decision and, generally, prefers reading pages stating that the medical treatment does not help in order to understand the opposite arguments. The participant adds that the date and time is important as recent information has more value than old ones. Further, he does not trust websites that are not professionally designed (colors, boxes, etc.) and where authors do not have authoritative titles such as doctors. When the participant has a prior belief about the effectiveness of a treatment, he is more likely to open and look at content agreeing with the prior belief.

B.2 Participant 4

The participant judges the reliability of the websites by looking at the content length and the presence of references. If the content is short or if there are no citations, then the source is not reliable. Additionally, the participant finds personal blogs and news websites not authoritative when searching for medical information. The participant fees suspicious about the web page content whenever it claims that the information comes from research (such as when the content states:"research shows") and starts looking for actual citations to check the content for confirmation. Further, the participant finds reliable sources first, then quickly checks relevant less reliable websites (such as answers.com and Yahoo answers) in order to look for consistency. If the topic keywords (treatment and health issue) are present in the title, then the participant decides to click on the link (relevance is important when deciding which links to click on). Next, the participant believes that the information presentation is important in getting a quick idea about the content (such as bullet points presentation, color scheme, clip arts, text size and no ads, etc.) When exploring the SERP page, the participant starts by reading the title, then looks whether the topic is relevant in order to decide whether to open it or not. Finally, when the participant finds inconsistency between two reliable sources, he opens almost all the links in the SERP page to get an idea about what the majority states.

B.3 Participant 5

When looking at search results, the participant chooses the pages to click on by looking at their authoritativeness. Study and research-oriented webpages (like PubMed) are chosen over advertisement-oriented content and blogs. Further, the participant trusts websites found reliable in the past to answer questions in the present. Next, the participant believes that the first three or four results in the search results page are the most popular and, as a result, they will most probably have the correct information. The participant opens all search result pages in order to find out the most relevant ones to the topic, then goes back and opens the relevant ones to answer the question. Additionally, the participant believes that the number of citations of a webpage content gives an idea about the level of reliability and trustworthiness in the content. Having a general agreement/consensus between different websites about a medical treatment judgment makes it more reliable. The participant does not trust news websites due to their political biases. If search result pages contain a large amount of advertisement, then participant does the search again with a new modified query (keywords).

B.4 Participant 6

The participant does not rely on web forums and blogs when searching for health-related topics as they lack authoritativeness and regulation. On the contrary, government and educational websites are trustworthy as they contain more accurate, professional, non-profitable and research-oriented information. Additionally, the participant does not trust webpages containing a large amount of advertisement content. An educational website listing the background, objective or strategies i.e with a research layout is more trustworthy. The participant believes that short concise and clear content is more useful than a long detailed reports with many pages and, also, finds that consistency among websites is important to give an idea about the correct answer. The page design (title appearance) and the easiness to read the content are helpful when exploring webpages content while doing the online search. Participant 6 explores the SERP pages as follows: first, the participant reads the titles to check if they are relevant to the topic, then, he/she looks into the URL in order to determine whether the webpage is news, government, blog etc. Finally, from the more trustworthy websites, the participant starts exploring the most relevant search results first. Further, the participant finds that listing references and recognized organizations (such as the UN), usually found at the end of the page, makes the content more reliable and trustworthy. Having special characters or capital letters in the URL title (like repeated exclamation makes) is a method to attract people and the URL might not be reliable (clickbait). The date the content is posted is important to participant 6 because recent information is more reliable and trustworthy. Even if the information comes from a non-trustworthy website, participant 6 keeps it in mind and looks for consensus (cross-referencing) with other websites, then decides on majority. The participant finds it suspicious when the title claims an extremely positive result such as "..relieve your back-pain forever".

B.5 Participant 7

Participant 7 read first the title and snippets and then looks at the URLs to check whether they are reputable. Search results that promote products are biased so participant 7 does not prefer to click on them. Furthermore, participant 7 reads content to understand first the causes of the health issue, before reading content about the effectiveness of the medical treatment. The structure of information (summary at the top, sections to different points) and the representation (grammar and spelling mistakes) and readability effects how helpful the information is. Moreover, having known organizations (such as WHO) listed in the websites showers high authoritativeness. Majority and search results consensus is important in exploring search result pages. This is even more important when dealing with a serious disease such as cancer. Sometimes, the participant looks for opposite arguments (why the treatment is not helpful) and then compares arguments with ones suggesting the treatment is helpful.

B.6 Participant 8

The participant uses the first clicked on link as a base reference for all future websites he decides to look at and finds that lengthier snippets suggest that there is a good amount of information about the topic in the search result page. Information representation (special characters in the title and snippet, having images), references, statistics and studies are important in determining the reliability of the website. Websites having advertisement content (posting prices, asking credit card information...) are disturbing and this shows a low credibility as per the participant. Further, The participant does not trust news websites and personal blogs because they do not contain information that is medically accurate. Different from the other participations, participant 8 finds that capital letters in titles catch attention in order to open the website and check the content. The date the content was published in the search result page and the author credentials show how credible the content is.

B.7 Participant 9

The participant describes deciding which search results to open by, first, looking at titles in order to check for relevance. Then, the participant checks the credibility which is particularly important during a medical search. Participant 9 relies on prior knowledge and past experience in order to evaluate websites authoritativeness. The participant believes that top search results, might not be reliable, but they are definitely the most relevant/helpful results to the topic so it is helpful to open them. Having citations and known health organizations gives more credibility while having advertisement content reduces the trustworthiness level in the source. News websites are not the best place to get medical related information as they lack authoritativeness when dealing with domain specific information such as medicine. Consistency is a key when it comes to online search. The participant compares different websites in order to reach a final decision. In case no consistency exists, the participant tries initiating a new search query with different keywords. In case of no consistency, the participant looks at the dates the information was published in order to check whether the non-agreement happens because of time difference. The participant believes that search results appearing more than once in the search result page shows that the source is important and reliable.

B.8 Participant 10

Participant 10 finds that world known organizations, news and medical sources are trustworthy when searching for medical information. The participant believe that rank is not a sign of the search results' reliability and is just a sign of how popular the page is. It is important for the participant to find consistency between reliable sources. Whenever there is an inconsistency, it is important to figure out the reason. Studies and references are important in determining the reliability of the search result content. Whenever the participant is skeptical about the content of a webpage such as natural and homeopathic treatments, he/she opens the listed references to read more about the topic.

B.9 Participant 12

Participant 12 starts exploring search results by opening the top ranked links while ignoring top advertisement links. The participant checks the content credibility by looking at the authors (in the medical domain, whether they are doctors) and finds short concise content more helpful and does not trust websites with a large content of advertisement. The participant's prior knowledge of the medical treatment effects the judgment of information in the search result page. The website design is an important factor when judging information helpfulness: if the website is not well designed (for example it does not look professional: black background, no images), then the participant will spend less time going through the content. The participant does not trust search result pages promoting health fads where certain keywords are used (such as using "healthy living" wording). Looking at the URL domain, the participant determines the level of credibility of the website. For example, .gov means that the website is governmental and has a high level of trustworthiness.

B.10 Participant 13

Participant 13 decides to click on the websites by looking at the titles to check whether they contain exact words as the search keywords (most relevant search result pages). The participant picks the most relevant URL to the topic, if it is legitimate, then the final decision will be based on the content of that URL and the participant will not check further resources. Whenever there is a website that does not give a clear answer about the treatment efficacy (discussion -when it helps and when it does not), the participant opens the URL and reads more details about both sides of the story. Further, the participant does not trust personal blogs when performing an online medical search.

B.11 Participant 14

Participant 14 explores the SERP results by reading the titles starting with the top search results . In case the title is relevant, the participant goes through the snippets. If the snippets contain interesting information, the participant opens the link to read more. If the title is not relevant, then the participant goes to the next search results page. The participant does not trust websites containing advertisement content, news websites and personal blogs and finds research-oriented content to be more credible. Further, the participant does not find complicated medical scientific information helpful when doing online medical search and feels that listing contact information in the URL gives it more credibility. The participant checks for consensus between different sources in order to make up the final decision. Having more than one website refereeing the same resource makes it more reliable.

B.12 Participant 15

Participant 15 searches for keywords in the titles and skims through the snippets in order to decide which URLs to click on. The participant does not trust personal blogs and advertisement content as a source of medical information and prefers research-based resources. Further, The participant does not rely on websites containing spelling mistakes and finds long content not to be helpful when searching online. The website layout (such as background color) is important to determine how trustworthy the website is. Finally, the participant counts the number of websites that state the treatment helps versus the number of websites that claim it does not. Finally, the majority will be the participant's decision.

B.13 Participant 16

Participant 16 starts exploring the search results by checking the first website in the SERP page. Later, the participant evaluates the authoritativeness of the website by looking at its domain. Educational and government websites are usually very trustworthy. The participant always looks for research papers, educational websites and world recognized health organizations when searching for medical related topics and avoids advertising content. The page layout/design gives a general idea about the reliability of the website. The authors' affiliation and mentioning known health organizations play an important role in evaluating the credibility of the website content. The participant looks for the references list in order to check whether there are reliable citation and trust the information-even though the website might not be credible. If the content is very scientific and has many technical terms, this makes the website less helpful and harder to read. The participant opens non-credible website when the majority of search results agree with the content of that website.

B.14 Participant 17

Participant 17 opens websites based on prior experience i.e. the participant will only open websites that have been reliable and helpful in the past. As a result, the participant does not trust and does not open new unknown search results. The participant does not trust information in personal blogs and search result pages with long hostnames containing multiple words or punctuations (such as health-women-coffee.command). The participant usually opens the top search result pages and ignores the ones in the bottom of the SERP page or the ones in the next pages because he/she believes that higher ranked websites are the most viewed and searched for. If the top websites are not relevant, then do the search again with different words. When opening a search result page, the participant reads carefully the first and last paragraphs to understand the major idea of the website. The participant usually trusts news websites. However, he/she does not trust some medical sources like WebMD because they might exaggerate symptoms and, if used for diagnosis, it may lead you to believing that you are sick of a dangerous disease. In case of no consistency between search results, the participant looks at the dates the information was published in order to check whether the non-agreement happens because of time difference.

B.15 Participant 18

Participant 18 avoids websites selling products, personal blogs and general-purpose search result pages (such as eHow and Yahoo Answers). The participant is more interested to read about why the treatment does not help and always clicks on the link explaining why the treatment fails. The participant checks for authors credentials, known health organizations and the page design in order to evaluate the credibility of the information. Furthermore, the participant looks for research-based resources and relies on the majority in order to decide whether the treatment is helpful or not. When the participant has a prior belief, even after going through the search results, he/she does not change the prior belief. The participant checks the list of references in order to make sure they are relevant to the information in the search results page.

B.16 Participant 19

The participant looks at the URL in order to determine the credibility of the web page. Participant 19 trusts scientific articles and journals and usually opens familiar known resources and looks for references when reading the page content. The participant does not trust news websites, personal blogs and general question answering websites (such as Wiki how or eHow etc.) as they are not a credible source for health information. Additionally, the participant does not trust celebrity doctors because he/she believes that the main purpose of their online content is advertising. Next, participant 19 finds that the page layout is a good indicator of the reliability of the web page. The participant pays specific attention to the writing style of the content and finds that, for example, very dramatic or very flowery writing style is less trustworthy and reliable. Having consistent information among page results helps the participant formulate the final decision

Appendix C

Terminology

Advocacy websites Advocating specific actions or policies, or claiming to be the best in providing the related information without official ties.

Authoritativeness The trustworthiness and reliability of the source of content in the search result pages.

Bottom-up approach An inductive approach where the codes are generated from the meaning discovery of specific instances, contexts, and individuals of the think-aloud data (Gu, 2014a).

Coding Applying a pre-established set of categories to the data according to explicit, unambiguous rules, with the primary goal being to generate frequency counts of the items in each category (Maxwell, 2012).

Crowdsourcing: Using a crowd-source online platform to recruit individuals to perform specific tasks.

Eye-tracking Recording eye-movements while performing a specific task then using the recorded data to perform evaluation measures (Kelly et al., 2009).

Interval entropy Measure the distribution of the time interval between successive retweets and distribution of distinct users involved in retweeting (Ghosh et al., 2011).

Linguistic Inquiry and Word Count (LIWC) A transparent text analysis program that counts words in psychologically meaningful categories (Tausczik and Pennebaker, 2010).

Majority The majority of the search results stating that the treatment helps or that the treatment does_not_help or looking for a consensus of different search results.

Misinformation A piece of information spreading in the online media confirmed to be false by reliable sources (Coady, 2006).

Optimism bias Expecting positive events in the future even when there is no evidence to support such expectations (Sharot et al., 2007).

Prior belief Trusting the information that agrees with our prior knowledge and disregarding facts that contradict with it, regardless of the actual truth (White and Horvitz, 2015).

Quality A set of criteria to define the quality of information on a search result page including the presence of statistics and studies, advertisements and list of references, the date the information was posted, the way the information is presented (images, text length), and the level of content readability.

Readability A measure of how easy it is to understand a piece of text (Feng et al., 2010b).

Rumor A piece of information spreading in the online media confirmed to be false by reliable sources (Coady, 2006). (Similar to the term *misinformation*)

Social factor Make a decision based on other people's beliefs and experiences. For example, whether a friend or a family member's opinion effects our decision about the effectiveness of a medical treatment.

Think-aloud The method of asking subjects to articulate their thinking and decision-making as they perform a specific task (Kelly et al., 2009).

Top-down approach A theory-driven approach where the codes are generated based on theoretical background and existing research (Gu, 2014a).