Spring 2014

Computational methods for mining health communications in web 2.0

Sanmitra Bhattacharya
University of Iowa

COMPUTATIONAL METHODS FOR MINING HEALTH COMMUNICATIONS IN WEB 2.0

by

Sanmitra Bhattacharya

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Computer Science
in the Graduate College of
The University of Iowa

May 2014

Thesis Supervisor: Professor Padmini Srinivasan

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Sanmitra Bhattacharya

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Computer Science at the May 2014 graduation.

Thesis Committee: _____
Padmini Srinivasan, Thesis Supervisor

_____
Alberto Maria Segre

_____
Philip Polgreen

_____
Nick Street

_____
Gautam Pant

To my family

ACKNOWLEDGEMENTS

Pursuing a PhD is an exciting but challenging journey. It cannot be achieved without the help and support of a lot of people from both professional and personal life. I would like to thank them all for helping me make this happen.

First and foremost, I would like to thank my advisor and mentor, Padmini Srinivasan for all her support since my first day at Iowa. She not only provided great guidance but also motivated and encouraged me in working towards various research problems.

I am grateful to other thesis committee members, Alberto Segre, Phil Polgreen, Nick Street and Gautam Pant for their valuable feedback and comments on my research. In particular, I would also like to acknowledge Dr. Polgreen for motivating the research problem of studying engagement of health agencies.

I would also like to thank Jerry Suls (Psychology) for collaborating on the belief work and Bob Boynton (Political Science) for giving access to various datasets.

I am thankful to have an opportunity to work with great colleagues from the Text Retrieval and Text Mining group. I would like to thank Hung Tran, Chao Yang, Yelena Mejova, Viet Ha-Thuc, and Chris Harris for their help and support, especially Hung who helped me a lot during the initial phase of the belief surveillance work. I would also like to thank my internship supervisors Michael Cantor (Pfizer) and Luca Toldo (Merck KGaA). I have learnt at lot from them and I thank them for their continued support.

Outside of academic life, I would like to thank my closest friends at Iowa, Gaurav Kanade and Biswanath Maity.

Last but not least, I would like to thank to my family for their love, support and encouragement. Without them it would have been impossible for me to finish this five-year long journey.

ABSTRACT

Data from social media platforms are being actively mined for trends and patterns of interests. Problems such as sentiment analysis and prediction of election outcomes have become tremendously popular due to the unprecedented availability of social interactivity data of different types. In this thesis we address two problems that have been relatively unexplored. The first problem relates to mining beliefs, in particular health beliefs, and their surveillance using social media. The second problem relates to investigation of factors associated with engagement of U.S. Federal Health Agencies via Twitter and Facebook.

In addressing the first problem we propose a novel computational framework for belief surveillance. This framework can be used for 1) surveillance of any given belief in the form of a probe, and 2) automatically harvesting health-related probes. We present our estimates of support, opposition and doubt for these probes some of which represent true information, in the sense that they are supported by scientific evidence, others represent false information and the remaining represent debatable propositions. We show for example that the levels of support in false and debatable probes are surprisingly high. We also study the scientific novelty of these probes and find that some of the harvested probes with sparse scientific evidence may indicate novel hypothesis. We also show the suitability of off-the-shelf classifiers for belief surveillance. We find these classifiers are quite generalizable and can be used for classifying newly harvested probes. Finally, we show the ability of harvesting and

tracking probes over time. Although our work is focused in health care, the approach is broadly applicable to other domains as well.

For the second problem, our specific goals are to study factors associated with the amount and duration of engagement of organizations. We use negative binomial hurdle regression models and Cox proportional hazards survival models for these. For Twitter, the hurdle analysis shows that presence of user-mention is positively associated with the amount of engagement while negative sentiment has inverse association. Content of tweets is also equally important for engagement. The survival analyses indicate that engagement duration is positively associated with follower count. For Facebook, both hurdle and survival analyses show that number of page likes and positive sentiment are correlated with higher and prolonged engagement while few content types are negatively correlated with engagement. We also find patterns of engagement that are consistent across Twitter and Facebook.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1
INTRODUCTION

Research in the health sciences has seen unprecedented growth in the past few decades. There has been an increasing emphasis on health sciences research using digital media. The use of electronic medical records for automatically identifying patient population for clinical trials or the use of wireless wearable sensors and smart phone applications to track patient vitals in real time are some examples. There has also been an increasing focus on the use of digital media such as online news papers and blogs, Youtube, Facebook, Twitter, Tumblr or Flickr as communication tools, especially in the health domain. A recent survey shows that almost 18% of online adults use Twitter[1], which generates over 500 million tweets[2] per day from over 500 million users around the globe[3]. Around 67% of online adults use Facebook[4] which has over 1.11 billion active users[5] who spend around 20 minutes per Facebook visit[6]. A recent study by PricewaterhouseCoopers showed that in the United States, 24% of adults post about their health experiences and updates on social media with 16% of

[1]http://pewinternet.org/Reports/2013/social-networking-sites.aspx

[2]https://blog.twitter.com/2013/new-tweets-per-second-record-and-how

[3]http://twopcharts.com/twitter500million.php

[4]http://pewinternet.org/Commentary/2012/March/Pew-Internet-Social-Networking-full-detail.aspx

[5]http://news.yahoo.com/number-active-users-facebook-over-230449748.html

[6]http://expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats/

them posting reviews of medications, treatments, doctors or health[7].

Two most popular social networking websites[8], Twitter and Facebook, have invigorated a wide range of health sciences studies in recent years. The scope of such studies is quite varied. The use of Twitter for disease surveillance [68] or analyzing the potential of Facebook in improving public health [53] takes a much broader scope compared to disease specific studies. Studies to monitor health information dissemination for specific health-related issues like dental pain [37] or concussion [73] show how one might use social media to get insight into the specific health problems. While most of these studies deal with information disseminated by individual users of social media platforms, few have also focused on how health organizations are involved in social media-based communications [75, 35].

First, at the individual level, we ask the general question: *What are the beliefs held by social media users?* While sentiment analysis and opinion mining from social media have been studied extensively, surveillance of beliefs, especially those related to public health, have received considerably less attention. There is a long-standing recognition that social and bio-behavioral scientists and policy makers need accurate and up-to-date information about the broad spectrum of beliefs and opinions voiced in the population [21]. As an example, having an accurate estimate of the frequency of people who believe the HPV increases the risk of cervical cancer [54] or that using deodorant increases the risk of cancer [29] allows public health scientists to decide

[7]http://www.pwc.com/us/en/health-industries/publications/health-care-social-media.jhtml

[8]http://www.ebizmba.com/articles/social-networking-websites

whether there is a need to mount a special health campaign to correct those beliefs. Large-scale survey approaches using mail, telephone, and special websites can provide useful data, but such approaches, by definition, having already formulated the content of their questions, do not tap the naturally occurring opinions or beliefs expressed by people. Moreover, typically there are always time-delays between preparation of the survey questions and administration. The development of a methodology that assesses the prevalence of the naturally occurring expression of beliefs and opinions would be immensely useful. This motivates our first question stated earlier.

Second, at the organizational level, we ask the general question: *How can organizations be more engaging on social media?* The importance of social media for communicating messages to a broad audience is well acknowledged in many domains such as journalism, politics, marketing, education or entertainment. In healthcare as well organizations such as the Centers for Disease Control and Prevention (CDC) or Food and Drug Administration (FDA) play a crucial role in informing the public of critical pandemic events like the spread of H1N1 or Coronavirus. They also communicate general public health information such as drug recall or sexual health information to the public. It is generally recognized that different types of social media postings get different levels of attention or response from the population. What causes these differences is not well understood. We use the term 'engagement' for studying interactions designed to promote some common goal [58]. The nature of engagement among the general population is an active focus of research in health sciences and in marketing [43] traditionally conducted by surveying populations of

health-information seekers. Studying engagement can inform organizations about topics the public is interested in or changes they need to employ to reach more people for topics fostering insufficient engagement. Motivated by this, in the social media context we ask the second general question stated above.

In brief, we address two broad questions about health communication at the individual and organization levels. We do this by 1) developing a computational framework for belief surveillance and 2) developing a computational framework for assessing the engagement of organizations. We then answer a series of more specific questions using these two frameworks. The two questions are connected in that we aim to identify health notions for which the public holds misconception (belief research). We then suggest techniques in which health organizations can effectively communicate relevant information (engagement research). Overall, this thesis contributes to the broader computer science research area of text mining and data mining applied to social media.

CHAPTER 2
RELATED RESEARCH

Social networking websites such as Twitter have invigorated a wide range of studies in recent years ranging from consumer opinions on products [40] to prediction of box-office revenue of movies [6] or tracking the political discourse [19]. Since tweets represent various levels of personal communication they can be used for conveying a wide range of information — from vetting out personal thoughts to broadcasting social awareness messages. Because of the 140-character length limitation and use of colloquial terms and slang expressions on Twitter, tweets tend to be noisy (in the form of abbreviations, misspellings, out-of-vocabulary words, etc.). In spite of these linguistic challenges that Twitter poses, patterns and trends aggregated from Twitter can be meaningful. This is indeed the rationale behind several studies and implementations involving Twitter [41, 46, 39, 87].

2.1 Health Communication of Individual Users

Recent studies in biomedicine and healthcare informatics have seen an increasing use of Twitter. In one of the earliest studies on the use of social media for health care, Hawn [36] describes an online diagnosis and treatment program by a Brooklyn-based primary care company called "Hello Health"[1]. It provides a secure interface for physicians and patients to directly interact though web videos, instant messages or tweets. Along these lines, studies have also shown the importance of Twitter as

[1]http://hellohealth.com/

a platform for sharing and spreading health related information by health librarians in the U.S. [30]. A recent article in *Oncology Times* reported the power of Twitter in spreading cancer-related information by prominent oncologists and biomedical researchers[2]. Researchers not only use Twitter for disseminating public health information but also for gathering information, correcting misinformation, communicating with patients, etc.

Researchers have also proposed models of health-related information flow on Twitter. For example, Murthy et al. [56] study information flow in cancer-related networks on Twitter and propose methods for inferring authority of individual authors and their tweets using network analysis and visualization techniques. They conclude that analysis of social networks can predict health outcomes for cancer.

Another direction of Twitter-based research that has found prominence in life sciences is the study of tweets on particular diseases/treatments to find sub-topics of discussion. The general framework for these studies is to collect a set of tweets on a given topic and manually categorize them into several sub-topics. For example, studies have been conducted to find the prevalence of tweets on problem drinking. Tweets collected during a two week period were categorized by time periods to show that problem drinking is present at a larger scale during the New Year's Eve holiday compared to a non-festive weekend [84]. Also tweets on problem drinking are more common from Friday to Sunday with a tendency among users to establish such behaviors as acceptable and expected. Similar studies were conducted for identifying

[2]http://bit.ly/1rdkHFj

sub-categories of dental pain [37], concussion [73], flu [20, 45] or use of antibiotics [67]. In a recent study Golder and Macy [31], over 500 million Twitter messages were studied to identify the diurnal and seasonal mood rhythms of individuals across the globe. Another recent study analyzed the nature of tweets on epilepsy for a 7-day period and found that epilepsy-related tweets could be classified into 4 predominant categories — metaphorical, individual experiences, informational and satirical [51]. 41% of these tweets were found to be derogatory and related to stigmatization of epilepsy in the society!

Another area which has seen an increasing use of social media is disease surveillance. Popular social networking websites (e.g. Twitter, Facebook, etc.) have been shown to be important sources for monitoring real time data for disease outbreaks. Twitter-based influenza epidemics detection by Aramaki et al. [1] shows the importance of mining social media for early stage detection of a disease outbreak. In a similar study Culotta [20] reinforces the timeliness of outbreak detection using tweets and found high correlation between patterns gathered from Twitter messages and CDC statistics. The use of Twitter traffic and tweets has also been shown for not only gauging public interest about a particular disease (H1N1), but also for tracking disease outbreak in real time [68].

In comparison to such disease specific studies, few studies take a broader scope. For example, Paul and Dredze [62] propose a topic model based approach to explore Twitter for public health research. They identify various public health topics that can be studied using Twitter such as, syndromic surveillance, behavioral risk factors, and

symptoms and medications taking into account the geographic distribution for each such topic. Donelle and Booth [24] studied the topics of health communication from a set of 2400 tweets following guidelines of the Public Health Agency of Canada's Determinant of Health to find that the most predominant topics are health services, personal health issues and advertising. Similarly, Prier et al. [63] study methods for mining general health topics from Twitter. Using an LDA topic model they identify health-related issues that garner a lot of attention on Twitter, such as, weight loss programs, Obama's health reform policy, marijuana uses, etc.

In contrast to the above studies, some studies have also shown the ineffectiveness of Twitter in influencing public health concerns. For example, Prochaska et al. [64] studied the impact of twitter on the cessation of smoking habits based on Twitter accounts that promote anti-smoking causes. They found that majority of tweets from these accounts link to some commercial websites for quitting smoking while a significant number of them had tweets on e-cigarettes. The remaining tweets were largely inconsistent with clinical guidelines and thereby irrelevant to the promotion of awareness against ill effects of smoking.

In spite of flourish of research in various health-related topics on Twitter, one direction that is relatively unexplored is the use of Twitter for mining health beliefs in the population. There is a long-standing recognition that social and bio-behavioral scientists and policy makers need accurate and up-to-date information about the broad spectrum of beliefs and opinions voiced in the population [21]. As an example, having an accurate estimate of the frequency of people who believe the

HPV increases the risk of cervical cancer [54] or that using deodorant increases the risk of cancer [29] allows public health scientists to decide whether there is a need to mount a special health campaign to correct those beliefs. Large-scale survey approaches using mail, telephone, and special websites can provide useful data, but such approaches, by definition, having already formulated the content of their questions, do not tap the naturally occurring opinions or beliefs expressed by people. Moreover, typically there are always time-delays between preparation of the survey questions and administration. The development of a methodology that assesses the prevalence of the naturally occurring expression of beliefs and opinions would be of substantial benefit to bio-behavioral scientists. To address this problem we proposed a novel surveillance framework and show its applicability in mining and guaging naturally occurring health beliefs in the population.

2.2 Health Communication of Organizations

Besides individuals, many organizations are also very active on various social media platforms. Social media provides a unique opportunity for organizations to directly interact with their stakeholders or potential customers. Social media enables them to not only promote new products but also get instant feedback on existing products. Lovejoy et al. [49] studied the use of Twitter by 73 nonprofit organizations and found that most of these organizations use Twitter not as an interactive platform but more for dissemination of information. These findings are also corroborated by another study [81], where the researchers found that nonprofit philanthropic

organizations seldom engage with the social media population and most of the information flow is unidirectional. Similar observations were also made from 275 nonprofit organizations on Facebook [80].

Researchers [66] have also investigated the use of Twitter in a set of 93 Fortune 500 organizations and found that most of the "dialogic" conversation stem from response to users, posting questions, etc. However, they found less than one percent of posts were targeted to specific users and generally posts were targeted to broader audiences. The analysis of the tweeting behavior of 60 government agencies is also in line with the above studies as most of the communication is one-way with lack of any interaction with the social media users[82].

Armstrong and Gao [2] studied the use of Twitter by nine news organizations over a four month period and found that the nature of tweets from regional, local and national media differ widely and crime and public affairs are the most prevalent topics for tweets from these organizations. Analysis of Facebook presence and activity of national tourism organizations in Europe reveal that these organizations have minimal presence on Facebook and they rarely take advantage of the richness of user-generated content for promoting tourism on Facebook [70]. Across all these studies we find that social media platforms present unique opportunities for user engagement but organizations seldom realize their full potential.

Very few studies have focused on studying the engagement of health organizations on social media. Few of these have focused on the role of health departments, primarily at the state level, to explore the extent of their social media adoption, their

interaction with patients and topics of discussion. Thackeray et al. [75] found that 60% of state health departments have a presence in at least one social media platform with Twitter and Facebook being the most preferred platforms, posting mostly informative messages but having little interactivity, i.e. engagement, with patients. The analysis of connections among state health departments on Twitter and Facebook reveals that health departments are more connected on Twitter than on Facebook with geographic proximity being one of the primary determinants of connectivity [35]. In an effort to study the determinants of audience engagement in health promotion on social media Neiger et al. [59] defined and discussed various general hierarchical metrics for engagement. In contrast to these studies, this thesis focuses on studying 'engagement' of Federal health departments with general Twitter and Facebook users. We go beyond simple network analysis or content analysis of tweets and study the factors influencing engagement using count data regression and survival models. Additionally, we also study the temporal aspects of 'engagement' which has not been studied in the context of health departments.

CHAPTER 3
SURVEILLANCE OF BELIEFS OF INDIVIDUAL USERS

In this chapter we discuss belief surveillance of individual users on social media. We address the general research question: *What are the beliefs expressed by social media users?* Before exploring our approach towards this problem we introduce some of the key concepts.

Belief is formally defined as "a feeling of being sure that someone or something exists or that something is true"[1]. *The earth revolves around the sun* is an example of a widely-accepted belief. Belief formation is one of the most elementary functions of the human mind and is thus central to studying the philosophy of mind as well as knowledge (called 'epistemology')[2].

Belief is considered as "propositional attitude" by contemporary philosophers. A proposition is simply an idea or concept expressed as a sentence. A propositional attitude can then be viewed as a mental stance or opinion about a proposition.

In our discussions we also refer to propositions as probes. We explore two distinct types of probes; those related to *causes* of ill health and those related to their *treatment*. These are categories of information typically sought by individuals afflicted in some way by disease or ill health. 'Talcum powder causes breast cancer' and 'garlic treats high blood pressure' are two examples. Each reflects a directed binary relationship linking two concepts. Of course one might explore probes involving

[1]http://www.merriam-webster.com/dictionary/belief

[2]http://plato.stanford.edu/entries/belief/

more complex relationships. However, in our study we simply wish to demonstrate the opportunity for belief surveillance using social media and show its application in health care.

Our probes are also precise rather than general. Consider for example, the more general statement 'There are many causes of cancer' or 'Food treats illness'. These 'catch-all' statements are both less interesting and more difficult to use as probes. First, people are less likely to converse at this general level. People are more interested in specific causes and particular food items as treatments for specific problems. Second, even if we do find conversations around these general probes, our results would not be useful. Instead, estimates of public beliefs can inform public health information and educational strategies. So we use specific propositions as probes.

A probe may also have a truth status as determined by scientific evidence. For example, there is indisputable scientific evidence that *the earth revolves around the sun* and it's truth status is *true*. For some others, such as *cell phone causes cancer* the scientific community has divided views and it's truth status is *debatable*. Some probes such as *vaccine causes autism* have no supporting scientific evidence and it's truth status is *false*. Truth status for probes can be determined by consulting an expert in the respective field. For health related probes, such as *vaccine causes autism*, a physician may be consulted to determine the truth status. However, we note that even experts (physicians, for example) may differ in opinion. These cases are particularly tricky and consensus about a truth status may be obtained by consulting multiple

experts.

A layperson may also express an opinion about a given proposition or probe. One may agree with a probe and *support* it or disagree with a probe and *oppose* it. If unsure about the probe one may even *doubt* it. For example, for the probe *cell phone causes cancer*, Alice may say "I know for sure that cell phone causes cancer" and thus support the probe. Bob may say "There's no connection between cellphones and cancer" and oppose it. Carol may ask "Is there any link between cancer and cell phones?" and express doubt about the probe.

We coin the term *belief surveillance* for studying the level of *support, oppose or doubt* for specific probes. As previously mentioned, this is in part inspired by efforts on disease surveillance using social media [68]. We use Twitter as a platform for studying belief surveillance. Unlike other social media platforms, namely Facebook, tweets are short and precise and fit well with our surveillance framework which focuses on very specific probes. Given tweets that pertain to the topic of a probe, we determine the position taken by each tweet vis-à-vis the probe. The positions we consider are *support, opposition or doubt*. We can then summarize across the tweets. For this we propose summary measurements of support, opposition and doubt. We use these to compare public attitudes towards a set of true, false and debatable probes.

Table 3.1 provides several example tweets to illustrate support, opposition and doubt w.r.t. the probe: *Vaccine causes autism*. The first tweet supports the probe, the second opposes while the third questions it, i.e., expresses doubt. The fourth tweet does not express a position while the last tweet, at least on the surface, is not

relevant.

Table 3.1: Example Probes & Tweets.

Probes: *Vaccine causes Autism*		
TW1	Terrible! Vaccines - Government Scientists Hide Vaccine-Autism Link #health	Support
TW2	Listen Bachman, you cannot get autism from a VACCINE	Oppose
TW3	Are the MMR vaccine links to autism unfounded?	Doubt
TW4	The Vaccine and Autism Debate	Other
TW5	Is there a vaccine against JennyMcCarthy? I'd even take autism to get her off my TV.	Not Relevant

Belief is different from sentiment [76]. Notice that the first tweet expresses negative sentiment while the second tweet maintains a neutral tone. If we took 'Terrible!' out of the first tweet the belief would be the same, but sentiment would reduce towards neutral. Likewise if we preface the tweets with a smiley icon, it does not change the level of belief expressed but it does modify the sentiment. Thus sentiment and belief are somewhat orthogonal aspects. Our goal is to study expressions of health beliefs in social media both in terms of probes and in terms of the poster's position.

3.1 Computational Belief Surveillance Framework

Our computational framework for belief surveillance comprises of two distinct components (see Figure 3.1). The first component deals with surveillance of probes

while the second component deals with harvesting of probes.

3.1.1 Probe Surveillance

Given an input proposition in the form of a probe our belief surveillance begins by retrieving tweets related to the probe (e.g. *vaccine causes autism*). Each probe comprises of two concepts connected by cause/causes or treat/treats. In the above example, the concepts are *vaccine* and *autism*. We first tried a direct search using a conjunctive query made of the two concepts along with various synonymous forms of cause and treat. For 'cause' we identified synonyms such as creates, raises, increases, associates, relates, etc. But then we had to also consider the various tenses (raised, will increase). Further, because this is social media, we had to consider likely misspellings. Then for each linking term we had to consider negation in it's various forms (does not, did not, won't, cannot, certainly doesn't etc.). The treats relationship had its own seemingly unbounded set of linking terms. This approach quickly became non-productive. It would also not scale with probes emphasizing other types of relationships (precedes, interacts with, accelerates, etc.). Thus, we designed a simpler strategy using just a single query which is the conjunction between the two concepts. All searches are performed using the Twitter Search API[3] which limits retrieval to a maximum of 1500 tweets from a prior 7-day segment.

The dashed box marked 'Probe Surveillance' in Figure 3.1 shows the two-step decision flow applied to each retrieved tweet. Retrieval using the two concepts alone

[3]https://dev.twitter.com/docs/api/1/get/search

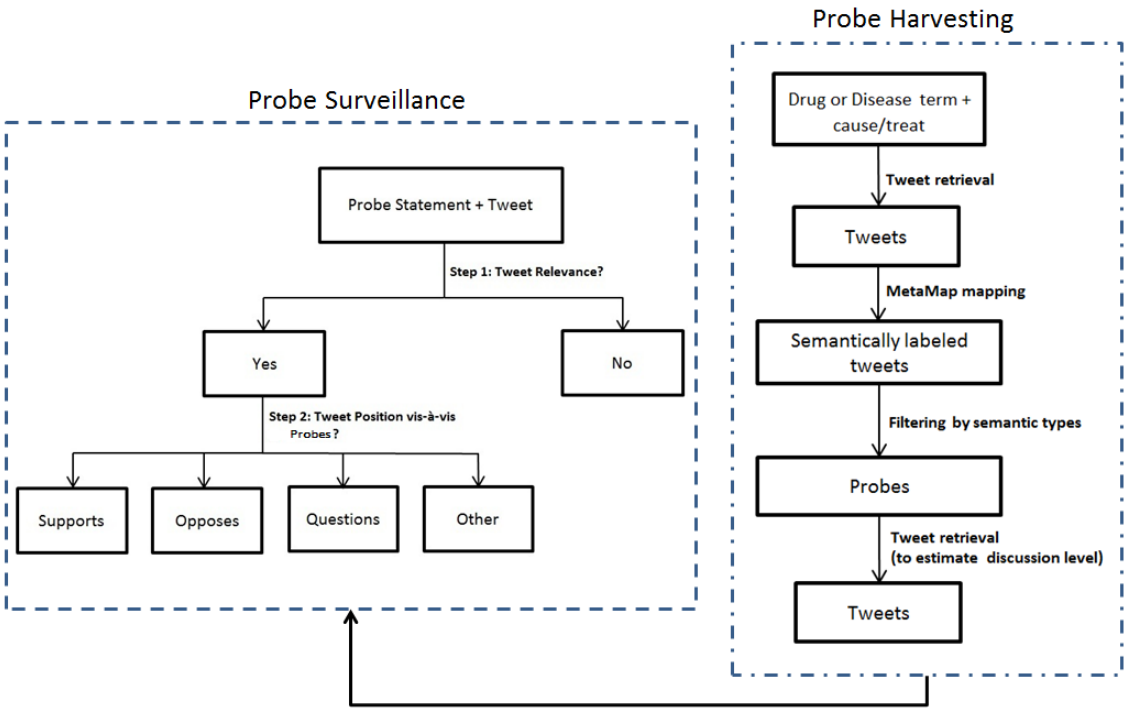


Figure 3.1: Two-Step Belief Surveillance

is clearly risky in terms of precision. We follow this with a refinement step (Step 1 in the figure) to find those relevant to the probe. In the example of Table 3.1 TW1, TW2, TW3 and TW4 are relevant whereas TW5 though it satisfies the query is not. After we are confident in our retrieved tweets Step 2 (Figure 3.1) follows to determine tweet position vis-à-vis the probe. Each relevant tweet's position can be: support, opposition or doubt (i.e. questioning). A tweet may also be none of the above and fall into an 'other' catch-all category.

In our experiments, which focus on health-related probes, a physician was consulted to determine the truth status of probes. Each probe was classified into one of three categories – true (T), false (F) and debatable (D).

For a probe (S_i) we provide intuitive measures for assessing public support, opposition and doubt from tweets. A simple estimate for support i.e. the proportion of relevant tweets agreeing with the probe is shown in Equation 3.1. Similarly, 'opposition' defined as public disagreement to S_i has a parallel measure (Equation 3.2). Lastly, 'doubt' defined as public questioning of the probe is estimated using Equation 3.3.

Besides these measures using individual tweets, one may also be interested in identifying support, opposition and doubt for a group of probes. For example, it may be interesting to group probes by drugs or disease types and see if the level of support, opposition or doubt for one group differ from another. It may be that people have more support for alternative medicines on social media compared to prescription drugs. In this thesis we are interested in groups based on truth status. With this notion, we define aggregations again using intuitive approaches. First, given a set of probes, we can aggregate the extent of public support, opposition and doubt in them as a whole with averages (Equations 3.4, 3.5 and 3.6). These allow us to compare our estimates of public support for one set of probes versus another set. For example we could compare a set of probes known to be true with another acknowledged as false. Similar comparisons may be made, for example, concerning the level of opposition and of doubt. True probes with low estimated support and false probes with high support may be marked for special attention.

$$Degree\_Support(S_i) = \frac{\#\ relevant\ tweets\ supporting\ S_i}{\#\ tweets\ relevant\ to\ S_i} \tag{3.1}$$

$$Degree\_Opposition(S_i) = \frac{\#\ relevant\ tweets\ opposing\ S_i}{\#\ tweets\ relevant\ to\ S_i} \tag{3.2}$$

$$Degree\_Doubt(S_i) = \frac{\#\ relevant\ tweets\ questioning\ S_i}{\#\ tweets\ relevant\ to\ S_i} \tag{3.3}$$

$$Support(S_1, S_2, ..., S_N) = \frac{1}{N}\sum_{i=1}^{N} Degree\_Support(S_i) \tag{3.4}$$

$$Opposition(S_1, S_2, ..., S_N) = \frac{1}{N}\sum_{i=1}^{N} Degree\_Opposition(S_i) \tag{3.5}$$

$$Doubt(S_1, S_2, ..., S_N) = \frac{1}{N}\sum_{i=1}^{N} Degree\_Doubt(S_i) \tag{3.6}$$

N above is the number of probes in the set. These notions and measures are sufficient for probe surveillance for a set of probes.

3.1.2 Probe Harvesting

Besides the surveillance of pre-selected probes, our computational framework can also identify probes from the spontaneous and naturally expressed discussions on health topics in Twitter (see Figure 3.1). While there is a wide range of health topics that are discussed on social media we are particularly interested in probes related to causes and treatments of illnesses. Probes representing drug usage for treatment of diseases or causing adverse reactions [78] are of particular interest. A probe about an entity causing a disease or used for its treatment are also of interest. Using specific drug or disease names (such as aspirin, influenza, etc.) or more general hashtags (like #disease) along with 'cause/causes' and 'treat/treats' verbs we can retrieve relevant tweets from which we can identify our specific probes of interest.

Since we are interested in harvesting drug and disease-related probes we needed to identify these terms from tweets. The scale of social media poses challenge for man-

ual identification of these terms and thus we needed to employ automated methods. We chose a widely used tool, MetaMap [3], for identifying biomedical concepts and corresponding semantic types from tweets. MetaMap can identify coherent words or phrases from a particular sentence (tweet in our case) and map them to Unified Medical Language System[4] (UMLS) metathesaurus concepts. UMLS provides a standard set of health and biomedical vocabularies and related semantic concepts. The semantic concepts related to drugs or diseases are used to identify corresponding terms. When associated with 'causes' and 'treats' verbs the identified terms form probes of interest.

We can now monitor the discussion on these harvested probes using the 'Probe Surveillance' part of our framework.

3.2 Research Questions

As previously mentioned, the general research question that we address in this chapter is: *What are the beliefs expressed by social media users?* However, we chose to constrain the problem to only health beliefs pertaining to causes and treatment of illnesses. We ask several specific research questions that are outlined below.

- RQ1: Can we identify the levels of support, opposition or doubt for specific propositions from a population of social media users?
- RQ2: Can we build automatic classifiers for belief surveillance using off-the-shelf tools?

[4]https://uts.nlm.nih.gov/

- RQ3: Can we use our belief classifiers trained on a set of probes to classify new probes?
- RQ4: Can we identify naturally expressed beliefs in a population of social media users?
- RQ5: Can we estimate the scientific novelty of a probe identified from social media conversations?
- RQ6: Does the level of discussion or the level of support, opposition or doubt for probes change over time?

We address each of these research questions, RQ1-RQ6 in Experiments 1-6 (Section 3.3-Section 3.8), respectively. In Experiment 7 (Section 3.9) we extend Experiment 4 by *harvesting* probes for an extended period of time, specifically over an eleven month period.

3.3 Experiment 1

3.3.1 Goals

In this experiment we address RQ1: *Can we identify the levels of support, opposition or doubt for specific propositions from a population of social media users?* The primary goal of this experiment is to estimate the levels of support, opposition and doubt for a set of hand-picked probes. To achieve this we create a gold standard dataset for this set of probes using crowdsourcing. We then use various measures (Equations 3.1-3.6 in Section 3.1) to estimate the levels of support, opposition and doubt for this set of probes.

3.3.2 Dataset

Thirty-two specific probes regarding illness (Table 3.2) were manually identified by consulting several sources: a) sites listing medical rumors such as for cancer (CDC, NCI, FDA, WebMD, etc.), b) physicians, and c) current news (Google News). Each probe was judged as true (T), false (F) or debatable (D) by a physician. In most cases (e.g. True: DES causes cancer [61], False: vaccine causes autism [23], Debatable: chocolate causes acne [22]) we were able to find supporting scientific publications. An important point is that a probe is interpreted contextually, especially considering current events. For example, 'cantaloupe causes listeria' is clearly not true in a historic sense. However, the deadly events of listeria due to tainted cantaloupe at the time of data collection drives the decision of true, especially since we are exploring current social media. Other probes such as 'milk causes mucous' are generally believed but lack scientific validity[5]. Other myths include, for example, probes 13, 20, 22, and 26. Some probes represent widely held opinions, at least in some parts of the world (19 and 22). Probe 9 has a special position. Radiation is a known treatment for some cancers. However, the probe as structured focuses on radiation's other side as a causal agent for cancer. This *dual nature* of the causal agent makes this probe stand apart as we will see in our analysis later. Our list also includes highly controversial probes, as for example the first.

[5]http://www.dairycouncilofca.org/Milk-Dairy/Milk-Myths/Myth-2.aspx

Table 3.2: Set of 32 pre-defined probes.

#	**Probe**	**Truth Status** (T/F/D)
1	vaccine causes autism	F
2	smoking causes cancer	T
3	zocor causes muscle injury	T
4	aspirin causes bleeding	T
5	stress causes cancer	D
6	tea causes cancer	F
7	asbestos causes cancer	T
8	abortion causes breast cancer	F
9	radiation causes cancer	T
10	sunscreen causes cancer	D
11	radon causes cancer	T
12	aspartame causes cancer	F
13	lemon treats cancer	F
14	actos causes bladder cancer	D
15	alcohol causes cancer	T
16	milk causes mucous	F
17	DES causes cancer	T
18	avandia causes heart attack	D
19	honey treats allergy	F
20	cracking knuckles causes arthritis	F
21	obesity causes cancer	T
22	garlic treats blood pressure	F
23	plastic causes cancer	D
24	cantaloupe causes listeria	T
25	chocolate causes acne	D
26	vinegar treats blood pressure	F
27	cocaine causes glaucoma	T
28	tanning causes cancer	T
29	anesthesia causes learning disabilities	T
30	drospirenone causes blood clot	T
31	coffee causes cancer	F
32	cell phone causes cancer	D

3.3.3 Methods

Using the tweet retrieval strategy shown in Section 3.1 we retrieved 11,591 tweets retrieved for the 32 probes. These data are results from a 7-day time span ending on October 10, 2011. An average of 362 tweets were retrieved per probe (minimum 2, maximum 1,365). We identified a subset of these probes that had a maximum of 100 retrieved tweets per probe. We removed duplicates and near duplicates (using cosine similarity, threshold 0.8) and then selected tweets randomly. If less than 100 tweets are retrieved for a probe all are included in this dataset. The resulting subset of 2105 tweets was used to study public support, opposition and doubt for our 32 probes.

We manually annotated the 2105 tweets for relevance and position. This is referred to as gold-standard dataset in further discussions.

We obtained annotations using the oDesk[6] crowdsourcing platform. Compared to other popular crowdsourcing platforms like Amazon's Mechanical Turk, oDesk has been found to be superior in terms of quality of service [14]. Most importantly, oDesk allows employers to supervise, communicate and provide feedback directly to their employees. We took advantage of these features to hand-pick the annotators for tweet labeling. To qualify for this task, prospective employees were required to annotate a set of 10 tweets following our annotation guidelines (available online[7]). Out of 30 applicants 3 were selected based on their performance on the trial run.

[6]https://www.odesk.com

[7]www.cs.uiowa.edu/∼sbhttcha/Annotation_Guidelines.pdf

The dataset containing 2105 tweets for 32 probes was hosted on an online labeling platform designed specifically for this purpose. All three annotators worked on the same set of 2105 tweets and it took approximately 4 days to get the annotations and cost us 135 USD.

If all three annotators agreed we took it as the final judgment. Otherwise, we took the majority vote (2 out of 3) to decide the final labels. For the relevance judgment, all mismatches could be resolved using this strategy. For the decision on tweet position, it was possible for 3 annotators to select 3 different labels ('supports', 'opposes', 'other'[8]) for a particular tweet. In such cases, the tie-breaking decision was made by a fourth individual.

3.3.4 Results & Analysis

The Kappa score of inter-annotator agreement for relevance was 83.4%. The Kappa score for tweet position vis-à-vis probe was 84.08%. Kappa scores greater than 75% signify excellent inter-annotator agreement [27]. We attribute these high values at least in part to the quality of the oDesk platform and our selection mechanism. Table 3.3 shows the distribution of relevant and non relevant tweets. Overall the retrieved set was 70% relevant which is reasonable for a simple retrieval strategy. Note that we use only the relevant portions for gauging support, opposition or doubt.

Figure 3.2 presents the distribution of relevant versus non-relevant tweets retrieved by each probe. Precision is noted by the number above the relevance bar. All

[8]'doubt' is identified from 'other' category tweet if it contains a question mark (?)

Table 3.3: Tweet Relevance.

	Relevant	**Not Relevant**
Full Set	1483 (70%)	622 (30%)
Treat Probes	56 (61%)	35 (39%)
Cause Probes	1427 (70%)	587 (30%)
True Probes	789 (77%)	227 (23%)
False Probes	366 (67%)	178 (33%)
Debatable Probes	328 (60%)	217 (40%)

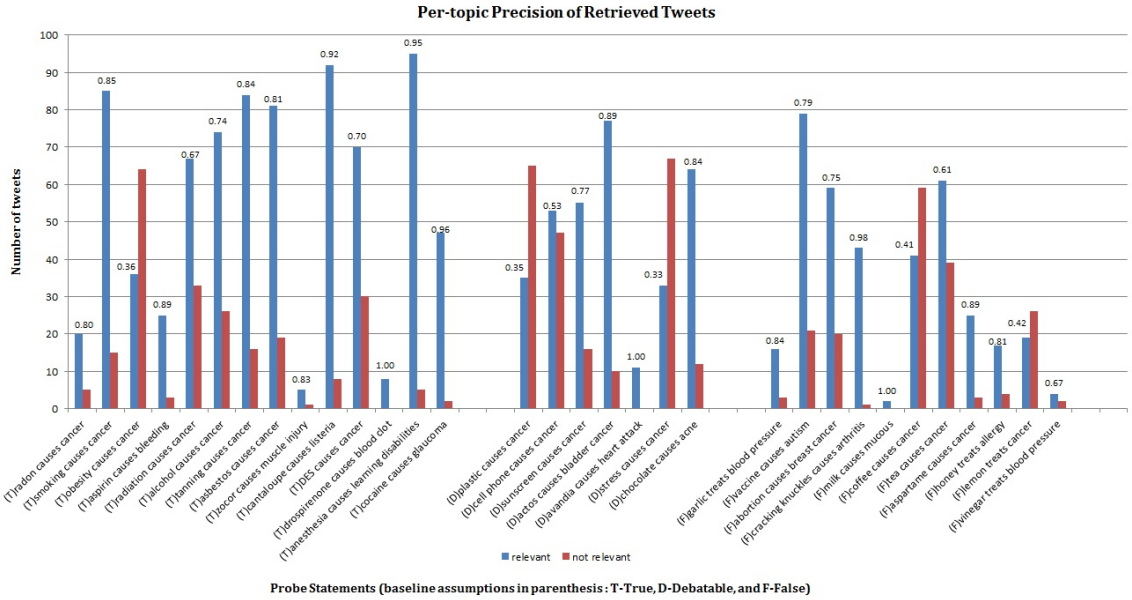


Figure 3.2: Per-topic Precision of Retrieved Tweets

our charts are clustered into 3 groups; the leftmost represents true probes (T), the rightmost represents false probes (F) and the middle probes are debatable (D).

While most of the probes in this experiment have high retrieval precision, for a few this is lower than 0.5. 'Obesity causes cancer' and 'plastic causes cancer' are two such examples. Not surprisingly we find that probes discussed recently in news generate a lot of conversation in Twitter and their tweets are generally relevant. The time period of our tweet collection spans events such as the listeria outbreak due to tainted cantaloupe, the publication of an article (in *Pediatrics*) reporting that multiple anesthesia exposures received by children under two may lead to learning disabilities, and the Actos class action lawsuit filed over bladder cancer risks. For all of these 'prominent' probes retrieval precision is very high. However, despite the appearance of a recent study from the WHO's International Agency for Research on Cancer reporting increased risk of cancer due to cell phone use, the precision for 'cell phone causes cancer' is only 0.53. On manual inspection we note that 39 out of the 47 false-positive cases appear to be random co-occurrence and usually in a spamming context. Perhaps terms such as 'cell phone' have a higher chance of appearing in a spam tweet. However, this could not have been predicted. Another reason for high precision in some cases is that some of the query term pairs (such as, anesthesia and learning disability) are unlikely to co-occur in any other context. Consider in this regard again listeria and cantaloupe. Hence we expect high precision of tweet retrieval for such pairs. On the opposite side, there are several contexts in which plastic and cancer may co-occur. For example with 'plastic + cancer' we get tweets like: *Plastic*

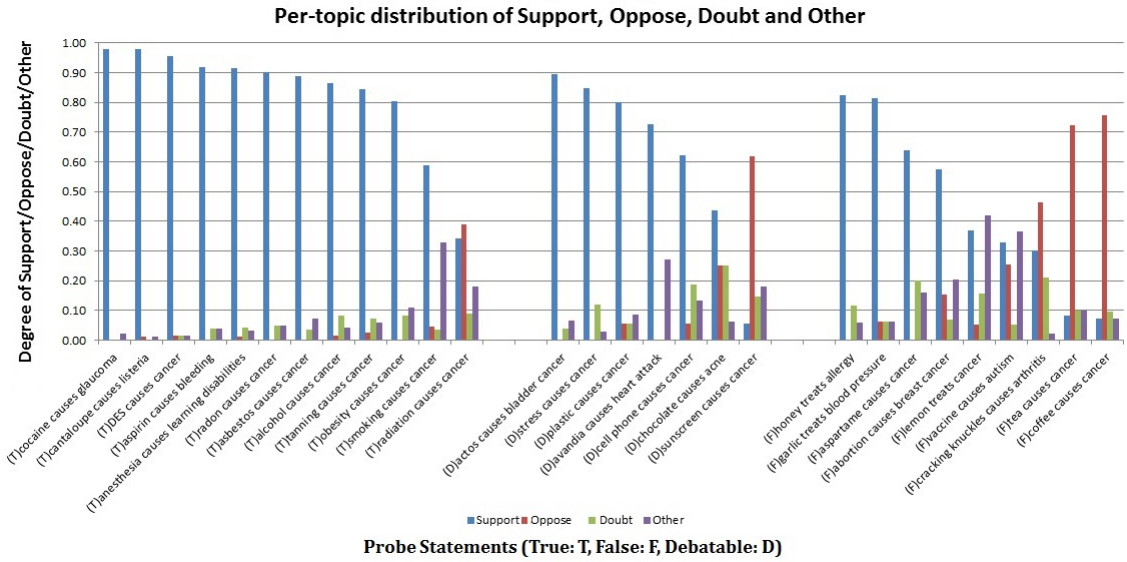


Figure 3.3: Tweet Position vis-à-vis Probe

surgery blog:Men can get breast cancer too (although rare). Thus the low precision for 'plastic causes cancer' may be explained.

Table 3.4 presents a summary of the positions taken by the 1483 relevant tweets in the dataset. We note that around two-thirds of our tweets support the input probe. 8% of the tweets were deemed as expressing 'doubt', reducing the size of the 'other' category from 284 to 165.

Table 3.4: Tweet Position vis-à-vis Probe.

Support	**Oppose**	**Doubt**	**Other**
982 (66%)	217 (15%)	119 (8%)	165 (11%)

Figure 3.3 presents support, opposition and doubt (and other) by each probe. The analysis considers only those 28 probes that have at least 10 relevant tweets. The y-axis conveys the degree (of support, opposition, doubt and other, estimated using Equations (3.1) to (3.3)). Within each group the probes are ordered by the degree of support. First we observe that the ideal situation would be if the tweets convey 100% support in true probes and 100% opposition in false probes. For the debatable probes, we can expect a mixture of support, opposition and some degree of doubt via questioning. However, our data indicate deviations from this ideal. We find that though true probes are supported for the most part, there are exceptions. Not surprisingly, 'radiation causes cancer' generates almost equal support and opposition. This can be explained as due to radiation's dual nature as a cause and a treatment for certain cancers (as discussed before). But of note, there is still a non-trivial degree of doubt on the issue. Similar levels of doubt are expressed regarding both obesity and tanning as causes of cancer. Interestingly there is still some opposition that smoking causing cancer. We even found a tweet conveying that smoking treats cancer; perhaps the tweeter was being facetious.

More interesting is the data for our false probes. For the first four probes (from the left), support is close to 0.6 or greater! For the next two probes (lemon treats cancer and vaccine causes autism) the level of support exceeds the level of opposition. The situation is even more pronounced if we combine support and doubt and then compare with opposition. The last 3 false probes generate far more opposition than support, though we still have some non trivial doubt on these issues. Some observa-

tions are a bit alarming, as for instance the high level of support in the false notions that abortion causes breast cancer and aspartame causes cancer. The National Cancer Institute has debunked these theories and their site shows that these are false[9]. Other false notions generating high levels of support such as honey treating allergies may be viewed as comparatively less worrisome.

In the debatable probes group, we observe strong support in the first 4 probes including 'cell phone causes cancer'. Since this is a charged notion that has attracted much attention in the press, we expected more of an even split between support and opposition. However, the data indicate definite inclinations towards support amongst these users.

'Chocolate causes acne' obtained the highest level of doubt in our dataset. Perhaps this is partly due to the high participation of youth in Twitter. Sunscreen causing cancer raises the most opposition despite the fact that the probe is still debated. Overall a few probes such as cocaine causing glaucoma (T), avandia causes heart attack (D) and honey treats allergy (F) generate no opposition.

Finally we aggregate support, opposition and doubt using equations (3.4) to (3.6) (Table 3.5). We see again that true probes are mostly supported. Surprisingly the rate of support in false probes is 0.45 and the public is inclined to support (0.63) debatable probes as well. Opposition is highest in aggregate for false probes but so is doubt. If we assess deviations from the ideal, the level of opposition or doubt regarding true probes is small at 0.09, but support or doubt about false probes

[9]http://www.cancer.gov/cancertopics/factsheet/Risk/

is high at 0.67.

Table 3.5: Support, Opposition and Doubt Measures for Experiment 1.

	True Probes	**False Probes**	**Debatable Probes**
Support	**0.83**	0.45	0.63
Opposition	0.04	**0.27**	0.14
Doubt	0.05	**0.12**	0.11
Other	0.08	0.16	0.12

3.3.5 Conclusions

In this experiment we showed that for a set of handpicked true, false and debatable probes and using manual annotations we can study the levels of support, opposition and doubt. We did this using belief plots for individual probes as well as using aggregation methods for a set of probes. Overall we found high levels of support for not just true probes but also for false and debatable probes. There was considerable doubt for both true and false probes.

3.4 Experiment 2

3.4.1 Goals

In the previous experiment we used manual annotation to determine tweet relevance and tweet position with respect to given probes. In this experiment we address *RQ2: Can we build automatic classifiers for belief surveillance using off-*

the-shelf tools? We show that we can build belief classifiers using Weka [32]. We experiment with various classification and feature selection strategies.

3.4.2 Dataset

We use the same dataset that was used in Experiment 1 (Section 3.3). This dataset comprises of 2105 tweets that were manually labeled for relevance and position using crowdsourcing.

3.4.3 Methods

In line with the two-step decision process (box on left in Figure 3.1, Section 3.1) in our surveillance framework there are two classifiers we need to build: (a) a binary classifier for classifying retrieved tweets as relevant or not relevant to the probe and (b) a three-way classifier for classifying the position of relevant tweets in terms vis-à-vis the probe (supports, opposes, other[10]).

We keep our approach general and build probe independent classifiers. That is, we assume that probes regarding causes and treatments for illnesses have sufficient properties in common that may be capitalized by classifiers. We make our tweet collection independent by replacing the two key concepts with non-word constants. Since the probes are directional we replace the agent concept by AAAA and the target concept by BBBB. Therefore tweet TW2 in Table 3.1 is modified to 'Listen Bachman, you cannot get BBBB from a AAAA.' Similarly the tweet 'Now they say that cell

[10]As previously mentioned, note that 'doubt' is identified from the 'other' category by postprocessing

phones use can give you cancer' for the probe 'cell phones cause cancer' is modified to 'Now they say that AAAA use can give you BBBB'.

Table 3.6: Features & Algorithms Explored (C1= agent, C2= target).

Features	**Algorithms**
1. Unigram 2. Unigram + Bigram 3. Unigram + C1's position w.r.t. C2 + Words between C1 and C2 + # words in tweet + Hashtags yes/no + Hashtag text + URL present/absent + Re-tweet yes/no *(Combined)*	1. J48 Decision Tree 2. Naïve Bayes 3. SVM with PolyKernel 4. SVM with RBF Kernel 5. SVM with PolyKernel and Bagging 6. SVM with PolyKernel and Boosting

Although Weka offers a standard and well-reputed toolkit, every new problem requires some exploration of feature space and algorithms. Thus our first goal is exactly that, to find a reasonable set of features and algorithms. In all of our experiments, unless stated otherwise, we use a standard 10-fold cross-validation for training and testing classifiers. Table 3.6 lists the categories of features and algorithms explored using Weka. We note that for both classifiers we find that the addition of features more sophisticated than unigrams gives almost no benefit. We understand that there are many more features that may be explored such as type of source (drug company, medical organization etc.) and including these may give better results. So the classifier results we present may be viewed as a lower estimate on performance.

3.4.4 Results & Analysis

Tables 3.7 and 3.8 show the algorithms and feature sets explored for the relevance and position classifiers. In Table 3.7 we find that the highest F-score achieved is 0.83 using SVM with PolyKernel and Bagging and *Combined* features. However these results are not statistically significantly better when we use the same algorithm but with unigram features only (0.823). We decided to use the combination of SVM with PolyKernel and Bagging with unigram features for all our relevance classification problems because of its lesser complexity and comparable performance. Choosing a more general approach is also preferred because of lower chances of over-fitting of the training data.

We use similar arguments in choosing SVM with PolyKernel and Bagging with unigram features as the position classifier (F-score: 0.734). Although we find (in Table 3.8) better performance with more complex features sets but the differences in F-scores are not statistically significant. Here again we prefer a more general approach to prevent over-fitting.

Overall, these results are quite encouraging given that they are achieved using off-the-shelf tools with basic feature space and algorithm exploration. Several other tweet features of potential value such as type of Twitter user (health organization versus individual etc.) are yet to be exploited. This experiment provides preliminary evidence demonstrating the potential for using classifiers for belief surveillance.

Table 3.7: Algorithms and feature sets explored for Relevance Classifier.

Algorithm	**Features**	**Precision**	**Recall**	**F-Measure**
J48 Decision Tree	Unigram	0.767	0.769	0.768
Naïve Bayes		0.761	0.773	0.758
SVM with PolyKernel		0.82	0.823	0.821
SVM with RBF Kernel		0.789	0.747	0.677
SVM with PolyKernel and Bagging		**0.823**	**0.828**	**0.823**
SVM with PolyKernel and Boosting		0.797	0.8	0.798
J48 Decision Tree	Unigram + Bigram	0.767	0.769	0.768
Naïve Bayes		0.769	0.78	0.764
SVM with PolyKernel		0.819	0.823	0.82
SVM with RBF Kernel		0.81	0.769	0.714
SVM with PolyKernel and Bagging		0.828	0.832	0.826
SVM with PolyKernel and Boosting		0.814	0.819	0.815
J48 Decision Tree	*Combined*	0.759	0.763	0.761
Naïve Bayes		0.774	0.784	0.77
SVM with PolyKernel		0.826	0.829	0.827
SVM with RBF Kernel		0.809	0.77	0.716
SVM with PolyKernel and Bagging		0.831	0.835	0.83
SVM with PolyKernel and Boosting		0.81	0.815	0.811

Table 3.8: Algorithms and feature sets explored for Position Classifier.

Algorithm	**Features**	**Precision**	**Recall**	**F-Measure**
J48 Decision Tree	Unigram	0.686	0.707	0.69
Naive Bayes		0.667	0.655	0.659
SVM with PolyKernel		0.734	0.747	0.734
SVM with RBF Kernel		0.757	0.671	0.549
SVM with PolyKernel and Bagging		**0.737**	**0.751**	**0.734**
SVM with PolyKernel and Boosting		0.717	0.724	0.717
J48 Decision Tree	Unigram + Bigram	0.686	0.709	0.69
Naive Bayes		0.691	0.672	0.678
SVM with PolyKernel		0.761	0.765	0.738
SVM with RBF Kernel		0.759	0.717	0.643
SVM with PolyKernel and Bagging		0.775	0.77	0.741
SVM with PolyKernel and Boosting		0.761	0.765	0.738
J48 Decision Tree	*Combined*	0.687	0.708	0.691
Naive Bayes		0.694	0.672	0.679
SVM with PolyKernel		0.767	0.768	0.741
SVM with RBF Kernel		0.768	0.722	0.653
SVM with PolyKernel and Bagging		0.775	0.768	0.737
SVM with PolyKernel and Boosting		0.767	0.768	0.741

3.4.5 Conclusions

In this experiment we show that we can use off-the-shelf machine learning tools to classify tweets into relevant and non-relevant categories and further classify the relevant tweets into one of the three categories: support, opposition and doubt. The results are encouraging for both these classifiers.

3.5 Experiment 3

3.5.1 Goals

In this experiment we address *RQ3: Can we use our belief classifiers trained on one set of probes to classify tweets for new probes?* In essence we want to test the generalizability of the relevance and position classifiers that we built in Experiment 2 (Section 3.4). Given the wide-ranging nature of health beliefs it will not be possible to get human annotations that cover all of them. So the ability of our classifiers to accommodate tweets from new probes becomes important. Motivated by transfer learning for classification we explore two approaches for building classifiers – one using a leave-one-probe-out strategy and the other using probe clustering followed by the leave-one-probe-out approach.

3.5.2 Dataset

We begin by extending the gold standard dataset from Experiment 1 (Section 3.3) which has 2105 annotated tweets labeled for relevance and position.

3.5.3 Methods

3.5.3.1 Extending Dataset and Classifiers

We extend the dataset from Experiment 1 (*gold standard dataset*) by copying over the annotations from tweets to those that are very similar but not annotated. We do this because in Experiment 1 (Section 3.3) we removed duplicates and near duplicates of tweets that were being labeled to increase diversity of annotated dataset. To compensate for that, in this experiment we copy over the annotations from labeled tweets to those that are similar. We use a high cosine similarity threshold of 0.8 (determined by running several similarity experiments for the optimal threshold). This set is the basis of all remaining classifier experiments. We refer to this dataset, comprised of 7957 tweets, as *ExtAnnotatedDataset*.

Before using this annotated data for classifying new data or testing their generalizability we implemented a standard 10-fold classification experiment using all of our 7957 labeled tweets (*ExtAnnotatedDataset*). We explore various training algorithms built in Weka with different feature vectors extracted from tweets' content (Section 3.4) ranging from simple unigram to richer Twitter-based features like hashtags, URLs, etc.

3.5.3.2 Generalizability of Classifiers

Using a Leave-one-probe-out Approach: We use the 15 probes in *ExtAnnotatedDataset* with at least 100 labeled tweets. Treating one probe as new we use its tweets as test data. These are classified with classifiers built from the tweets annotated for

the remaining 14 probes. These probes are marked by asterisk in Table 3.9.

Table 3.9: Probe Clusters.

Treats Cluster	**Therapeutic Drug Cluster**	**Food/Drink/Recreational Drug Related Cluster**	**Life/Nature Cluster**
Lemon treats cancer	DES causes cancer*	Alcohol causes cancer*	Cell phone causes cancer*
Honey treats allergy	Vaccine causes autism*	Cantaloupe causes listeria*	Stress causes cancer*
Garlic treats blood pressure	Zocor causes muscle injury	Smoking causes cancer*	Obesity causes cancer*
Vinegar treats blood pressure	Aspirin causes bleeding	Coffee causes cancer*	Radiation causes cancer*
	Avandia causes heart attack	Chocolate causes acne	Asbestos causes cancer*
	Drospirenone causes blood clot	Tea causes cancer*	Tanning causes cancer*
	Actos causes bladder cancer	Plastic causes cancer*	Sunscreen causes cancer
	Anesthesia causes learning disabilities*	Milk causes mucous	Cracking knuckles causes arthritis
		Aspartame causes cancer	Radon cancer causes
		Cocaine causes glaucoma	Abortion causes breast cancer

Using A Clustered Leave-one-probe-out Approach: The idea is that we hone in on the most related probes for our training data. In essence we cluster the probes and then when classifying tweets of a probe we limit the training data to other probes from the same cluster. For this we manually place the 32 probes (Section 3.4) into four clusters shown in Table 3.9. The probes used in this study are marked with an asterisk. The manual clustering in effect imposes a ceiling on performance as we could expect automatic clustering to yield up to the same results. The 'treats cluster' consists of probes containing term pairs related by the 'treats' verb. The 'therapeutic drug cluster' consists of probes containing therapeutic drugs or other drug-related terms. The 'food/drink/recreational drug-related cluster' consists of terms referring to food, drinks and recreational drug usage. The 'life/nature cluster' consists of terms that are related to our daily life or life style choices and natural products. In some

cases we made exceptions based on the nature of the tweets retrieved for a certain probe. For example, while it can be argued that 'plastic causes cancer' should be considered in the 'life/nature cluster', we noticed that most of the tweets for this probe conveyed the idea that plastic food containers cause cancer. Because of the nature of it's similarity with other tweets of the 'food/drink/recreational drug-related cluster' it was placed there. This experiment design is identical to the leave one out approach, except that when classifying tweets of a probe, the training data used is limited to the probe's cluster.

3.5.4 Results & Analysis

3.5.4.1 Extending Dataset and Classifiers

The *ExtAnnotatedDataset* contains 7957 labeled tweets out of which of 6088 were relevant (76.5%).

Table 3.10: Performance of Classifiers (Wt. Avg.: Weighted Average).

Relevance Classifier				**Position Classifier**			
TP rate	FP rate	F-Score	Class	TP rate	FP rate	F-Score	Class
0.982	0.113	0.974	Relevant	0.986	0.072	0.972	Supports
0.887	0.018	0.912	Non-relevant	0.944	0.005	0.964	Opposes
0.960	0.091	0.959	Wt. Avg.	0.860	0.011	0.886	Other
				0.960	0.048	0.960	Wt. Avg.

Our results show that SVMs with Polykernel and Bagging performs best for both relevance and position. Results presented in Table 3.10 show excellent perfor-

mance in F-score for both classifiers. The position classifier's ability to classify tweets that take no position (i.e. 'other') vis-à-vis the probe seems to be the most challenging task. Compared to the Results of Experiment 2 (Section 3.4), we find that both relevance and position classifiers of the extended dataset outperform those of original dataset. The position classifier receives a higher boost in performance (0.960 vs 0.734) compared to the relevance classifier (0.959 vs 0.823) using the extended dataset. We remind the reader that these experiments were strictly 10 fold over the union of all tweets for all probes in the dataset.

3.5.4.2 Generalizability of Classifiers

The first two numerical columns in Table 3.11 present our results for the leave-one-probe-out approach. We note that deciding relevance is relatively easier than deciding tweet position. The results are overall quite encouraging. We find that except for the first probe, the F-scores are between 0.553 and 0.893. For tweet position, the F-score is below 0.5 for five of the fifteen probes. Average F-scores are 0.71 for relevance and 0.62 for position.

The results for the clustered leave-one-probe-out approach are presented in the second set of columns of Table 3.11. As with the previous approach, the relevance classifiers perform better than the position classifiers. Except a couple of probes (*cell phone causes cancer* and *coffee causes cancer*), the F-scores are between 0.554 and 0.96. For tweet position, the F-score is below 0.5 for 6 of the 15 probes. The average F-scores for cluster classifiers are 0.69 for relevance and 0.57 for position.

Table 3.11: Generalizing to New Probes: Comparing Two Strategies.

Probe	**Leave-one-probe-out Classifiers**		**Cluster Classifiers**	
	Relevance F-Score:	**Position F-Score:**	**Relevance F-Score:**	**Position F-Score:**
cell phone causes cancer	0.414	0.674	0.442	0.625
smoking causes cancer	0.76	0.399	0.726	0.392
vaccine causes autism	0.698	0.339	0.7	0.14
obesity causes cancer	0.556	0.711	0.609	0.717
plastic causes cancer	0.735	0.768	0.638	0.674
radiation causes cancer	0.674	0.4	0.65	0.261
stress causes cancer	0.553	0.74	0.554	0.778
alcohol causes cancer	0.753	0.8	0.658	0.783
tanning causes cancer	0.85	0.541	0.868	0.476
coffee causes cancer	0.608	0.229	0.476	0.117
tea causes cancer	0.801	0.255	0.814	0.208
asbestos causes cancer	0.814	0.727	0.777	0.687
cantaloupe causes listeria	0.819	0.908	0.799	0.956
DES causes cancer	0.719	0.929	0.684	0.891
anesthesia causes learning disabilities	0.893	0.835	0.96	0.88
Average F-Score	0.71	0.617	0.69	0.572

We perform tests of significance (paired t-test) to compare the results from the two approaches. The difference in *relevance* scores is insignificant ($p>0.01$) while the difference is significant for *position* ($p<0.01$) scores. We conclude that clustered leave-one-probe-out based classification approach does not work well for our dataset. One likely reason for this is the relatively fewer number of labeled tweets in the training data for some of the clusters.

We compare these generalizability results with the 10-fold cross-validation results of the *ExtAnnotatedDataset*, where there is no notion of classifying a new probe. As expected we find that results for the 10-fold cross validation are much better than the generalizability classifiers for both relevance and position.

3.5.5 Conclusions

In this experiment we tested the generalizability of our classifiers using two approaches – a leave-one-probe-out approach and a clustered leave-one-probe-out approach. In both these approaches we find that our classifiers perform well, independent of the training probes and are quite generalizable. However, not surprisingly, the results are not as strong as where the classifier has tweets related to the probe of interest in its training data.

3.6 Experiment 4

3.6.1 Goals

In all the previous experiments we used a set of pre-defined probes for studying beliefs. In this experiment we address *RQ4: Can we identify naturally expressed beliefs in a population of social media users?* We use the probe harvesting technique outlined in Section 3.1 to achieve this task. Similar to the previous experiments we study the levels of support, opposition and doubt for these new probes and also aggregate them using techniques shown in Section 3.1.

Table 3.12: Health Related Hashtags.

Hashtags			
#disease	#medicine	#doctor	#patient
#doctors	#patients	#health	#pharma
#healthcare	#pharmacy	#hospital	#physician
#hospitals	#physicians	#medical	#therapy

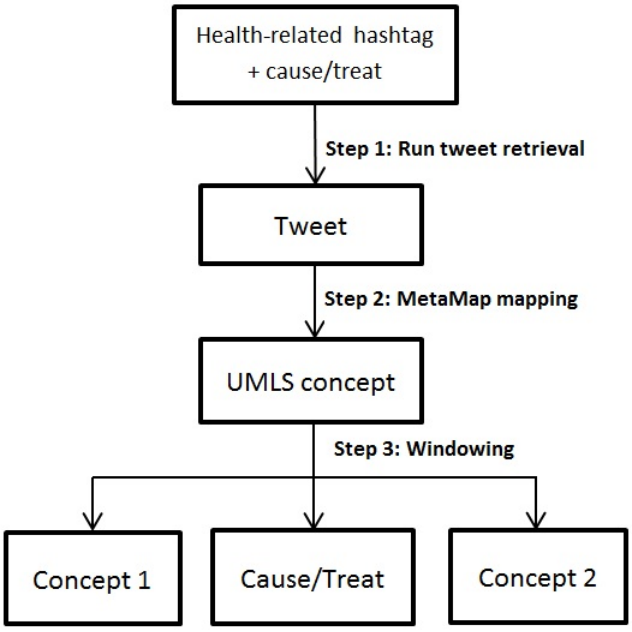


Figure 3.4: Flowchart of Health-related Tweet Retrieval and Concept Extraction (example of Concepts 1 and 2 may be 'smoking' and 'cancer')

3.6.2 Dataset

We propose two approaches to identify discussions related to causes and treatments of illnesses in Twitter. In the first approach we want a general strategy to identify associations of drugs, diseases and other entities with each other using a set of hashtags that capture broader health-related tweets. In the second approach we want to be more specific by identifying drug interactions with specific diseases and vice-versa.

3.6.2.1 Hashtag Dataset

We identify a set of 16 general health-related hashtags (Table 3.12) from Fox's ePractice Healthcare Hashtag Project[11]. We do not select country (e.g. #cdnhealth), organization (e.g. #FDA) or technology-specific (e.g. #HealthIT) hashtags. We also choose not to explore disease-specific (e.g. #hepatitis) hashtags; though relevant we are leaving these for future research. Each hashtag is then coupled with cause/causes and treat/treats verbs as search terms. These pairs are searched using the Twitter Search API (as in previous work). The *HashtagDataset* was built on October 13, 2011 and contains 1,313 non-unique tweets. After removal of user mentions (@ prefixed), retweet mentions (@RT) and URLs we get 613 unique tweets.

Table 3.13: Drugs and Diseases Explored.

Top 10 DTC Drugs	Lipitor, Cialis, Advair, Abilify, Cymbalta Symbicort, Pristiq, Plavix, Chantix, Lyrica
OTC Drugs	Aspirin, Advil, Prilosec, Centrum, Robitussin Tylenol, Nyquil, Dramamine, Zantac, Benadryl
Chronic Diseases	Diabetes, Asthma, Arthritis, Schizophrenia, Cardiac failure Glaucoma, Haemophilia, Hypertension, Multiple sclerosis, Parkinson's disease, Osteoporosis, Psoriasis, Obesity, Epilepsy
Infectious Diseases	HIV/AIDS, Dengue, Malaria, Anthrax, Cholera Bubonic plague, Influenza, Typhoid, Smallpox, Pneumonia Tuberculosis, Yellow fever, Bird flu, Ebola, Leprosy, Hepatitis

[11]http://www.foxepractice.com/healthcare-hashtags/

3.6.2.2 Drug and Disease Dataset

Here we select a list of drugs and diseases (*DrugDiseaseDataset*) (Table 3.13). For drugs, we selected the top 10 most Direct-to-Consumer (DTC) advertised drugs[12] for heart diseases, neuropathic pain, etc. We also selected 10 of over-the-counter (OTC) or generic drugs[13] that are less advertised but frequently used for common problems like fever, pain, heartburn, etc. For diseases we selected chronic[14] and infectious diseases[15] from the World health Organization's (WHO) fact sheets for such diseases (Table 3.13). As with the hashtags, we combined each drug and disease term from Table 3.13 with relationship terms cause/causes and treat/treats for search on Twitter. This was done on February 24, 2012 using the Twitter Search API. Similar to the previous approach, here also we removed URLs, user mentions and re-tweet mentions and took the unique instances of the remaining tweets. For the drugs and diseases this resulted in 942 and 3722 unique tweets respectively.

3.6.3 Methods

We process the 613 tweets of the *HashtagDataset* using the procedure outlined in Figure 3.4. After processing each tweet with National Library of Medicine's MetaMap [3] program we extract pairs of concepts belonging to key semantic types ('Disease or Syndrome', 'Finding', 'Pharmacologic Substance', 'Manufactured Ob-

[12]http://gaia.adage.com/images/bin/pdf/WPpharmmarketing_revise.pdf

[13]http://www.uihealthcare.com/pharmacy/OTCmedications.html

[14]http://www.who.int/topics/chronic_diseases/factsheets/en/

[15]http://www.who.int/topics/infectious_diseases/factsheets/en

ject', etc.) appearing within a specific window size of 4 words[16]. These are the types that represent drugs and diseases.

The *DrugDiseaseDataset* was processed using a similar approach (Figure 3.4). However given the considerably large number of tweets retrieved using this approach and the potential variability in tweet expressions in this set of tweets, we considered a wider array of semantic type to fetch the probes. Similar to the hashtag-based approach, each tweet was examined for the presence of drug and disease-specific semantic types within a tweet. Such tweets are expected to portray drug-disease relationships being discussed. Additionally, we also consider certain drug-unrelated semantic types which may have important association with diseases (e.g. [Food], [Mammal], etc.).

3.6.4 Results & Analysis

3.6.4.1 Results for *HashtagDataset*

The *HashtagDataset* results in 49 new concept pairs linked by 'cause/causes' or 'treat/treats'. We now conduct surveillance on these probes as described in Section 3.1)[17]. Table 3.14 lists the 17 pairs that retrieved at least 10 tweets and that did not appear in our initial probe set from Experiment 1 (Section 3.3). So these are the probes we have discovered in our data. They represent the conversations (of at least 10 tweets) around causes and treatments that were occurring in Twitter. Each

[16]This parameter was tuned on a training set of 30 MetaMapped tweets

[17]This was done on November 1, 2011; retrieved tweets dated back to the previous 7 days as per the API.

Table 3.14: *HashtagDataset* mined probes (T: True; F: False; D: Debatable).

Mined Probes	#Tweets	Truth Status
smoking causes death	1181	T
skin product causes aging	673	F
chemotherapy treats breast cancer	392	T
oral sex causes throat cancer	272	T
marijuana treats PTSD	235	D
smokeless tobacco causes cancer	150	T
antidepressant causes depression	149	T
stress causes sickness	129	T
medication causes hair loss	84	T
milk causes acne	52	T
milk causes osteoporosis	29	F
magic mushroom causes personality change	23	D
nasal polyp causes nasal block	17	T
tea tree oil treats infection	17	D
cialis treats enlarged prostrate	16	T
diet causes bad breath	16	T
listeria causes miscarriage	14	T

probe was judged as either true, false or debatable by two physicians. Disagreements in judgments were resolved by consulting a third physician. The identified probes are used for belief surveillance.

As part of the surveillance we utilize our two general classifiers (SVM with PolyKernel and Bagging for relevance and for position, developed using *ExtAnnotatedDataset* in Experiment 3 (Section 3.5)). We classified the tweets retrieved for the 17 probes to produce the belief chart shown in Figure 3.5. The chart is clustered into 3 groups; the leftmost represents true probes (T), the rightmost represents false probes (F) and the middle probes are debatable (D). The y-axis conveys the degree of support, opposition, doubt and other (estimated using Equation (3.1) and its analogues from Section 3.1). Within each group the probes are ordered by the degree of support.

We see high levels of support for false and debatable probes. For example, there is almost no support to the false notion that *skin products cause aging* and little opposition to the debatable notion that *tea tree oil treats infection*. Aggregating support, opposition and doubt using Equations (3.1) and its analogues (Section 3.1) gives us some very alarming results (Table 3.15). Not surprisingly, there is high support (0.80) for true probes. However, there is almost equal support for both false and debatable probes. Opposition is surprisingly low for false (0.05) probes where the deviation from ideal is 0.84. These results are even more striking compared to the findings from Experiment 1 (Section 3.3). There is considerably less opposition and doubt for both false and debatable probes. The differences in the design of the two

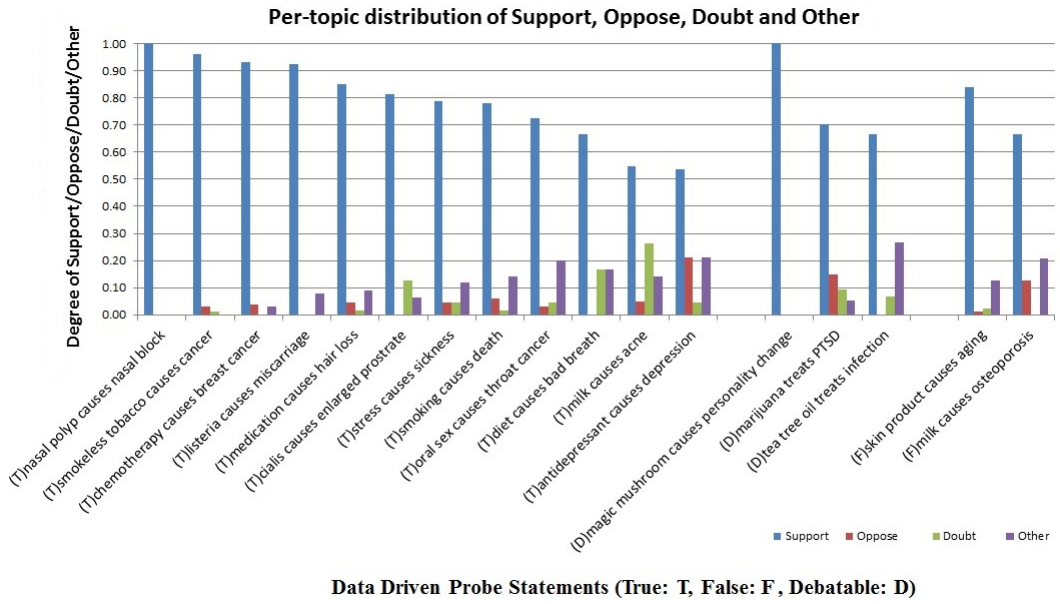


Figure 3.5: Belief Plot for *HashtagDataset*

experiments is that the probes considered here are naturally expressed as opposed to pre-defined probes in Experiment 1. This possibly leads to differences in results of two experiments. Additionally, the number of probes in the true, false and debatable categories differ between these experiments. The results of this experiment likely reflects the more common scenario since we considered naturally expressed probes here. In a later study (Experiment 7, Section 3.9) we look at more harvested probes.

3.6.4.2 Results for *DrugDiseaseDataset*

We identified 361 and 978 term pairs from the drug and disease sets respectively. Manual inspection of these pairs revealed certain anomalies in the identified term pairs. For example, for the tweet "The girl in my class is giving a speech and said weed causes schizophrenia", MetaMap identifies the verb 'said' as 'said (Simian

Table 3.15: Support, Opposition and Doubt Measures for *HashtagDataset*.

	True Probes	**False Probes**	**Debatable Probes**
Support	**0.80**	0.79	0.75
Opposition	0.04	0.05	**0.07**
Doubt	**0.06**	0.05	0.01
Other	0.10	0.11	0.17

Acquired Immunodeficiency Syndrome) [Disease or Syndrome]'. Other examples of frequently appearing but incorrectly mapped terms are 'I' (identified as 'I NOS (Blood group antibody I)'), dnt (abbreviated don't) (identified as 'DNT (Dysembryoplastic neuroepithelial tumor)'), etc. On the other hand, we also identified some very interesting probes such as 'armadillos[18] cause leprosy' because of the use of wider range of semantic types. After manually filtering out pairs containing incorrectly mapped terms we got 209 and 556 pairs from the drugs and diseases datasets respectively. Table 3.16 summarizes these dataset characteristics. We note that disease-related search terms retrieve more tweets than drug-related terms. Consequently, the number of probes mined from disease-related tweets is also greater than that of drug-related tweets dataset.

Here we follow the same retrieval strategy shown in the box on the left in Figure 3.1 (Section 3.1). We collected 88048 tweets for over 700 filtered probes (Ta-

[18]Armadillos (*Armadillo officinalis*) are mammals primarily found in Central and South America.

Table 3.16: Drugs and Diseases Datasets.

Dataset: *Drugs*	
Number of tweets retrieved for Disease set	1226
Number of unique tweets	942
Number of pairs (before filtering)	361
Number of pairs (after filtering)	209
Dataset: *Diseases*	
Number of tweets retrieved for Drugs set	8679
Number of unique tweets	3722
Number of pairs (before filtering)	978
Number of pairs (after filtering)	556

Table 3.17: Select probes from *DrugDiseaseDataset* (truth status: True: T, False: F, Debatable: D).

Harvested Probes	
Advil treats hangover (T)	Ginkgo treats diabetes (F)
Viagra treats anxiety (F)	Viagra treats hypertension (D)
Advil causes stomach bleeding (T)	Marijuana causes schizophrenia (T)
Armadillos causes leprosy (T)	Methotrexate treats cancer (T)
Video causes seizure (T)	Water treats hangover (T)
Benadryl causes itching (F)	Nigella sativa treats diabetes (D)
Bilberry treats diabetes (D)	Nyquil causes coma (F)
Weed treats AIDS (D)	Weed treats asthma (F)
BPA causes obesity (D)	Overeating causes memory loss (D)
Cannabis treats bronchitis (F)	Seroquel causes diabetes (D)
Weed treats cancer (D)	Weed treats depression (D)
Cialis treats impotency (T)	Stress causes schizophrenia (T)
Coffee causes diabetes (F)	Viagra causes hearing loss (D)
Weed treats glaucoma (D)	

ble 3.16)[19]. 117 probes of the drugs dataset and 324 probes of the diseases dataset retrieved more than 10 tweets. We selected a subset of 27 probes based on interestingness to analyze further. Table 3.17 shows the selected probes that retrieved at least 10 relevant tweets. As mentioned earlier, note that the mined probes may involve entities that are not drugs or diseases (e.g. water, coffee, weed, armadillos, etc.). For each probe in our subset a physician provided decisions on whether the probe is true, false or debatable. One-third of the probes pertain to effects/side-effects of therapeutic drugs. Quite a few refer to recreational drugs. There were a number of probes in alternative medicine (herbal therapy, homeopathy, etc.) and dietary substances with causal or curative relationships with diseases. Not surprisingly, we find a large number of probes relate to recent studies with animal models that might have generated a buzz. Naturally probes mined depend upon current events and developments. This is because social media often correlates to current events in news media or even pop-culture[20].

Figure 3.6 shows the plots of support, opposition, doubt, and other for the true, false and debatable probes (Table 3.17) having at least 10 relevant tweets. Here we notice that there is surprisingly low support for the true probe *Cialis treats impotency* which is a known prescribed medication for impotency. For this probe there is a high proportion of tweets in the 'Other' category. This is perhaps because drugs of this category (i.e. sexual health) are highly targeted by spammers on social media

[19]The Twitter search was performed on February 28, 2011

[20]Several tweets related to the probe "video causes seizure" refer to a popular music video that might cause epileptic seizure and contains a related disclaimer

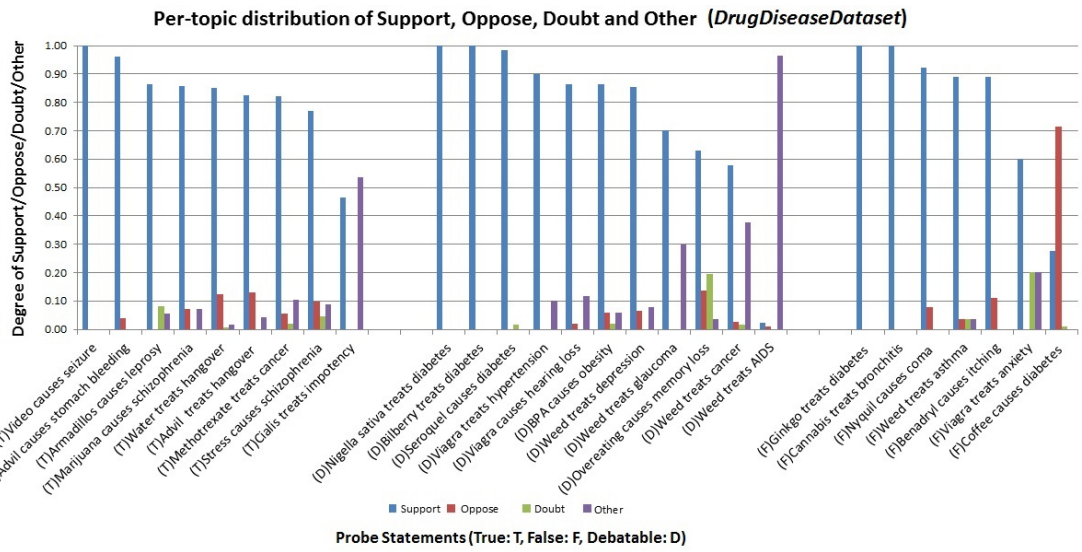


Figure 3.6: Belief Plot for *DrugDiseaseDataset*

and these spam tweets are captured in the 'Other' category. Additionally there is a high level of support in debatable probes, especially those pertaining to alternative medicines and recreational drugs (e.g. weed). Similar to the probe *Cialis treats impotency* we also notice that a significant proportion of tweets retrieved for probes like *weed treats glaucoma* contain large number of spam tweets (advertisements) which are classified in the 'other' category. These were apparently quite challenging for our relevance classifier. We see high support in false probes related to alternative medicine, recreational drugs and even therapeutic drugs. For example, tweets like "*smokin weed helps people wit asthma #fact*" emphasize such support for false probes.

The aggregate calculations of support, opposition, doubt, and other are shown in Table 3.18. Similar to the belief plot, we see high support (0.82) for true probes. Moreover, there is almost equally high support (0.80) for false probes and debatable

Table 3.18: Support, Opposition and Doubt Measures for *DrugDiseaseDataset*.

	True Probes	**False Probes**	**Debatable Probes**
Support	**0.82**	0.80	0.76
Opposition	0.06	**0.13**	0.03
Doubt	0.02	**0.04**	0.02
Other	0.10	0.03	0.19

probes (0.76). We also note that the level of support for false probes in the harvested set is actually almost double that of the pre-defined probes (0.49) of Experiment 1 (Section 3.3). We speculate that this difference is probably because of the bias involved in selecting the pre-defined probes and the lack of it in harvesting naturally expressed probes. We find very low opposition (only 0.13) to false probes. This is alarming and emphasizes the need to embark on public health campaigns to correct such misinformed beliefs.

3.6.5 Conclusions

In this experiment we show that we can harvest naturally expressed health beliefs from Twitter. The findings for the automatically mined probes mirror our earlier findings using the 32 probes in several aspects. With both the *HashtagDataset* and *DrugDiseaseDataset* we have shown that support for false and debatable probes is quite high and sometimes even comparable to support for true probes. However, the levels of opposition for false and debatable probes is much lower for the mined tweets compared to the 32 probes from Experiment 1 (Section 3.3). The low levels

of opposition for certain probes like *vaccine causes autism* [10] may be deemed more worrisome compared to *weed treats asthma* based on the reach and impact of such beliefs. Similarly, probes that generate a lot of doubt in the form of questioning such as *chocolate causes acne* or *Viagra treats anxiety* may be identified as equally important for disseminating public health knowledge in the population.

3.7 Experiment 5

3.7.1 Goals

In this experiment we address *RQ5: Can we estimate the scientific novelty of a probe identified from social media conversations?* Here our primary goal is to investigate methods for estimating novelty of probes harvested using Twitter.

Novelty refers to whether the idea underlying the probe has already been explored in the scientific research. One may of course adopt a broader perspective and assess if the idea has been even discussed or utilized outside of the scientific arena. We take an initial step here by exploring the presence of these probes in MEDLINE through a PubMed search. Note that the probe's absence (or low presence) in MEDLINE does not necessarily indicate a reasonable or interesting hypothesis. But yet it is a start towards hypothesis discovery. One decision we can be confident about is that if a probe has an 'appreciable' MEDLINE footprint then one must remove it from further consideration. This effort caters to the broader research area of Literature-based Discovery (or LBD), which has operated largely off bibliographic data as in PubMed or full-text collections as for example PubMed Central.

3.7.2 Dataset

We use the *DrugDiseaseDataset* from Experiment 4 (Section 3.6). This dataset comprises of over 700 probes, few of which are selected for studying in this experiment.

3.7.3 Methods

We conduct our PubMed search using the following strategy. First we search the PubMed title/abstract fields using the concept terms limited to publications dated prior to Feb 28, 2012[21]. For example the PubMed search for a probe such as 'aspirin treats poor leg circulation' is **(aspirin [Title/Abstract] AND poor leg circulation [Title/Abstract]) AND ("1800" [Date - Publication] : "2012/02/28" [Date - Publication])**. If we find multiple hits with this strategy then we add the relationship term 'treats[Title/Abstract]' to the search query. On the other hand, if we do not find any hits for the previous search strategy, we relax the search query to a simple keyword search using the concepts which helps us identify possible synonyms for the concepts. These are then replaced in the original query and search is again executed. It is important to note here that consumer vocabulary (as mined in probes) is significantly different from standardized or scientific vocabulary [89]. While the PubMed search algorithm implicitly does query expansion to include standardized or scientific terms for common terms, it is not comprehensive. The following search results are limited to the use of common terms used in the probes and alternative scientific terms suggested by PubMed.

[21]This date was chosen as all probes were harvested by this date.

3.7.4 Results & Analysis

We manually analyzed a sample of two groups of probes (relationships) based on the number of PubMed documents retrieved. In the first category, *Probes with Sparse PubMed Support*, we have probes that retrieved low to moderate number of PubMed records. In the second category, *Probes with No Explicit PubMed Support*, we have probes for which we could not find any PubMed records. When we did find records, we also took a look at them to see if the relationship expressed in the probe was being discussed or if it was a false positive retrieval.

3.7.4.1 Probes with Sparse PubMed Support

- Curcumin treats multiple sclerosis (MS):

  Sample Tweet: "Curcumin has bright prospects for the treatment of multiple sclerosis [URL]"

  While we find only few tweets retrieved for this probe, 10 articles on this probe are retrieved using PubMed search. Adding the relationship term 'treat' however results in only one article. A manual analysis of the abstracts of these articles reveal various indications of its benefits – from discussion of its efficacy in animal models to its anti-inflammatory properties in specific scenarios – without any concrete evidence of its use in curing MS in humans.

- Cilantro treats diabetes:

  Sample Tweet: "Apparently cilantro is used to treat diabetes...well I hope to god I don't get it cause I can't stand cilantro."

Only the above tweet was retrieved for this probe. While a PubMed search of the exact concept term pairs did not return any results, replacing cilantro with its scientific name coriandrum resulted in 10 hits. These articles covered various topics including other traditional plant treatments for diabetes to its efficacy in animal models.

- Coconut oil treats psoriasis:

  Sample Tweet: "RT @Psoriasisclub: Coconut Oil: a fantastic natural moisturiser for any dry skin and especially helpful for psoriasis. [URL]"

  While 13 tweets referred to this probe, only 2 studies could be found using PubMed search. One of the studies found no significant benefits in using coconut oil for psoriasis clearance, while the other discusses the process of a drug preparation ("77 oil") which uses coconut oil as a base and used in the treatment of psoriasis.

- Cialis causes hearing loss:

  Sample Tweet: "RT @Iamsuperbrad: One side effect of Cialis can be hearing loss. [expletive satire] It's every man's dream in pill form."

  38 tweets were retrieved supporting this probe. A PubMed search on this probe (using generic name Tadalafil) resulted in 2 retrieved records both indicating hearing loss due to various Phosphodiesterase type 5 inhibitors.

- Cialis treats high blood pressure:

  Sample Tweet: "cialis treat high blood pressure [URL]"

A large number of tweets (97) supporting this particular probe was mined from Twitter. Using our strict PubMed search strategy we did not find any evidence of this association. However using the relaxed search strategy we found several instances where Cialis (generic name Tadalafil) is used as a treatment for pulmonary arterial hypertension.

- Krill oil treats Rheumatoid Arthritis (RA):

  Sample Tweet: "krill Oil Supplements Can Treat the Symptoms of Rheumatoid Arthritis [URL]"

  We found around 10 tweets discussing this probe. A PubMed search of the probe returns two records. One of these records, an older study found "Neptune Krill Oil (NKO)" to be beneficial for RA, while a more recent study from 2010 demonstrates its efficacy in animal models.

- Bergamot treats psoriasis:

  Sample Tweet: "Fischer-Rizzi sugests blending bergamot with rock rose and everlasting to treat eczema and psoriasis. #aromatherapy #skincare"

  This probe results in only 3 PubMed hits. A manual inspection of the PubMed records reveal direct or indirect relationship between bergamot (specifically bergamot oil) and psoriasis. A search of bergamot and eczema using the strict or relaxed PubMed search shows no results.

- Cialis causes muscle pain:

  Sample Tweet: "cialis side effects muscle pain [URL]"

16 tweets were retrieved for this probe. However, a PubMed search with the generic name Tadalafil resulted in only 2 hits, both related to adverse effect for this drug.

- Benadryl causes hallucinations:

  Sample Tweet: "OMG!!!! benadryl causes hallucination! #hallucinate"

  This probe mined from Twitter fetches only 2 PubMed records when searched as-is. However, using 'diphenhydramine', the generic name for Benadryl, we get 19 hits. Appending the relationship term 'cause' to the search results in only 5 hits.

3.7.4.2 Probes with No PubMed Support

- Lavender oil treats acne/psoriasis:

  Sample Tweet: "#Natural #Health: Lavender oil has been used for centuries to treat acne, wrinkles, psoriasis + skin irritants [URL] #beauty"

  While quite a few tweets (8) were found relating lavender oil to treatment for acne and psoriasis, we did not find any PubMed records supporting such claims. However, 'wrinkles', which is also mentioned in the same tweet, retrieves one PubMed record when associated with Lavender oil. Manual inspection of the article reveals the use of lavender oil aroma for easing anxiety of patients undergoing BOTOX treatment for wrinkled skin.

- Triphala treats obesity:

  Sample Tweet: "Triphala treats obesity miraculously. As triphala regularizes

the functioning of our digestive system, it directly reduces body fat."

While a single tweet referred to this particular probe, PubMed search of triphala (an Ayurvedic medicine comprised of three myrobalans) and its treatment potential for obesity did not return any results.

- Clove oil treats colds/bronchitis/asthma/tuberculosis:

  Sample Tweet: "Clove leaf oil is also clearing nasal passage & treat colds, bronchitis, asthma, and tuberculosis."

  While we find evidence of this probe in Twitter discussions, PubMed searches of the association of clove oil with any of the diseases or symptoms do not return any result.

- Neem treats psoriasis:

  Sample Tweet: "Using Neem to Treat Psoriasis — 21st Century Apothecary [URL]".

  We did not find any documents relating neem *(Azadirachta indica)* with psoriasis.

- Lyrica causes hair loss:

  Sample Tweet: "@CraigHeff Lyrica, Topamax, Lamictal are all used for neuropathic pain releif. Side effects are, suicidal thoughts, memory and hair loss."

  While we find sparse evidences of association of Lyrica (generic Pregabalin) with suicidal thoughts in PubMed (13 tweets), there is no evidence of the adverse effect of hair-loss in association with Lyrica in PubMed search (2 tweets).

- Cialis causes heartburn:

  Sample Tweet: "why does cialis cause heartburn [URL]".

  41 tweets reporting this side-effect were found in our dataset. However a PubMed search of this probe using both the brand name and the generic name of the drug returned no results.

- Ginkgo treats bronchitis:

  Sample Tweet: "[URL] Ginkgo leaves and seeds are utilized to treat asthma, bronchitis, allergies, cardiac arrhythmia and to improve memory".

  A PubMed search of the various treatment related probes of Ginkgo, namely for treatment of asthma or allergy or cardiac arrhythmia , return at least a few articles. However no article could be found on the efficacy of Ginkgo for bronchitis.

- Rosehip oil treats acne/eczema/psoriasis:

  Sample Tweet: "Rosehip oil used to treat stretch marks, burns, sunburn surgery scars, acne, eczema, psoriasis [URL]".

  No evidence of associations between rosehip oil and any of the skin conditions listed was found in PubMed.

3.7.5 Conclusions

In conclusion, we find that several probes mined from Twitter are either not present or sparsely present in PubMed. Note this statement is made within the constraints of our search strategy. At first glance these probes representing proto-ideas

have some potential towards developing new hypotheses for scientific research. However, these need further validation especially regarding reasonableness or rationale and this validation may involve downstream text mining processes. Social media being the source underlines this need. In fact, a natural strategy, which we propose for the future, is to put the mined probes through a closed discovery process [74] to extract any underlying rationale (for instance between Triphala and obesity). Overall we show that, our method exploiting the semantics of concepts is capable of mining specific types of relationships from Twitter discussions that could feed into a more general LBD process.

3.8 Experiment 6

3.8.1 Goals

In this experiment we address *RQ6: Does the level of discussion or the level of support, opposition or doubt for probes change over time?* One of the limitations of our previous experiments is the dataset which spanned only one week. Here our goal is to conduct a surveillance study using data collected at two different time spans. To address this we conduct the same 'Probe Surveillance' technique (Section 3.1) for the 32 probes (studied in Experiment 1) at two different time spans. From this study we hope to estimate whether the levels of discussion for particular probes change over time, or the direction of positions change or if they stay the same over time. If there are changes, we also wish to investigate factors that may be influencing such changes.

3.8.2 Dataset

We use the same set of 32 predefined probes (Table 3.2) as in Experiment 1 (Section 3.3). Here we retrieve tweets for a second week and compare the levels of support, opposition and doubt. The selected weeks are separated by approximately 4 months and are referred to as the *Oct11Dataset* (11,591 tweets) and *Feb12Dataset*. The *Feb12Dataset* dataset was collected on Feb 22, 2012 and comprises of 9665 tweets. Retrieval details are in Section 3.1.

3.8.3 Methods

For the new *Feb12Dataset* we use the relevance and position classifiers (*ExtAnnotatedDataset*) from Experiment 3 (Section 3.5) to classify new tweets.

3.8.4 Results & Analysis

It is interesting to see how the support, opposition, and doubt for the same probes change over time. Figure 3.7[22] and Figure 3.8 show support, opposition, and doubt by each probe with at least 10 relevant tweets from two datasets – *Oct11Dataset* and *Feb12Dataset*. Our results from both *Oct11Dataset* and *Feb12Dataset* show the consistent pattern that the degrees of support for probes in the true group are very high, over 0.5 for all probes. Surprisingly, the degrees of support are also high for both debatable and false groups. For both plots, most of the tweets associated with the debatable probe *Actos causes bladder cancer* indicate high support, which is in

[22]Note that we recreate the belief plot from Experiment 2 here using the extended dataset and classifiers from Experiment 3.

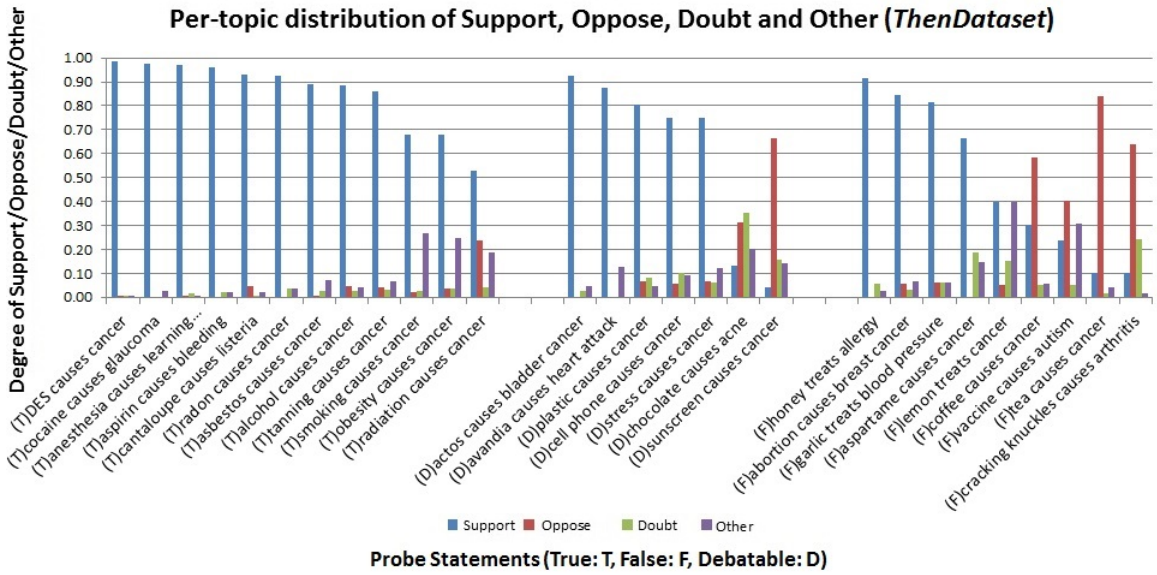


Figure 3.7: Belief Plot for *Oct11Dataset*

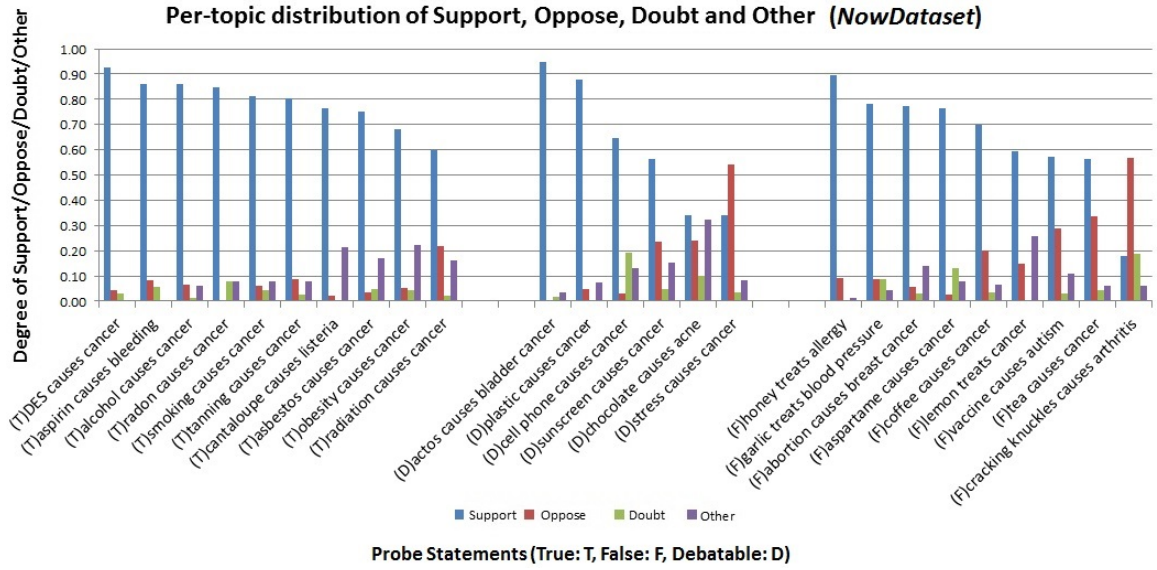


Figure 3.8: Belief Plot for *Feb12Dataset*

line with recent warning reports from the FDA and the class action lawsuit. While support is high for the true probe *radiation causes cancer*, opposition is also noticeable which perhaps reflects the dual nature of radiation in it's use as a therapy for cancer. We note that support for the false probe *abortion causes breast cancer* is dominant for both time spans, though we see an increasing number of politics infused tweets in *Feb12Dataset* consistent with the current events relating to the US presidential primaries. (e.g. "*<candidate name>, on the floor of the u.s. senate: 'abortion increases a woman's risk of breast cancer by 30 percent.'*")

It is also interesting to see how the volume of discussion for the same probes change over time. We see that the number of tweets retrieved for probes such as *cocaine causes glaucoma* and *anesthesia causes learning disabilities* drops to less than 10 tweets in the *Feb12Dataset*. The spike of tweets on these topics in the *Oct11Dataset* co-occurred with buzz created by new findings and corresponding publications which has since subsided. Hence we omit these probes in our belief plot for *Feb12Dataset*. The degrees of support for few probes in the true group reduce (e.g. *cantaloupe causes listeria*; given fewer incidences of this casualty in February 2012) while degrees of opposition increase for some probes (e.g. *tanning causes cancer*). In the debatable group, *stress causes cancer* has degree of support reducing from around 0.7 to around 0.3. In the false group, the degrees of support increase and degrees of opposition decrease for several probes (e.g. *tea/coffee causes cancer*). This means that some false probes are more believable now. Such changes are important to monitor and if significant may indicate the need for educational strategies. Also interestingly we see

that the false probe "coffee causes cancer" garners more support now than previously. An examination of tweets retrieved for this probe reveals a lot of discussion on the following topic: "*There are more than 1,000 chemicals in a cup of coffee. Of these, only 26 have been tested, and half caused cancer in rats*". As discussed previously, here also we see the influence of recent scientific studies on animal models creating a buzz in Twitter.

Aggregated support, opposition, and doubt for the *Oct11Dataset* and *Feb12-Dataset* are shown in Table 3.19. While the number of probes studied in the belief plots for these two datasets are different, the aggregate measures (Section 3.1), based on the average of *Degree_of_Support*, etc., are still comparable. We note that support for true probes in the *Feb12Dataset* is 7% lower compared to the *Oct11Dataset*. Moreover, we find that there has been a 16% increase in support for false probes. Support, opposition, and doubt regarding debatable probes remain almost consistent over this span of time. While these results hold for this particular set of probes and the specific time spans, nevertheless, it shows that the direction of support, opposition or doubt may change over time and it may be for better or for worse. As discussed above, we also note that most of these changes correspond to external events such as recent news stories, publications, etc.

3.8.5 Conclusions

In this experiment we tracked a set of pre-defined probes over different time spans. Overall, we find that for the set of selected probes, the amount of discussion as

Table 3.19: Aggregated Support, Opposition and Doubt over time.

	Oct11Dataset			**Feb12Dataset**		
	True Probes	False Probes	Debatable Probes	True Probes	False Probes	Debatable Probes
Support	0.86	0.49	0.61	0.79 (−7%)	**0.65 (+16%)**	0.62 (+1%)
Opposition	0.04	0.29	0.17	0.07 (+3%)	**0.20 (−9%)**	0.18 (+1%)
Doubt	0.02	0.09	0.11	0.03 (+1%)	0.06 (−3%)	**0.07 (−4%)**
Other	0.08	0.13	0.11	0.11 (+3%)	**0.09 (−4%)**	0.13 (+2%)

Note: Numbers in the parenthesis indicate percentage changes with highest percentage changes bolded.

well as positions (i.e. levels of support, opposition and doubt) change over time. These changes are generally triggered by external events such as news stories, publications, etc.

3.9 Experiment 7

3.9.1 Goals

In Experiment 4 (Section 3.6) we harvested new probes from a week's tweet stream. Here our goal is to extend probe harvesting to a longer time period. Specifically we harvest probes over a 43 week data stream. We then conduct a surveillance study for a few harvested probes.

3.9.2 Dataset

Using the retrieval technique outlined in the 'Probe Harvesting' framework (Section 3.1) we collected tweets for a set of 200 drugs, 78 diseases and 932 hashtags focusing on cause/causes and treat/treats relationships. These were identified using the most prescribed over-the-counter[23] and prescription medicines[24], WHO's list of

[23]http://www.drugs.com/otc/

[24]http://www.drugs.com/stats/top100/2012/sales

chronic[25] and infectious diseases[26], and a list of popular healthcare hashtags[27] respectively. For each drug, disease and hashtag we performed weekly searches using the Twitter Search API starting from March 28, 2013 to January 14, 2014 (43 weeks). We collected 1,771,601 tweets during this time period.

3.9.3 Methods

In this experiment, due to the large number of tweets from which we aim to harvest probes, we deviate slightly from the 'Probe Harvesting' framework. In the MetaMap based probe harvesting approach we typically used heuristics to associate semantically-labeled biomedical concepts to form probes. These heuristics don't scale well and are thus not viable for identification of thousands of potential probes. Hence in stead of using MetaMap we use NLM's SemRep[28] for processing each tweet. SemRep automatically extracts semantic predictions from tweets as subject-relation-object triples, a format which is consistent with our probes.

As a second step, a handful of these harvested probes (300) were selected based on their frequency of occurrence in the tweet collection. We manually eliminated those which were widely known or vague. The remaining are then fed into the 'Probe Surveillance' component of the computational framework. The retrieval of tweets for this stage was conducted on March 18, 2014. Probes retrieving less than 10 tweets are

[25]http://www.who.int/chp/en/

[26]http://www.who.int/topics/infectious_diseases/en/

[27]http://www.symplur.com/healthcare-hashtags/

[28]http://semrep.nlm.nih.gov/

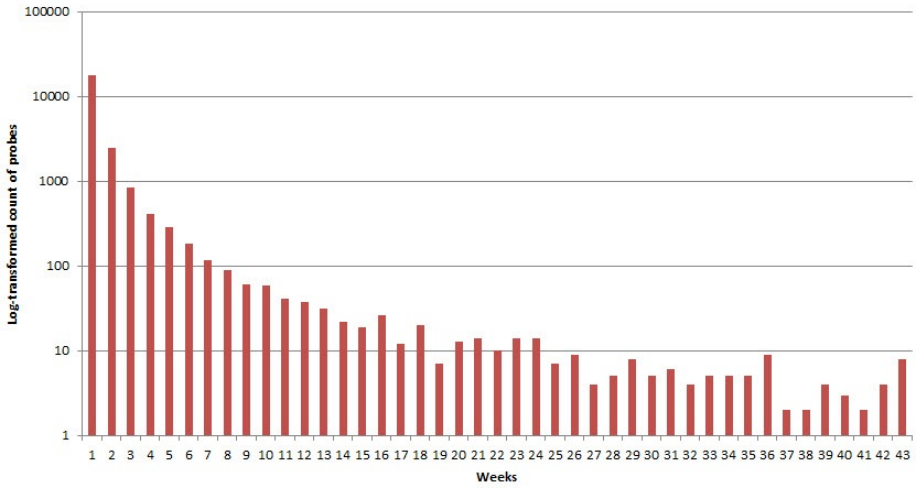


Figure 3.9: Harvested probe distribution over time (in weeks)

omitted from our analysis. The remaining probes were classified by a physician into true, false and debatable categories. Finally, we conduct belief surveillance and draw belief plots for these harvested probes (true probes clustered in the left, debatable in the middle and false in the right) and also calculate the levels of support, opposition and doubt using aggregate measures.

3.9.4 Results & Analysis

3.9.4.1 Harvested Probe Analysis

In our dataset of over 1.7 million tweets, the majority of tweets were retrieved using hashtags (59.8%), followed by disease (34.6%) and drug (5.6%) terms. We harvested 140,751 non-unique probes from these tweets. 22,475 of these probes were unique. While 2,502 (9.8%) of them appear for more than 5 times, only 166 (0.6%) of these appear 100 times or more.

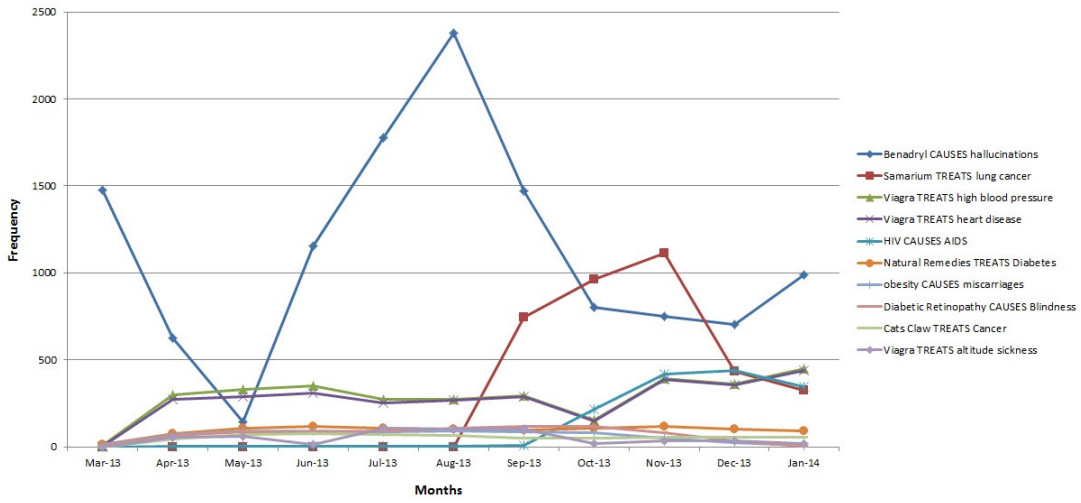


Figure 3.10: Distribution of most frequent probes over time (in months)

A temporal analysis of harvested probes reveals various interesting observations. In Figure 3.9 we find that only 8 probes are actually persistent through the 43 week span of our dataset. These probes are: dengue hemorrhagic fever causes death, diabetes causes blindness, natural remedies treats diabetes, sugar causes diabetes, sweet causes diabetes, vaccines causes autism, Viagra treats heart disease and Viagra treats high blood pressure. The majority of probes (78%) appear in only one week.

Further analysis of the most frequently identified probes (Figure 3.10) over the 11 month period (3-month rolling averages) reveals that probes such as 'Benadryl causes hallucinations' or 'Viagra treats high blood pressure' are harvested every month. Some probes such as 'Samarium treats lung cancer' start appearing at an intermediate time point but sees a subsequent rise in popularity. For 'Samarium treats lung cancer' the spike could be attributed to tweets like 'Breaking Bad has 62 episodes. The 62nd element on the periodic table is Samarium, which is used to treat

lung cancer', which became very popular after the season finale of the popular TV show 'Breaking Bad'. The probe 'HIV causes AIDS' is also harvested minimally till it sees a spike around November-December 2013. Inspecting the tweets that resulted in this probe we found that most tweets were about raising awareness for HIV/AIDS around World AIDS Day observed on December 1.

Manual analysis of a set of 100 randomly selected probes reveal that while a majority of the probes (65%) are quite specific (e.g. *Benadryl causes hallucinations*), many of them are also quite broad (e.g. *pills treat cancer*). Since we are primarily interested in specific probes we filtered our set of probes using stopwords that represent broader categories of drugs (e.g. pills, medicines, etc.), diseases (e.g. illnesses, complications, etc.) and other general terms (e.g. food, drinks, etc.). A total of 44 stopwords[29] were used for this purpose. This resulted in a filtered set of 15,675 probes which are analyzed further.

We found that the most discussed drugs in our probes were: Benadryl, Viagra, Samarium, marijuana, vaccines, Taurine, Anabesol, Aspirin, acetaminophen and Tylenol. These drugs belong to OTC, prescribed and recreational drug categories. Diseases that were most frequent in these probes were: hallucinations, blindness, obesity, diabetes, lung cancer, high blood pressure, heart disease, AIDS, asthma and autism. These diseases belong to broader categories of sensory, chronic, genetic and infectious diseases.

We also analyzed the probes based on the multiplicity of association of drugs

[29]List of 44 stopwords:http://geordi.cs.uiowa.edu/sbhttcha/44_stoplist.txt

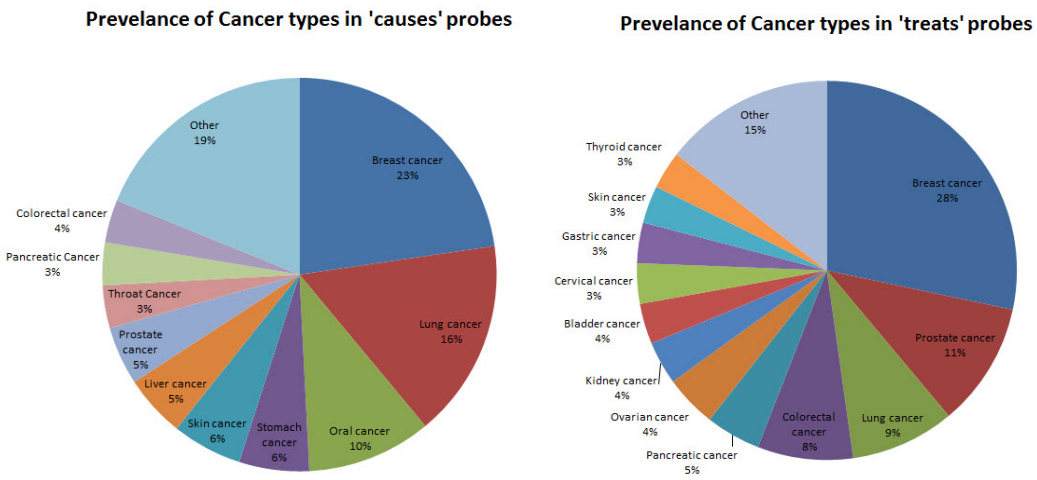
or diseases. For example, besides being one of the most discussed drugs, Viagra also features in more probes (147) than any other drugs. A few example probes where Viagra appears are: Viagra treats altitude sickness, Viagra treats neck tumor, Viagra treats menstrual cramps, etc. Aspirin (128), Advil (110), marijuana (98), Tylenol (97), NyQuil (79), and Benadryl (69) are the other drugs which appear in multiple probes.

We find cancer (587) to be the most featured disease in probes, followed by diabetes (541), obesity (510), blindness (497), asthma (323), pain (182), arthritis (179), epilepsy (178), depression (131) and pneumonia (131). We also find several subtypes of cancer which are quite prevalent in probes.

Figure 3.11 shows the types of cancer that are being discussed in the 'causes' and 'treats' probes. We find that the breast cancer is the most discussed type of cancer related to both causation and treatment. This is in line with Nation Cancer Institute's (NCI) estimates of new cancer cases[30] where breast cancer is ranked as the second leading cause. Lung cancer, which is the third leading cause of new cancer cases, is also one of the most frequent cancer types found in our probes. Prostrate cancer, which is the leading cause of new cancer cases is the second most prevalent cancer type in relation with treatment. Other cancer types, such as gallbladder or scrotal cancer which are sparingly discussed in probes are grouped in the 'Other' category.

We also identified the top 10 causes of cancer as discussed in our probes.

[30]http://www.cancer.gov/cancertopics/types/commoncancers



(a) Cancer types in 'causes' probes (b) Cancer types in 'treats' probes

Figure 3.11: Types of cancer identified in probes

These are: HPV, alcohol, deodorant, aspartame, toxins, Actos, marijuana, sunscreen, tobacco and BPA. The top 10 treatments for cancer are: surgery, chemotherapy, radiation therapy, Cisplatin, immunotherapy, Avastin, marijuana, proton therapy, mastectomy and Docetaxel. These causal or treatment agents found in our probes correspond to some of the most common carcinogens[31] and treatments[32] of cancer.

3.9.4.2 Probe Surveillance

We selected an initial list of 200 most frequent probes for probe surveillance. This list is then filtered by eliminating widely known (e.g. HIV causes AIDS) as well as vague probes (e.g. Bacteria causes leprosy). This gives us a list of 27 probes which are fed into the Probe Surveillance component (Section 3.1) of the computational

[31]http://www.cancer.org/cancer/cancercauses/

[32]http://www.cancer.org/treatment/treatmentsandsideeffects/treatmenttypes/index

framework.

Table 3.20: Retrieved tweets for harvested probes.

Probe	**# tweets**	**Truth Status**
Epilepsy causes seizures	1588	T
Serotonin causes depression	1532	D
Marijuana treats Epilepsy	1348	T
Marijuana treats multiple sclerosis	857	D
HIV causes heart disease	392	T
HPV causes cervical cancer	297	T
MMR causes autism	227	F
Diabetes causes kidney failure	203	T
Iodine treats breast cancer	133	D
Diabetes causes blindness	125	T
Natural Remedies treats Diabetes	107	D
Marijuana treats Diabetes	107	D
Energy drinks causes miscarriages	102	F
Marijuana causes Schizophrenia	74	D
Samarium treats lung cancer	59	T
Viagra treats high blood pressure	50	T
Diabetes causes erectile dysfunction	46	T
Acrylamide causes cancer	41	D
Acupuncture treats arthritis	40	T
Cats Claw treats Cancer	40	F
Viagra treats heart disease	24	D
Diabetic Retinopathy causes Blindness	16	T
Total	**7408**	

Table 3.20 shows the 22 probes that retrieved at least 10 relevant tweets. The total number of tweets retrieved by these probes is 7408. For each probe a physician provided decisions on whether the probe is true, false or debatable. We note here that while we considered the most frequent harvested probes for surveillance many of them

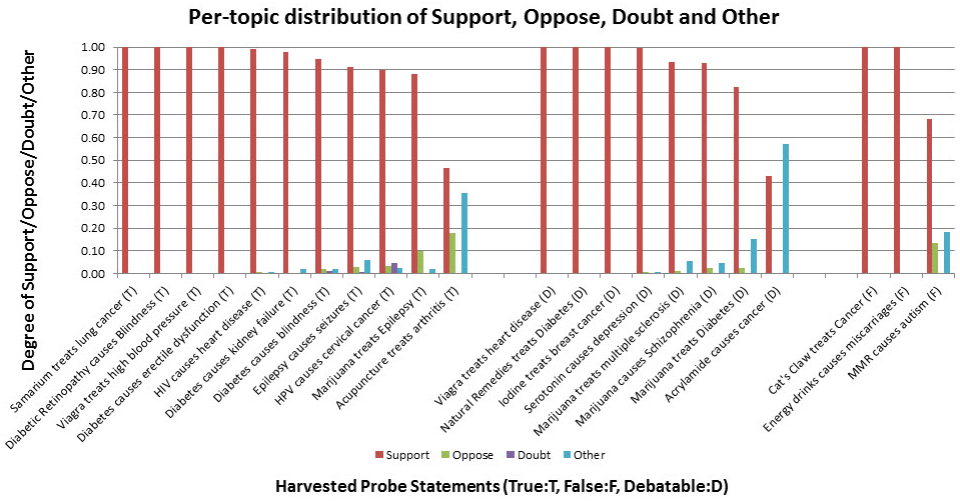


Figure 3.12: Belief Plot for Harvested Probes

retrieved very few tweets. We believe that this is because of the temporal differences in Twitter activity around these probes. For example, while 'Viagra treats altitude sickness' was one of the most frequent probes during the 2013, there were very few tweets on this topic in early 2014, when the data for surveillance was gathered.

More than one-third of the 22 probes pertain to recreational drugs or alternative medicines. Few probes associate viruses, prescription drugs and dietary substances with causal or curative relationships with diseases. As observed before, probes mined depend upon current events and developments. For example, some recent research discussing the role of Iodine in the treatment of breast cancer invigorated a lot of discussions in this topic.

Figure 3.12 shows the plots of support, opposition, doubt, and other for the true, false and debatable probes having at least 10 relevant tweets. Here we notice that there is surprisingly low support for some true probes such as *acupuncture treats*

arthritis which is a known therapy for arthritis. Additionally there is a high level of support in debatable probes, especially for those pertaining to recreational drugs (e.g. marijuana). We also notice that a significant proportion of tweets retrieved for probes like *acrylamide causes cancer* contain advertisements which are classified in the 'other' category. We see high support in false probes related to alternative medicine and energy drinks. In comparison to our findings from Experiment 1 (Section 3.3) relating vaccine to autism, we find that there is even stronger of support for *MMR causes autism*. This might have been influenced by some recent comments made by a popular television personality [33].

Table 3.21: Support, Opposition and Doubt Measures for Harvested Probes.

	True Probes	**False Probes**	**Debatable Probes**
Support	**0.92**	0.89	0.84
Opposition	0.03	**0.05**	0.03
Doubt	**0.01**	0	0
Other	0.04	0.06	**0.13**

The aggregate calculations of support, opposition, doubt, and other are shown in Table 3.21. Similar to the belief plot, we see high support (0.92) for true probes. Moreover, there is almost equally high support (0.89) for false probes and debatable probes (0.84). Low opposition (only 0.05) to false probes is alarming and emphasizes

[33]http://nypost.com/2014/03/18/anti-vaccine-activist-jenny-mccarthy-mother-of-plagues/

the need to embark on public health campaigns to correct such misinformed beliefs. Similar to Experiment 4 (Section 3.6) we find extremely high level of support for false probes. We speculate that people are more likely to proactively express about their support for something rather than their opposition.

3.9.5 Conclusions

In this experiment we showed that we can harvest probes over time and study them using our surveillance framework. We find that the frequency of harvested probes can vary considerably with time and there may be wide fluctuations triggered by external events. We also find that prevalence of certain probes harvested from social media correspond to national statistics. For a handful of the harvested probes we estimated the levels of support, opposition and doubt. For these probes, we found that false and debatable probes are supported as highly as true probes. We also note that there is no doubt in the debatable and false harvested probes. These observations probably correspond to the nature of the probe harvesting technique.

3.10 Conclusions

In this chapter we asked the general research question: *What are the beliefs held by social media users?* Since beliefs can be of a wide variety we limited our analysis to only health beliefs. We then asked several specific research questions (see Section 3.2) pertaining to health beliefs. To address each of these questions we designed corresponding experiments.

To begin with, we proposed a belief surveillance framework comprising of

two primary components – probe surveillance and probe harvesting [10]. The probe surveillance component deals with measuring the levels of support, opposition and doubt for pre-defined probes.The harvesting component deals with identification of naturally expressed probes from tweets. These newly identified probes in turn may be fed into the surveillance component.

In Experiment 1 the goal was to show the applicability of the Probe Surveillance component for surveillance of a set of handpicked true, false and debatable probes. We showed that we can study the levels of support, opposition and doubt for these statements using manually annotated tweets. We used belief plots for visualizing the levels of support, opposition and doubt for individual probes. Aggregation methods were used to get composite measures for a set of probes. Overall we found high levels of support for not just true probes but also for false and debatable probes. There was considerable doubt for both true and false probes.

In Experiment 2 the goal was to develop automated methods for Probe Surveillance. We showed that we can build classifiers using off-the-shelf machine learning tools to classify tweets into relevant and non-relevant categories and further classify the relevant tweets into one of the three categories: support, opposition and doubt. The results were encouraging for both these classifiers in 10-fold cross-validation experiments.

In Experiment 3 we aimed to develop generalizable classifiers that we could use to classify new probes. We showed that we are able to build classifiers that are quite generalizable and can label tweets from unseen probes fairly well.

Next we focused on the 'Probe Harvesting' for identification of naturally expressed beliefs from Twitter (Experiment 4) [9]. Using two different approaches for probe harvesting we showed the use of biomedical concept identification and semantic mapping of tweets as effective ways for harvesting probes. The harvested probes, fed into the surveillance framework, show high level of support for false and debatable probes, sometimes even comparable to support for true probes. Overall, we found that the level of support for false and debatable harvested probes was higher compared to the pre-defined probes from Experiment 1.

Estimating the novelty of harvested probes is also important (Experiment 5) [8]. Using PubMed as a reference for testing novelty, we found that several probes mined from Twitter are either not present or sparsely present in PubMed. While these probes appear to have potential towards developing new hypotheses for scientific research, further validation may be involved especially because the probes are harvested from social media.

We also experimented with the temporal variability of the levels of support, opposition and doubt for some specific probes (Experiment 6). Overall, we find that the level of support in true probes decreases over time while it increases for false and debatable probes. We also find that the level of doubt in false or debatable probes decreases over time. However, note that these observations are probe and time specific.

Experiment 7 is an extension of Experiment 4. Here we harvested probes over a longer period of time time and studied a handful of them using our surveillance

framework. We found that the frequency of harvested probes may change considerably with time. A widely popular probe in the past may never appear again in the future. Consistent with previous experiments, we find high support for false probes and low opposition to debatable probes.

In summary, we contribute a novel belief surveillance framework. We show through various experiments that Twitter is a good source for surveillance of known health beliefs as well as for finding new ones. Across various experiments we consistently find that there is high support for false and debatable probes. This is of concern especially for public health educators and shows the importance of using our surveillance framework for identifying areas of public health that need special attention for correcting false notions about health.

There are some limitations to our study. For example, for probe surveillance we have not considered tweet features such as type of source (drug company, medical organization etc.). More importantly, we do not distinguish between a statement that expresses certainty with another that expresses less confidence, even though both may take the same position w.r.t. the probe statement. Another angle that arises from our work is that there may be a baseline level of beliefs – a propensity. Belief positions extracted should probably be gauged against such baselines. One way to measure this would be to calculate baseline measures for these positions by averaging over the set of probes that we have already studied. We can then observe the variance in positions for new probes w.r.t. to this baseline to easily identify ones that are worth looking into. Another approach can be to identify the tendency to believe in the written word

(i.e., support probes irrespective of what they are). We can test this quite simply by considering a set of propositions that includes both the positive and the negative versions. For example, we would include both A causes B and A does not cause B. Then we would check the overall tendency to agree irrespective of the position taken by the proposition. This can be done via a simple crowdsourcing experiment. We also do not compare our results with other sources or surveys primarily because of the lack of comparable knowledge bases. In the harvesting framework we limit ourselves to specific drugs, diseases or hashtags to find probes related to causes and treatment of illnesses. However there may be more health-related discussions on social media that are not captured by our methods. In future research we would like to investigate more in these directions.

CHAPTER 4
ASSESSMENT OF SOCIAL MEDIA ENGAGEMENT FOR ORGANIZATIONS

Government agencies are increasingly interested in using social media to distribute information at the national, state and local levels. U.S Federal agencies, for example, routinely use a variety of social media sites including Twitter, Facebook, YouTube, Flickr, and Instagram to enhance communication[1]. In addition to distributing information, government agencies are increasingly interested in interacting with the populations they serve. For example, new guidelines entitled "Digital Governmental Strategy" outline specific steps for governmental agencies to make digital information more "customer centric"[2]. This bidirectional form of communication can be defined as engagement: interactions designed to promote some common goal [58].

To date no study has systematically explored factors associated with the levels of health agency engagement on social media. We consider two of the leading social media platforms – Twitter and Facebook to address this gap.

4.1 Engagement of Health Organizations in Twitter

The primary measure for gauging engagement on Twitter is retweeting activity. A retweet is an acknowledgment that the original tweet has been read and also that it is viewed as sufficiently interesting to merit a re-post. The followers of the retweeting account now have ready access to the original retweet. Retweets are in some sense

[1]http://govsm.com/w/Federal_Agencies

[2]http://1.usa.gov/MgEHYl

analogous to citations in an article. A second aspect to engagement relates to the time period over which retweeting occurs. A tweet with a longer retweeting time span compared to another is one where engagement occurs over a longer period of time. Thus, Twitter engagement for a federal agency is maximized when all of its tweets generate the highest possible number of retweets with retweets starting almost immediately after the tweet is posted and continuing on forever. While in practice these conditions are never achieved, it is clear that some tweets generate stronger responses than others. Our overarching goal is to determine whether there are features that relate to higher levels of retweeting and longer lifespans of tweets in order to offer insight into ways to improve Twitter engagement for health agencies.

Specifically we address the following three questions with respect to Twitter messages posted by US Federal Health agencies and their responses. First, which features are associated with the level of response in the form of retweets? Second, which features are associated with the interval between an agency's tweet and its first retweet? Third, which features are associated with the interval between an agency's tweet and the last retweet it generates? We address our goals by following almost all of the tweets ever posted by the 130 Twitter accounts of 25 Federal Health Agencies on Twitter. This allows us to present a close to complete picture of the levels of engagement achieved by these agency accounts. We explore levels of engagement using hurdle regression and Cox proportional hazards regression models. We consider several features affect engagement. Features include ones that are typically studied such as numbers of followers and friends and also ones that are rarely studied such

as the semantic content of a tweet.

4.1.1 Data Collection

4.1.1.1 Agencies & Handles

We selected health agencies through the HHS Social Hub website[3] which maintains a list of all official HHS-affiliated social media accounts across various platforms. We identified all agencies with Twitter accounts (also known as handles). A total of 134 Twitter accounts were identified out of which 4 were either deleted or suspended or had no tweets posted in their lifetime. We used the remaining 130 handles in our study. These correspond to 25 different health agencies, fifteen are NIH divisions such as NIH/NLM, NIH/NIAIA and NIH/NCI. Some agencies have quite a few handles such as NIH/NCI (13 handles: SmokefreeGove, NCIHINTS etc.), CDC (25 handles: CDCgov, CDCActEarly etc.), FDA (10 handles: US_FDA, FDATobacco etc.), and others have just one handle such as AHRQ, ACF and NIH/NEI. Table 4.1 lists the various agencies, the number of handles for each and a few examples of handles.

4.1.1.2 Tweets & Retweets

The Twitter REST API v1.1 (user_timeline)[4] was used to collect all tweets from a handle's timeline as of late November 2012 . A maximum of 3200 tweets from a handle's timeline can be retrieved using this method. We could collect all posted tweets for 112 handles; 18 handles had more than 3200 tweets at the time of data

[3]http://www.hhs.gov/socialhub/

[4]https://dev.twitter.com/docs/api/1.1/get/statuses/user_timeline

Table 4.1: Agencies and Handles.

Agency	Name	# handles	Examples of handles
ACF	Administration for Children & Families	1	HeadStartgov
AHRQ	Agency for Healthcare Research & Quality	1	AHRQNews
CDC	Center for Disease Control & Prevention	25	CDCgov, CDCActEarly, CDC_BioSense, etc.
CMS	Centers for Medicare & Medicaid Services	4	CMSGov, CMSinnovates, IKNGov, etc.
FDA	U.S. Food & Drug Administration	10	US_FDA, FDATobacco, FDADeviceInfo, etc.
HRSA	Health Resources & Services Administration	1	HRSAgov
NIH	National Institutes of Health	15	NIHforFunding, NIHprevents, NIHClinicalCntr, etc.
NIH/NIA	National Institute on Aging	1	NIAGo4Life
NIH/NCCAM	National Center for Complementary & Alternative Medicine	1	NCCAM
NIH/NCI	National Cancer Institute	13	SmokefreeGov, NCIHINTS, NCIBulletin, etc.
NIH/NEI	National Eye Institute	1	NEHEP
NIH/NHLBI	National Heart, Blood & Lung Institute	3	TheHeartTruth, nih_nhlbi, BreatheBetter
NIH/NIAAA	National Institute of Alcohol Abuse & Alcoholism	1	NIAAAnews
NIH/NIAID	National Institute of Allergy & Infectious Diseases	3	NIAIDNews, NIAIDCareers, NIAIDFunding
NIH/NIAIMS	National Institute of Arthritis & Musculoskeletal & Skin Diseases	1	NIH_NIAMS
NIH/NICRR	National Center for Research Resources	1	ncrr_nih_gov
NIH/NIDA	National Institute of Drug Abuse	1	NIDAnews
NIH/NIEHS	National Institute of Environmental Health Sciences	1	NIEHS
NIH/NIGMS	National Institute of General Medical Sciences	1	NIGMS
NIH/NIHGRI	National Human Genome Research Institute	1	DNAday
NIH/NIMH	National Institute of Mental Health	1	NIMHgov
NIH/NLM	National Library of Medicine	11	NLM_LHC, medlineplus, NCBI, etc.
OIG	Office of Inspector General	1	OIGatHHS
OS	Office of the Secretary	29	AIDSgov, bestbones4ever, BirdFluGov, etc.
SAMHSA	The Substance Abuse & Mental Health Services	2	samhsagov, distressline
Grand Total		130	

collection. Handles such as CDCSTD, womenshealth and CDCNPIN had posted over 9000 tweets by the time of the data collection. For such handles the most recent 3200 tweets were collected. For each agency tweet, we recorded its unique identifier and raw retweet count among other tweet-based data and metadata as described below.

We collected a total of 164,104 tweets from the timelines of the 130 handles. A third of the tweets (53,556) had zero retweets, i.e., generated no observable engagement. Less than 1% (613) had more than 100 retweets (total = 174,395, mean = 284). The remaining two-thirds (109,935) of tweets fell between these ranges (mean = 7.5, total = 826,052 retweets). Table 4.2 shows summary details about tweets and retweets per agency. Similar details per handle are displayed in Table 4.3.

In raw numbers we note that while the CDC posted the most tweets (37,136), it also has the highest raw number of tweets that are not retweeted (11,063). In

Table 4.2: Tweets and Retweets per Agency.

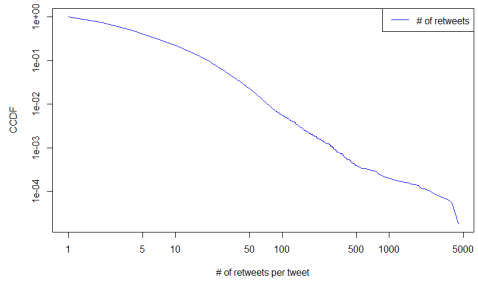
Agency	Date first handle was created	# tweets	# tweets with at zero retweets	# tweets with at least 1 retweet	# retweets	# retweets per tweet	# retweets per non-zero retweeted tweet
ACF	9/7/2011	605	219 (36.2%)	386 (63.8%)	1924	3.18	4.98
AHRQ	6/5/2009	1475	415 (28.14%)	1060 (71.86%)	3432	2.33	3.24
CDC	7/24/2008	37136	11063 (29.79%)	26073 (70.21%)	278885	7.51	10.70
CMS	9/1/2009	5620	2132 (37.94%)	3488 (62.06%)	11023	1.96	3.16
FDA	12/11/2008	10574	3007 (28.44%)	7567 (71.56%)	75245	7.12	9.94
HRSA	6/1/2009	1241	332 (26.75%)	909 (73.25%)	5391	4.34	5.93
NIH	6/16/2008	15550	7446 (47.88%)	8104 (52.12%)	49666	3.19	6.13
NIH/NIA	10/18/2011	1891	629 (33.26%)	1262 (66.74%)	10556	5.58	8.36
NIH/NCCAM	8/20/2009	1489	568 (38.15%)	921 (61.85%)	4102	2.75	4.45
NIH/NCI	4/28/2009	15679	5580 (35.59%)	10099 (64.41%)	46586	2.97	4.61
NIH/NEI	3/23/2011	401	249 (62.09%)	152 (37.91%)	331	0.83	2.18
NIH/NHLBI	2/26/2009	5135	1526 (29.72%)	3609 (70.28%)	29447	5.73	8.16
NIH/NIAAA	7/15/2010	424	122 (28.77%)	302 (71.23%)	2279	5.38	7.55
NIH/NIAID	7/24/2009	1725	830 (48.12%)	895 (51.88%)	2808	1.63	3.14
NIH/NIAIMS	8/31/2009	822	135 (16.42%)	687 (83.58%)	1850	2.25	2.69
NIH/NICRR	8/14/2009	1029	704 (68.42%)	325 (31.58%)	515	0.50	1.58
NIH/NIDA	1/5/2010	2191	669 (30.53%)	1522 (69.47%)	7484	3.42	4.92
NIH/NIEHS	12/17/2009	682	320 (46.92%)	362 (53.08%)	858	1.26	2.37
NIH/NIGMS	9/2/2009	983	420 (42.73%)	563 (57.27%)	1791	1.82	3.18
NIH/NIHGRI	2/25/2009	401	180 (44.89%)	221 (55.11%)	652	1.63	2.95
NIH/NIMH	5/11/2009	959	177 (18.46%)	782 (81.54%)	16779	17.50	21.46
NIH/NLM	2/12/2009	15058	6525 (43.33%)	8533 (56.67%)	48497	3.22	5.68
OIG	5/2/2011	1476	386 (26.15%)	1090 (73.85%)	2459	1.67	2.26
OS	5/30/2007	36587	8026 (21.94%)	28561 (78.06%)	376158	10.28	13.17
SAMHSA	3/17/2009	4971	1896 (38.14%)	3075 (61.86%)	21729	4.37	7.07
Total		164104	53556 (32.64%)	110548 (67.36%)	1000447	6.10	9.05
Mean (SD)		6564.16 (10355.66)	2142.24 (3055.12)	4421.92 (7499.23)	40017.88 (89880.72)	4.10 (3.64)	5.99 (4.39)

Table 4.3: Top 10 handles with most retweets per tweet.

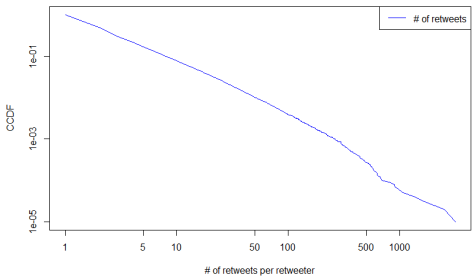
Handle	Date of creation	# tweets	# of tweets with non-zero retweets	# of tweets with zero retweets	# retweets	# of retweets per tweet	# of retweets per non-zero retweeted tweet
CDCemergency	1/28/2009	792	523 (66.04%)	269 (33.96%)	36756	46.41	70.28
FitnessGov	9/15/2011	935	834 (89.2%)	101 (10.8%)	23003	24.61	27.58
womenshealth	5/30/2007	3236	3163 (97.74%)	73 (2.26%)	85832	26.52	27.14
HealthCareGov	11/1/2009	409	404 (98.78%)	5 (1.22%)	10315	25.22	25.53
HHSGov	6/4/2009	1295	1103 (85.17%)	192 (14.83%)	26313	20.31	23.86
FDArecalls	12/11/2008	2118	1278 (60.34%)	840 (39.66%)	29764	14.05	23.29
CDCgov	5/21/2010	3226	2904 (90.02%)	322 (9.98%)	66204	20.52	22.80
CDC_eHealth	7/24/2008	1517	1255 (82.73%)	262 (17.27%)	27856	18.36	22.20
NIMHgov	5/11/2009	959	782 (81.54%)	177 (18.46%)	16779	17.49	21.46
PHEgov	4/26/2010	1356	998 (73.6%)	358 (26.4%)	20683	15.25	20.72

contrast, the Office of the Secretary (OS), a close second in the number of total tweets (36,587), has the highest number of retweeted tweets (28,561) and also the highest number of retweets (376,158). Each tweet from OS gets approximately 10 retweets. The agency with the most retweets per retweeted tweet is NIH/NIMH with about 18 retweets per tweet. Also, it leads the agencies with 82% of its tweets retweeted at least once. Interestingly, this agency has less than 1000 tweets. Table 4.3 shows the top 10 handles ranked by the number of retweets per tweet. These are: CDCemergency (CDC), FitnessGov (OS), womenshealth (OS), HealthCareGov (OS), HHSGov (OS), FDArecalls (FDA), CDCgov (CDC), CDC_eHealth (CDC), NIMHgov (NIH/NIMH), PHEgov (OS).

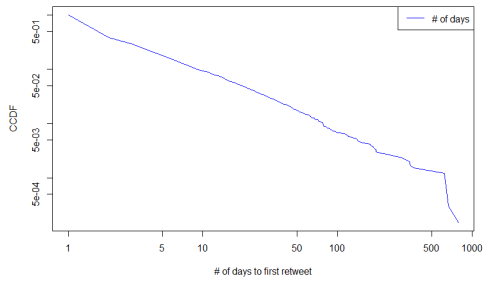
The four panels of Figure 4.1 plot different aspects of retweeting: a) retweets/



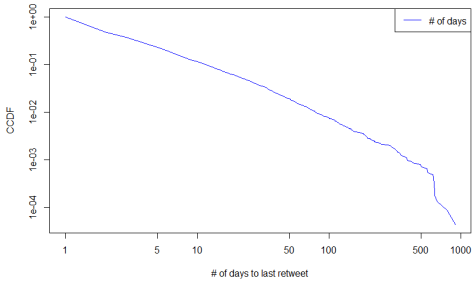
(a) Plot of # of retweets per tweet



(b) Plot of # of retweets per retweeter



(c) Plot of # of days to first retweet



(d) Plot of # of days to last retweet

Figure 4.1: Power-law plots of various retweet-based features

tweet, b) retweets/retweeter, c) time to first retweet and d) time to last retweet. In each plot the y-axis represents the complementary cumulative distribution function (CCDF) as used for example in [44]. With the exception of time to first retweet the plots fit power-law distributions with exponents in the range generally expected for most real-world networks (between 2 and 3) [18].

From Figure 4.1(a) we find that only a handful of tweets get multiple retweets; the majority get very few retweets. This is not surprising and corroborates previous findings [44, 28, 41]. A very small number of tweets get more than 500 retweets. The plot fits a power-law distribution with an exponent of 2.56. Figure 4.1(b) shows the plot of the number of retweets per retweeter. We note that a few Twitter users retweet extensively while the majority of them retweet sparingly. Only a few retweeters retweet more than 500 times. The plot fits a power-law distribution with an exponent of 2.35.

88.46% of the retweeted tweets get their first retweet on the day of the tweet (referred to as day zero in our discussion). The remaining tweets are plotted in Figure 4.1(c). We note that very few tweets are retweeted after 100 days. The plot, with an exponent of 1.87, does not fit a typical power-law distribution. 60.6% of the retweeted tweets get their last retweet on day zero (note that tweets with only one retweet get their first as well as last retweet on day zero). We note that very few tweets get their last retweet after day 500. The plot fits a power-law distribution with an exponent of 2.33.

For the sake of completeness we also look for HHS handles doing the retweet-

ing. We find that 117 of the 130 handles retweet each other's tweets. The top retweeting handles are womenshealth with 2500 retweets followed by the NIH/NCI with 1662 retweets. MedicareGov, NCITechTransfer, NEHEP, NIAIDFunding and NIOSHManuf have the lowest retweet counts with 1 retweet each. Apart from these HHS handles, OrleansCoHealth, the Twitter handle of Orleans County Health Department (New York), has the highest retweeting activity with 3154 retweets.

4.1.2 Tweet Features

First we decided which features we would use to represent each tweet. We included those examined commonly in Twitter-based studies as well as those that have not yet been considered. Table 4.4 lists 11 features we considered under 2 broad categories: handle-level features that are the same for all tweets issued by a handle (e.g., numbers of followers and friends) and tweet-specific features such as sentiment.

We also divided the features into two logical groups. Group 1 has features that cannot be changed or easily manipulated by an account holder. We include tweet age in this group as it represents a natural phenomenon that cannot be changed. The account holder has control over Group 2 features.

4.1.2.1 Handle Level Features

Several handle level features may be associated with engagement levels. In particular the numbers of friends and followers may be important. If user Y is a follower of user X then it means that Y receives all of X's tweets automatically. Also, X is regarded as a friend of Y. Relevant to us is that a tweet is displayed on the

Table 4.4: Features Examined.

Type	Group	Features	Description
Handle-level	1	Favorites	# of users favoriting tweets of a particular handle (log-transformed).
	1	Followers	# of users following a particular handle (log-transformed).
	1	Friends	# of users followed by a particular handle (log-transformed).
	1	Betweenness-centrality	Importance of node in network.
	2	Status count	# of tweets posted by a handle in its lifetime (log-transformed).
Tweet-level	1	Tweet age	# of days between handle creation and tweet post (log-transformed).
	2	Hashtag	Whether a tweet contains a hashtag, word prefixed with # (binary).
	2	URL	Whether a tweet contains an URL, http, ftp, etc. (binary).
	2	User-mention	Whether a tweet contains a user-mention, word prefixed with @ (binary).
	2	Sentiment	Two scores: one for positivity and another for negativity.
	2	Content (Semantic Groups)	Classification of each tweet into 15 semantic groups using MTI followed by post-processing. Multiple classes per tweet allowed.

timelines of all of its handle's followers, so these are the users most likely to retweet the post. Although the tweet is open and available to the entire Twitter community, it is only accessible by others through a search window. Once accessed by any user, retweeting is possible. It seems likely that the number of followers is associated with retweeting, but it is less clear whether the number of friends is associated as well. Previous studies show differing results. Some studies find that higher follower and friend counts imply higher retweetability [72, 71] while another study [15] claims that it is the content of tweets and not follower count that drives retweetability.

For each handle we identify the numbers of followers and friends using the

Twitter REST API v1.1 (users/show)[5]. Figure 4.2 shows a scatter plot of followers versus friends. We find that CDCemergency has the highest number of followers (1,432,424) but very few friends (393). On the other hand GoHealthyPeople has many friends (7,688) but few followers (34,913). NIAIDCareers (1008: 729) and distressline (1701: 1203) have relatively balanced number of followers and friends in comparison to the overall ratio of followers and friends for the different handles (49832: 405).

We also study the number of favorites, i.e., the number of users favoring a particular handle. NLM_DIMRC has the highest number (575) of favorites, followed by GoHealthyPeople (343) and AIDSgov (216). 50 handles (e.g. NIHLBI, DNADay, NCBI) did not have any favorites.

The top ranking handles in status are CDCSTD (12151), womenshealth (9419), CDCNPIN (9157), NIOSH (8936) and talkHIV (7663) and the lowest 5 are ncbi_pubmed (60), NCISymptomMgmt (144), NIOSH_FirRanges (150), FDACBER (162), and Medicare_Fraud (171).

Another property that may associate with engagement is the betweenness-centrality score for the handle. Betweenness centrality shows the extent to which a node acts as an intermediary in the shortest path between nodes in the network. This shows the importance of a particular node with respect to the network structure. Figure 4.3 shows a graph representing the betweenness-centrality scores of all the health agencies computed in a network with friends and followers links. Node size is representative of the betweenness-centrality score; NIHforHealth, CDCgov and HHSGov

[5]https://dev.twitter.com/docs/api/1.1/get/users/show

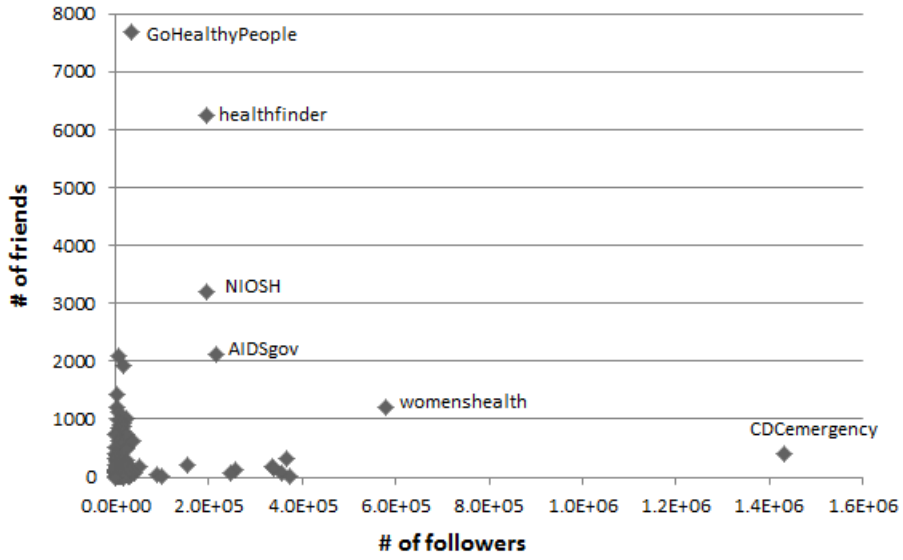


Figure 4.2: Plot of # of followers vs. # of friends for each handle (few handles with disparate distribution of followers and friends have been labeled)

have the highest betweenness-centrality values of 987.2, 851.51 and 717.54 respectively. Betweenness-centality does not apply to nodes with zero in- or out-degrees, namely for NIHforFunding and nlm_newsroom.

While betweenness-centrality has been used extensively to determine influential users in social media [69, 86, 12, 83] in various domains ranging from health to politics, in most cases it used as a metric of influence in a retweet or a reply network. To the best of our knowledge, researchers have not explored the direct association of betweenness centrality scores to retweeting activity.

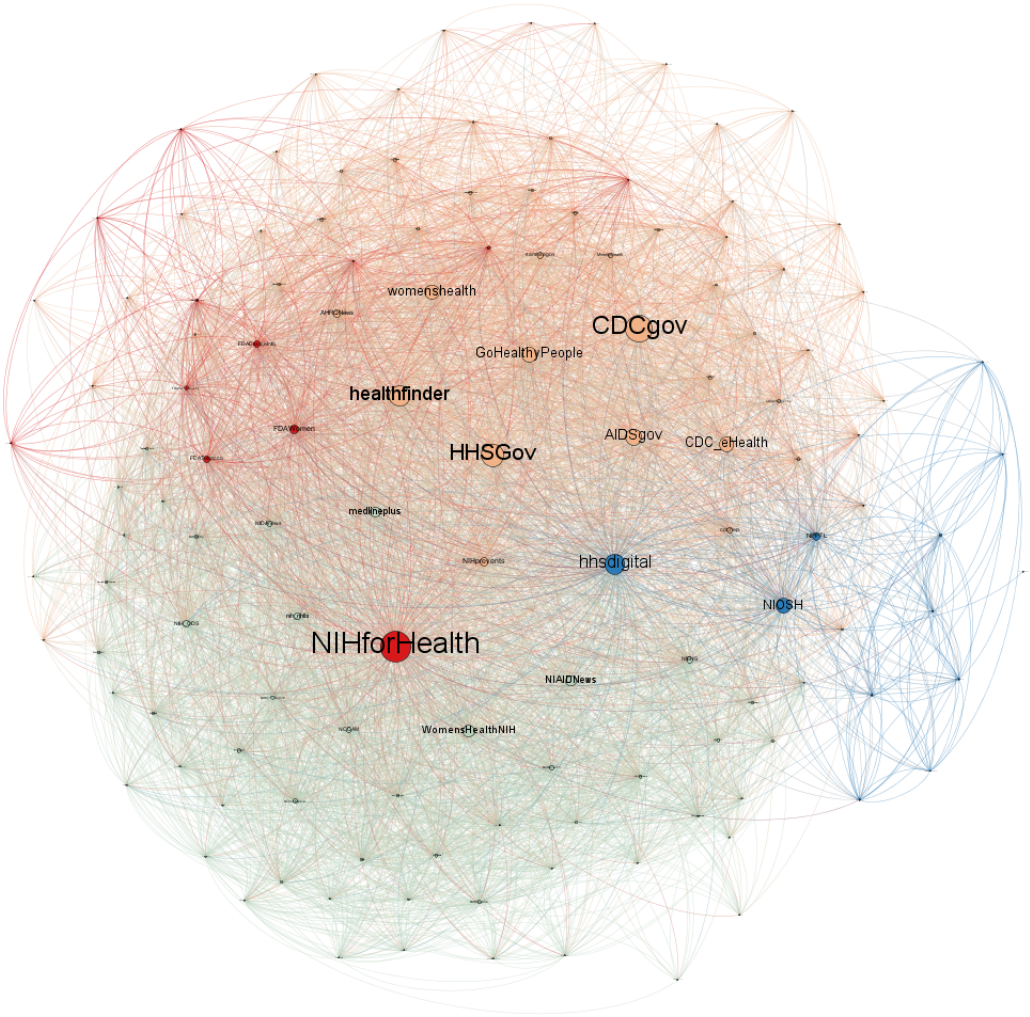


Figure 4.3: Graph displaying the betweenness-centrality for various agency handles (color-coded communities are also shown in the graph)

4.1.2.2 Tweet-specific Features

4.1.2.2.1 Tweet Age

It may be that the age of the handle when the tweet is posted relates to engagement. Handles that have been around longer might perhaps have a greater chance of exposure to users; this in turn might positively relate to the retweet rate. Thus we also study age of the twitter handle in relation to when the tweet is posted. We determine this by the number of days between the day the account was created to the day the tweet was written. Of note here is that previous studies have found age to be significantly correlated to retweetability [72].

4.1.2.2.2 URLs, Hashtags, User-mentions etc.

We will study whether hashtags (words prefixed with #), URLs (http, ftp, etc.) and user-mentions (words prefixed with @) are associated with engagement. An overwhelming portion, 75% of tweets (123,379) in our dataset contain URLs signifying the frequent use of tweets as gateways to external resources. This is in sharp contrast to previous research where only 19% [88] to 21% [72] of tweets were found to contain URLs. We speculate that this abundance of URLs for tweets from health agencies may be because in health communications references to sources and supporting materials are necessary. Elaborations cannot be provided in a short-text tweet. This is supported by another study on the use of twitter by local health departments where the authors found 74% of tweets contain URLs [59]. As regards hashtags and user-mentions, 93,031 tweets (around 57%) of our tweets contain hashtags and 65,180

(38%) contain user-mentions. These percentages are also considerably higher compared to [59], where the authors found hashtags in 16% of tweets and user-mentions in 20% of tweets. It is not clear how these features, appearing more frequently in our agency dataset than in general, might associate with engagement. Are tweets with URLs (or hash-tags or user-mentions) associated with higher levels of retweeting or longer retweet timelines? Previous research, such as [72] and [71] found that URLs and hashtags have significant positive associations with retweetability. The first study also found that user-mention had marginally significant negative association with retweetability. But given that these devices are much more prominent in tweets from our agencies when compared to the general domain, it is important to address these questions in the current context.

4.1.2.2.3 Tweet Sentiment

Tweet sentiment may also be associated with engagement. Perhaps more positive sentiment is linked with greater retweeting activity, or maybe the reverse holds. So we ask: Is there a difference in the response to a positive, negative, or neutral tweet? We analyze sentiment using a state-of-the-art lexicon-based sentiment classifier, SentiStrength[6] [77]. SentiStrength has been widely applied for sentiment analysis of tweets [76] and has been shown to outperform other lexical classifiers [60]. SentiStrength classifies each tweet into positive and negative sentiments on a scale of +/-1 (neutral) to +/-5 (extreme). Table 4.5 shows the distribution of tweets across

[6] sentistrength.wlv.ac.uk/

these sentiment scales. We find that in general slightly more tweets are classified as negative (percentage of moderate to extreme negative is 32.2% while for positive this percentage is 28.3%). Positive and negative sentiments are two different independent variables.

Researchers have studied the influence of sentiment on retweetability. Recent studies [57] have shown that tweets with negative valence values or negative emoticons are more likely to get retweeted. However, taking into account Twitter's role as both a social network and news media [44], researchers have found that negative news content as well as positive non-news content are both likely to be retweeted [34]. Again, we are interested in sentiment in the context of tweets posted by US health care agencies and not by individual users.

Table 4.5: Distribution of positive and negative sentiments for tweets on a 5-point scale.

Sentiment level	**# of positive tweets**	**# of negative tweets**
neutral	117599 (71.66%)	111233 (67.78%)
moderate-medium	36940 (22.51%)	31791 (19.37%)
medium	8502 (5.18%)	10143 (6.18%)
medium-extreme	1051 (0.64%)	10772 (6.56%)
extreme	12 (0.01%)	165 (0.10%)
Total	164104	164104

4.1.2.2.4 Tweet Semantics

One aspect of tweet analysis that is often overlooked is the content of the tweets. Content is important as it focuses on the subject matter of the tweet. It is possible that some subjects are more attractive than others to a broad audience. For example, a tweet about an emergency situation ("RT @fema: #Sandy East coast, search for open shelters by texting: SHELTER + a zip code to 43362 (4FEMA). Ex: Shelter 01234 (std rates apply)") got far more retweets compared to a job posting from NIH ("#NIH has unique & fascinating #sciencejobs working on the economics of aging as a #Health Scientist Administrator http://bit.ly/boDkwf").

It is highly challenging to build a generic, i.e., domain independent, content analyzer; this is likely a reason why content analysis is often omitted in Twitter-based studies. Domain or topic-specific tweets have been analyzed to identify the nature of tweet-based communications. For example, 5,395 tweets posted during the 2009 H1N1 outbreak were manually categorized into 6 categories such as resources, personal experiences and opinions, jokes, marketing and spam [17]. Other examples of content analysis on Twitter include analysis of tweeting behavior of professional athletes [33] and the manual coding of 1,000 concussion-related tweets along 9 board themes [73]. There are a few limitations to this type of content analysis. First, the manual analysis process limits the number of tweets that can be coded using the above methods. The additional overhead of time and cost is also a limiting factor. Second, and perhaps more important is that the selection of pre-defined coding categories limits the number of possible coding outcomes for a tweet and it ignores spontaneously

generated categories of interest.

In contrast we propose a fully automated method for content analysis of the 164,104 tweets in our dataset. We use the National Library of Medicine's Medical Text Indexer (MTI)[7] [5] for assigning Medical Subject Headings (MeSH) [48, 7] recommendations to each tweet. MTI is commonly used for recommending MeSH terms to titles and abstracts of biomedical literature and has been shown to be useful in other domains such as clinical text [4]. We show a novel application of MTI in the social media domain.

Consider the example tweet: "News: Weight loss does not lower heart disease risk from type 2 diabetes http://t.co/DNpukx6j" from nihforhealth. MTI assigns the following five MeSH terms: Weight Loss (T045556), Diabetes Mellitus, Type 2 (T011751), Obesity (T029022), Body Weight (T005291), and Heart Diseases (T019184). The MeSH identifiers in parenthesis are then mapped to Concept Unique Identifiers (CUIs) and corresponding semantic types are identified using the Unified Medical Language System (UMLS) Metathesaurus [11]. For example, Weight Loss (T045556) is mapped to the semantic type "Finding". Diabetes Mellitus, Type 2 (T011751), Obesity (T029022) and Heart Diseases (T019184) are mapped to the semantic type "Disease or Syndrome". Body Weight (T005291) is mapped to the semantic type "Organism Attribute". The semantic types are next mapped to semantic groups for higher level abstraction of the semantic types[8] [50]. Thus 134

[7] ii.nlm.nih.gov/mti.shtml

[8] http://semanticnetwork.nlm.nih.gov/SemGroups/SemGroups.txt

UMLS semantic types are reduced to 15 semantic groups. Table 4.6 shows the 15 semantic groups with examples of component semantic types and their prevalence in our dataset. Note that a particular tweet can be classified into multiple semantic groups. "Concept & Ideas" is the most prevalent semantic type in our dataset with 42% tweets containing terms that correspond to this semantic group. "Genes & Molecular Sequences" is the rarest semantic group with only 0.69% tweets containing terms corresponding to this semantic type. This is understandable since health agencies are more likely to discuss concepts and ideas or disorders than amino acid and carbohydrate sequences on social media.

4.1.2.3 News

A feature that we also considered is the occurrence of a news item in traditional media related to the agency's tweet. It may be that if a tweet is accompanied, either just before or immediately after, by a news item on the same topic then it promotes retweeting. The influence of traditional news sources on social media has been extensively studied [90, 38, 52] but not in the health domain. We collected Google Health News[9] headlines using its RSS feed twice a day (12 AM and 12 PM) from 11/11/2011 until one week past the last date of tweet collected in our dataset. The dataset comprised 7192 unique news headlines. Even with a fairly low cosine similarity threshold of 0.3, we were able to find matching news items for only 1601 tweets (<1% of the total). Of these, the tweets and the news appear on the same

[9]https://news.google.com/news/section?ned=us&topic=m

Table 4.6: Semantic groups with examples of component semantic types and their prevalence in the dataset.

Semantic Groups	**Example Semantic Types**	**# of tweets (%)**
Concepts & Ideas	Functional Concept, Regulation or Law, Temporal Concept, etc.	68391 (41.68%)
Disorders	Anatomical Abnormality, Disease or Syndrome, Neoplastic Process, etc.	59164 (36.05%)
Living Beings	Mammal, Eukaryote, Plant, etc.	57836 (35.24%)
Geographic Areas	Geographic Area	42133 (25.67%)
Chemicals & Drugs	Clinical Drug, Organic Chemical, Enzyme, etc.	39065 (23.81%)
Activities & Behaviors	Daily or Recreational Activity, Machine Activity, Social Behavior, etc.	38276 (23.32%)
Organizations	Health Care Related Organization, Professional Society, Self-help or Relief Organization	35163 (21.43%)
Physiology	Cell Function, Mental Process, Organ or Tissue Function, etc.	32308 (19.69%)
Objects	Entity, Food, Manufactured Object, etc.	23452 (14.29%)
Procedures	Diagnostic Procedure, Research Activity, Therapeutic or Preventive Procedure, etc.	23445 (14.29%)
Phenomena	Biologic Function, Human-caused Phenomenon or Process, Natural Phenomenon or Process	20252 (12.34%)
Anatomy	Anatomical Structure, Cell Component, Tissue, etc.	7925 (4.83%)
Occupations	Biomedical Occupation or Discipline, Occupation or Discipline	7633 (4.65%)
Devices	Drug Delivery Device, Medical Device, Research Device	1610 (0.98%)
Genes & Molecular Sequences	Amino Acid Sequence, Carbohydrate Sequence, Gene or Genome, etc.	1138 (0.69%)

day in 320 cases, the tweets precede the news in 610 cases (average lag 2.7 days) and news precedes the tweet in 671 cases (average lag 3.3 days). Because of paucity of data on matching news items, we do not consider this feature further.

4.1.3 Modeling Retweet Count using Hurdle Model

4.1.3.1 Choice of Model

As seen in Figure 4.1(a), the number of retweets per tweet in our dataset is highly skewed. There is an abundance of tweets with zero retweets (33%) and few tweets (less than 1%) with very high retweet count. This type of data distribution where the variance (1346.11) is much greater than the mean (6.09) is described as overdispersed data [79] with zero-inflation [16]. Typically linear models such as Poisson or negative binomial regression are used to model count data. However the zero-inflation and overdispersion ($p<0.001$) of the retweet count necessitates the use of two-part count data models such as the hurdle regression model [13, 85, 55].

Hurdle models use two separate components: a zero-portion used to fit the inflation of zero counts in the data and a count-portion to fit the non-zero counts of the data. The zero-portion of the hurdle model determines the binary outcome of whether a count is zero (no retweets) or not using a binomial probability model. The count portion of the model determines the conditional distribution of the non-zero count of the data using a zero-truncated negative binomial or Poisson model.

We formally compare different count data regression models (namely, the Poisson (P), negative binomial (NB), hurdle Poisson (HP) and hurdle negative binomial

(HNB)) using standard goodness-of-fit measures. The likelihood ratio test (LRT) is used to compare full and nested models (e.g. NB vs. P and HNB vs. HP). The Akaike information criterion (AIC) and Vuong statistics are used to compute goodness of fit for all pairs of non-nested models (e.g. NB vs. HP, etc.).

Table 4.7: Comparison of various count data regression models.

	P	**NB**	**HP**	**HNB**
AIC	2649779	813296.6	2274348	800270.2
P	–	73.81 *** (1836484***)	14.14***	73.89 ***
NB		–	-60.08***	14.43***
HP			–	59.36*** (1474079***)
HNB				–

Table 4.7 shows our comparisons of P, NB, HP and HNB for modeling retweet counts. The first row shows the AIC values for the different models. Since lower AIC values imply better model fit, HNB is deemed to be the best model for fitting our data. The rest of the table shows model comparison in terms of Vuong statistics. For comparison between nested and non-nested models (P vs. NB and HP vs. HNB) LRT scores are shown in the parenthesis. Significant positive values for the Vuong and LRT statistics imply a better fit for the model in the column than the one in the row. For all tests considered, the hurdle negative binomial fit best. For example, Vuong statistics for HNB compared to P, NB and HP were 73.89, 14.43 and 59.36 respectively all significant at $p<0.001$.

An important assumption in multiple regression analysis is that the variables used in the statistical models are independent of each other i.e. multicollinearity should not exist among them. We use the variance inflation factor (VIF) to check for the presence of multicollinearity in our experiments. VIF scores for all independent variables in our regression analysis were within the range of zero to 5 indicating no multicollinearity issues.

Table 4.8 presents results from the hurdle regression model applied to our data. The regression coefficients in the zero-portion are exponentiated as odds ratios (OR) while the exponentiated regression coefficients in the count portion are treated as incident rate ratios (IRR) [25]. In the analysis we assume that all other variables remain constant while we interpret the results of a particular variable.

4.1.3.2 Analysis for Retweet Presence

The coefficients of the logit regression in the zero portion of the model indicate how the features relate to crossing the 'hurdle' of obtaining at least 1 retweet. For continuous variables such as log-transformed counts of favorites, followers, friends and status, a unit increase in these values might change the odds of a tweet being retweeted. For binary variables (hashtags, URLs, user mentions and each semantic group), the odds of getting at least one retweet is increased or decreased based on the presence of the feature compared to its absence.

A unit increase in the log-transformed tweet age or follower count or favorite count increases the odds of getting at least one retweet by 202.9%, 151.5%

Table 4.8: Results of hurdle negative binomial model for Twitter data.

	Zero Portion				**Count Portion**			
	Estimate (SE)	**OR**	**z value**	**p**	**Estimate (SE)**	**IRR**	**z value**	**p**
(Intercept)	-3.295 (0.05)	0.037	-65.69	***	-1.361 (0.053)	0.256	-25.89	***
LT Favorite Count	0.207 (0.009)	1.23	23.668	***	0.074 (0.009)	1.077	8.025	***
LT Follower Count	0.922 (0.012)	2.515	74.148	***	0.939 (0.011)	2.559	85.831	***
LT Friend Count	0.002 (0.012)	1.002	0.168		-0.181 (0.013)	0.835	-13.717	***
LT Status Count	-1.242 (0.02)	0.289	-61.38	***	-0.712 (0.019)	0.491	-37.481	***
LT betweenness-centrality	0.016 (0.01)	1.016	1.679		0.099 (0.01)	1.105	10.347	***
LT tweet age	1.108 (0.016)	3.029	69.973	***	0.12 (0.018)	1.128	6.539	***
Hashtag	0.386 (0.012)	1.471	32.662	***	-0.034 (0.011)	0.966	-3.01	**
URL	0.529 (0.014)	1.697	38.084	***	-0.08 (0.014)	0.923	-5.581	***
User-mention	0.229 (0.012)	1.257	18.355	***	0.869 (0.012)	2.385	72.131	***
Positive Sentiment	-0.08 (0.009)	0.923	-8.473	***	-0.016 (0.01)	0.984	-1.695	
Negative Sentiment	-0.141 (0.008)	0.868	-18.747	***	-0.056 (0.007)	0.945	-8.222	***
Activities & Behaviors	0.32 (0.014)	1.377	22.869	***	0.175 (0.013)	1.191	13.213	***
Anatomy	0.195 (0.028)	1.215	6.959	***	-0.05 (0.025)	0.951	-2.026	*
Chemicals & Drugs	0.105 (0.014)	1.11	7.675	***	0.141 (0.013)	1.151	10.82	***
Concepts & Ideas	0.235 (0.012)	1.265	19.933	***	-0.022 (0.011)	0.978	-1.94	
Devices	0.273 (0.059)	1.314	4.653	***	-0.226 (0.054)	0.797	-4.224	***
Disorders	0.278 (0.014)	1.32	20.516	***	0.177 (0.013)	1.193	13.909	***
Genes & Molecular Sequences	0.058 (0.071)	1.059	0.812		-0.952 (0.065)	0.386	-14.674	***
Geographic Areas	-0.037 (0.018)	0.964	-1.986	*	-0.324 (0.018)	0.723	-18.216	***
Living Beings	0.083 (0.012)	1.086	6.643	***	0.082 (0.012)	1.085	6.927	***
Objects	0.14 (0.017)	1.15	8.331	***	0.192 (0.016)	1.212	11.979	***
Occupations	-0.057 (0.027)	0.945	-2.146	*	-0.134 (0.027)	0.875	-5	***
Organizations	-0.107 (0.02)	0.899	-5.394	***	-0.24 (0.019)	0.786	-12.683	***
Phenomena	0.07 (0.018)	1.073	3.939	***	0.436 (0.017)	1.547	25.043	***
Physiology	0.188 (0.015)	1.207	12.465	***	0.311 (0.014)	1.365	22.106	***
Procedures	0.046 (0.017)	1.047	2.733	**	-0.082 (0.016)	0.921	-5.157	***
Log(theta)					-1.8 (0.024)	0.165	-73.547	***

Note: The Coefficient (SE), hazard ratio (HR), z and p-values (*p<0.05, **p<0.01, ***p<0.001) for various independent variables are shown.

(OR=2.515) and 23% respectively, all other variables remaining constant. On the contrary, a unit increase in log-transformed status count decreases the odds of a retweet by over 71.1%. The log-transformed friend count or of betweenness-centrality score for a handle is not associated with the odds of a tweet getting a retweet.

A unit increase in either negative or positive sentiment decreases the odds of getting a retweet by 13.2% and 7.7% respectively. The presence of a URL or of a hashtag or of a user mention are each linked to an increase in the odds of a tweet getting at least one retweet but at different rates: 69.7%, 47.1% and 25.7% respectively, all other variables remaining constant.

Eleven of the 15 semantic groups increase the odds of getting a retweet with the group "Activities & Behavior" showing the highest increase (37.7%) followed by "Disorders" (32%). "Organizations", "Occupations" and "Geographic Areas" are the three semantic groups that decrease the odds of getting a retweet by 10.1%, 5.5% and 3.6%, respectively.

4.1.3.3 Analysis for Retweet Abundance

We now analyze the coefficients of the Negative Binomial regression in the count portion of the hurdle model. This allows us to study factors related to the rate of retweeting for tweets that succeed in getting at least one retweet. For continuous variables such as log-transformed counts of favorites, followers, friends and status, the coefficients give us estimates of incidence rate ratio for a unit increase in their values. For binary variables (hashtags, URLs, user mentions and each semantic group), we

get estimates of the rate ratio of retweets based on the presence compared to the absence of each feature, holding other variables constant in the model.

Given a unit increase in the log-transformed follower count of a handle, the rate of retweeting is expected to increase by a factor of 2.559, while holding all other variable in the model constant. Similarly, given a unit increase in the log-transformed tweet age, betweenness-centrality score or favorite count we expect the rate of retweeting to increase by factors of 1.128, 1.105 and 1.077 respectively. A unit increase in the log-transformed number of friends or of status goes in the opposite direction, these are expected to decrease the rate of retweeting by a factor of 0.835 and 0.491 respectively.

For the sentiment, a unit increase in negative sentiment decreases the rate of retweeting by 0.945. Positive sentiment has no significant association with the abundance of retweets.

The presence of a user mention increases the expected rate of retweeting by a factor of 2.385 all other variables remaining constant. On the other hand, the presence of a hashtag or of a URL decreases the rate of retweeting by a factor of 0.966 and 0.923 respectively.

Of the 15 semantic groups only 7 have significant positive association with the rate of retweeting. The presence of the semantic group “Phenomena” increases the rate of retweeting by a factor of 1.557 (highest amongst the semantic groups) followed by “Physiology” which increases the rate of retweeting by a factor of 1.37. Of the 7 semantic groups having significant negative associations with the abundance

of retweets, “Genes and Molecular Sequences” has the largest decrease in the rate of retweeting with a factor of 0.386. Examples of other groups negatively associated are “Anatomy”, “Geographic Areas” and “Occupations”.

4.1.3.4 Analysis across Hurdle Components

Looking across both components of the hurdle model several features show consistent benefit for engagement. These include numbers of favourites and followers, tweet age and the inclusion of user-mentions. Emphasizing semantic groups such as Activities & Behaviour, Chemicals & Drugs, Disorders and Living Beings increase engagement. Sentiment in tweets almost always lowers engagement. So do certain groups such as Geographic Areas, Occupations and Organizations. Hashtags and URLs are important for crossing the initial hurdle of getting at least 1 retweet but then their presence dampens retweet rate. Status is consistently negatively related to getting retweets, strongly so in the case of getting at least 1 retweet.

4.1.4 Modeling Retweet Life Span

As evident from Figure 4.1(c), the time to first retweet can vary considerably. While retweets usually begin on the very day the tweet is posted, there are instances where this first retweet occurs after 10, 50, or even 100 days. Therefore the characteristics of a tweet that influence such behavior are of great interest. We use methods from survival analysis [26, 42], the branch of statistics dedicated to modeling such temporal behavior. Typically in survival analysis we build models to analyze “time to events” such as death of an organism or failure of a machine [47]. In our case

the "event" refers to the appearance of the first retweet or last retweet of a tweet. Similar to previous Twitter research [79] we use the Cox proportional hazards regression model [16] to estimate how the different handle and tweet-based features (see Table 4.4) correlate with the time to the first and last retweets. Note that the ideal scenario for propagation of a tweet arises with an early first retweet and a late last retweet giving a tweet a prolonged lifespan.

4.1.4.1 Modeling Time to First Retweet

Table 4.9 shows the results of Cox proportional hazards regression model for time to first retweet. The regression coefficients are exponentiated as hazard ratios (HR) and used in the interpretation of the survival models.

For continuous variables such as log-transformed counts of favorites, followers, friends and status, a unit increase in these values may change the time to first retweet with all other variables remaining constant. Note that a negative association means that the time to first retweet is increased. For binary variables (hashtags, URLs, user-mentions, and each semantic group) the time to first retweet may increase or decrease based on the presence of a feature compared to its absence in a tweet.

We find that a unit increase in the log-transformed follower count or tweet age or favorite count of a tweet handle decreases the time to first retweet by 10.7%, 9.3% and 5.6% respectively. On the other hand, a unit increase in the log-transformed number of friends increases the time to first retweet by 2.6%. Interestingly, a unit increase in the log-transformed status count or betweenness-centrality are not associated with

Table 4.9: Results of Cox proportional hazards model for interval between a tweet and its first retweet.

	Interval Between Tweet and First Retweet			
	Coefficient (SE)	**HR**	**z**	**p**
Log-transformed Favorite Count	0.055 (0.006)	1.056	8.72	***
Log-transformed Follower Count	0.102 (0.009)	1.107	11.029	***
Log-transformed Friend Count	-0.026 (0.009)	0.974	-2.929	**
Log-transformed betweenness	-0.004 (0.008)	0.995	-0.566	
Log-transformed Status Count	0.017 (0.015)	1.017	1.176	
Log-transformed tweet age	0.089 (0.012)	1.093	7.204	***
Hashtag	0.116 (0.009)	1.123	12.873	***
URL	-0.021 (0.011)	0.978	-1.907	.
User-mention	-0.072 (0.01)	0.930	-7.186	***
Positive Sentiment	-0.02 (0.007)	0.979	-2.807	**
Negative Sentiment	-0.001 (0.005)	0.998	-0.284	
Activities & Behaviors	0.008 (0.01)	1.008	0.827	
Anatomy	0.001 (0.019)	1.001	0.077	
Chemicals & Drugs	0.013 (0.01)	1.013	1.347	
Concepts & Ideas	0.008 (0.009)	1.008	0.954	
Devices	-0.009 (0.04)	0.990	-0.231	
Disorders	0.02 (0.01)	1.020	2.037	*
Genes & Molecular Sequences	0.051 (0.047)	1.052	1.085	
Geographic Areas	-0.033 (0.014)	0.967	-2.296	*
Living Beings	0.006 (0.009)	1.006	0.667	
Objects	0 (0.012)	0.999	-0.033	
Occupations	-0.02 (0.02)	0.980	-0.998	
Organizations	0.006 (0.015)	1.006	0.416	
Phenomena	0.021 (0.013)	1.021	1.596	
Physiology	0.009 (0.011)	1.009	0.827	
Procedures	0.024 (0.012)	1.024	2.006	*

Note: The Coefficient (SE), hazard ratio (HR), z and p-values (*p<0.05, **p<0.01, ***p<0.001) for various independent variables are shown.

the interval between a tweet and it's first retweet.

For the sentiment features we find that a tweet with positive sentiment is more likely to see a delay of 2.1% in time to first retweet while negative sentiment does not have any effect, with all other variables remaining constant.

We also find that hashtags in a tweet reduce the time to first retweet by 12.3% while the presence of user-mentions increase this time by 7%. Contrary to a previous finding [79] we do not find URLs having significant association with time to first retweet. We remind the reader that our health agency tweets have more URLs, hashtags and user-mentions than tweets in the general domain.

Amongst the 15 semantic groups we find that only 3 have significant influence on this interval. Semantic groups "Procedures" and "Disorders" seem beneficial to the time to first retweet by 2.4% and 2% respectively while "Geographic Areas" is the only one that increases the time to first retweet and this is by 3.3%.

4.1.4.2 Modeling Time to Last Retweet

Next, we study the degree to which the different features relate to the interval between a tweet and its last retweet. Table 4.10 shows the results of Cox proportional hazards regression model used for this purpose. It is important to note here that while a shorter interval is desirable for the first retweet (as discussed in the previous section), a longer interval is desirable for the last retweet. Thus the features with negative association (in red) are the beneficial ones.

We find that a unit increase in the number of followers increases the time

Table 4.10: Results of Cox proportional hazards model for interval between a tweet and its last retweet.

	Interval Between Tweet and Last Retweet			
	Coefficient (SE)	**HR**	**z**	**p**
Log-transformed Favorite Count	0.037 (0.006)	1.037	5.912	***
Log-transformed Follower Count	-0.27 (0.009)	0.763	-29.249	***
Log-transformed Friend Count	-0.009 (0.009)	0.991	-0.972	
Log-transformed Status Count	0.351 (0.015)	1.420	23.994	***
Log-transformed tweet age	-0.014 (0.012)	0.986	-1.124	
Log-transformed betweenness	0.036 (0.008)	1.036	4.832	***
Hashtag	0.139 (0.009)	1.149	15.519	***
URL	-0.179 (0.011)	0.835	-15.915	***
User-mention	0.094 (0.01)	1.098	9.355	***
Positive Sentiment	-0.025 (0.007)	0.975	-3.411	***
Negative Sentiment	0.037 (0.005)	1.037	7.049	***
Activities & Behaviors	-0.043 (0.01)	0.957	-4.265	***
Anatomy	-0.038 (0.019)	0.962	-1.961	*
Chemicals & Drugs	-0.019 (0.01)	0.981	-1.936	
Concepts & Ideas	-0.011 (0.009)	0.989	-1.262	
Devices	-0.059 (0.04)	0.942	-1.471	
Disorders	-0.011 (0.01)	0.988	-1.156	
Genes & Molecular Sequences	0.059 (0.047)	1.060	1.252	
Geographic Areas	0.001 (0.014)	1.000	0.04	
Living Beings	-0.006 (0.009)	0.993	-0.721	
Objects	-0.049 (0.012)	0.951	-3.993	***
Occupations	0.04 (0.02)	1.040	1.977	*
Organizations	0.041 (0.015)	1.041	2.687	**
Phenomena	-0.012 (0.013)	0.987	-0.928	
Physiology	-0.013 (0.011)	0.986	-1.189	
Procedures	0.017 (0.012)	1.016	1.388	

Note: The Coefficient (SE), hazard ratio (HR), z and p-values (*p<0.05, **p<0.01, ***p<0.001) are shown.

to last retweet by 23.7%. On the contrary, a unit increase in the log-transformed status count, favorite count, betweenness-centrality score (decrease the time to last retweet by 42%, 3.7% and 3.6% respectively. Clearly the effect of status count far outweighs the other two features. Tweet age and the number of friends do not have any significant association with the time to last retweet.

For sentiment features we find that a unit increase in positive sentiment increases the time to last retweet by 2.5% while a unit increase in negative sentiment decreases the time to last retweet by 3.8%.

For the binary variables we find that the presence of a URL increases the time to last retweet by 16.5% while the presence of a hashtag or a user-mention decrease the time interval by 14.9% and 9.8% respectively. Amongst the 15 semantic groups, only five have significant relation to the time to last retweet. Tweets containing semantic groups "Objects", "Activities & Behavior" and "Anatomy" are positively related to an increase in the time to last retweet of 4.9%, 4.3% and 3.8% respectively. "Organizations" and "Occupations" are the only ones that decrease the time to last retweet by 4.1% and 4% respectively.

4.1.4.3 Analysis across Life Span

Considering the two survival models together we note that only follower count consistently benefits both temporal aspects of engagement. Several others benefit one aspect while remaining neutral to the other. These include tweet age, URLs, and semantic groups such as Activities & Behaviors, Anatomy, Disorders and Objects.

User-mentions are consistently negative while features such as Hashtags, status, and friend count are negative in one aspect and neutral in the other. Status in particular stands out with its strong negative effect on time to last retweet.

4.1.5 Discussion

Our results show that although multiple federal health agencies are using Twitter, there is a great deal of difference between levels of Twitter use and also retweets. For public health agencies, we found that a tiny minority of tweets gets more than 100 retweets; a two-thirds majority of tweets get on average 8 retweets. We also found that a handle's follower count and favorite count have strong positive relationships with retweeting behavior. While these features are not easy for agencies to improve, they are easy metrics to follow. In contrast, we found that having more friends on Twitter was negatively associated with the number of times a tweet is retweeted. Early adoption of Twitter by an agency is associated with our measures of engagement. As a handle ages the chances for engagement overall seem to improve. This is not something that agencies can change but it does provide support for health agencies thinking about starting Twitter accounts to do just that and not to wait and delay getting started.

Agencies that generated more Tweets than others did not necessarily have more retweets. In fact, we found that status, the number of tweets posted overall, is negatively associated with retweets. This suggests that the agency might consider only tweeting posts that it regards as important so as to not 'dilute' the public's

attention.

Health agencies can augment their tweets by adding hashtags, URLs, or user-mentions and this may increase the likelihood that users will find the information encoded in the tweet more useful and thus retweet it. Indeed, we found that the addition of hashtags, URLs, or user-mentions did indeed increase the likelihood that a given tweet would be retweeted. However, the inclusion of hashtags and URLs is also associated with decreased numbers of retweets, and user-mentions are associated with shorter times to last retweet. Thus, agencies may be able to increase retweets by using these conventions, but they might not increase the longevity of tweets.

Our observations regarding hashtags, user-mentions and URLs are also interesting because of differences in their prevalence between our dataset and Twitter data in general. We speculate that this abundance of URLs for tweets from health agencies may be because in health communications references to sources and supporting materials are necessary. Hashtags and user-mentions are also more prevalent in our dataset appearing in 57% and 38% of agency tweets respectively, while in the general domain hashtags were found in only 16% and user-mentions in only 20% of tweets [59].

Betweenness-centrality is positively related to the number of retweets and negatively related to the time to last retweet. To the best of our knowledge, researchers have not explored the direct association of betweenness centrality scores to retweeting activity.

Much work has been done involving mining sentiment from Twitter and it has

previously been demonstrated that the presence of sentiment of one kind or the other is associated with higher rates of retweeting. In contrast, we found that sentiment in tweets from government agencies, either positive or negative, is not associated with retweeting. It should also be noted that agency tweets are predominantly neutral (70%).

Semantic groups have not been studied in the context of retweet rates. We found that posts about activities and behaviors, chemicals and drugs, disorders, living beings, objects, phenomenon and physiology are positively associated with engagement. In contrast, posts about organizations, occupations, genes & sequences and geographic areas tend to lower engagement. But it may also be that the intent behind such posts are less to engage and more to just inform.

4.1.6 Predictive Modeling of Retweets

In addition to the count data regression models we also built predictive models for retweet counts. First, we built a classifier for predicting whether a tweet will generate a retweet or not. Second, we built various regressors for predicting the number of retweets for tweets that get retweeted.

Using the same features as shown in Table 4.4 we built various classifiers, namely naïve Bayes (NB), decision tree (J48) and Support Vector Machine (SVM), for predicting the retweetablity of a tweet. The results of these classifiers using standard evaluation metrics in 10-fold cross-validation experiments are shown in Table 4.11. We find that J48 outperforms the other methods in terms of precision (0.784), recall

(0.789) and F-score (0.785).

For modeling the count of retweets we implemented various regressors, namely Support Vector Regression (SVR), M5P (a decision tree with linear regression functions at the nodes)[65] and ZeroR. As before, we use the same set of features as listed in Table 4.4. The results of these regressors using standard evaluation metrics in 10-fold cross validation experiments are shown in Table 4.12. MAE measures the average absolute deviation between a predicted retweet count and the actual count. RMSE, which squares the error before summing it amplifies the presence of larger errors. Lower values of both these metrics signify better performance while the opposite is true for correlation coefficients. We find that SVR outperforms the other methods in terms of higher correlation coefficient and lower mean absolute error (MAE) and root mean squared error (RMSE).

We note that while it's relatively easier to classify tweets that are retweeted vs. those that are not, estimating the number of actual retweets for a tweet is a much more challenging task. Even the best regression technique in our experiment (SVR) perform quite poorly with low correlation and high MAE and RMSE. We believe that typical regressors such as those used in this experiment might not be well suited for modeling the long-tailed retweet counts. Instead, count data regression techniques such as negative binomial considered for modeling this data (Section 4.1.3) seems better suited.

Table 4.11: Results of classifiers for predicting retweetability

Algorithms	**Precision**	**Recall**	**F-score**
NB	0.676	0.603	0.615
J48	0.784	0.789	0.785
SVM	0.779	0.786	0.778

Table 4.12: Results of regressors for predicting retweet counts

Algorithms	**Correlation coefficient**	**MAE**	**RMSE**
ZeroR	-0.0127	9.2633	44.4026
M5P	0.1098	7.8693	45.2349
SVR	0.1726	6.3726	44.1741

4.1.7 Conclusions

To the best of our knowledge, our study is the first comprehensive analyses of Twitter engagement by public health agencies. The level of Twitter activity varies greatly by health agency: some health accounts are very active and others are not as much. However, it seems to be the content of the tweets (e.g., about activities and behaviors, disorders) and not the number of tweets alone that is associated with a higher level of engagement (number of re-tweets). Furthermore, although some of the factors associated with more engagement cannot be changed by the agency (e.g., the length of time they have been active on Twitter), several factors associated with higher re-tweets can be controlled (e.g., use of hashtags, URLs). Predictive modeling of retweets show that while it is relatively easier to classify tweets that get retweet vs. those that don't, predicting the actual retweet count is a challenging task. Our results provide a framework for future experiments designed to improve the public's

engagement with health agencies via Twitter.

4.2 Engagement of Health Organizations in Facebook

In the previous section we studied engagement on Twitter. Many of these agencies also have substantial presence on Facebook. Facebook presents not just another platform for studying engagement but also can provide critical cross-platform insights. Studying a second social media platform, in addition to Twitter, will help in identifying critical differences between the two platforms.

The methods adopted in this study parallel those of the Twitter-based study. Instead of retweets we use the sum of likes, shares and comments (referred to as 'activity' hereafter) as a measure of engagement in Facebook.

We address the following two questions with respect to Facebook posts from US Federal Health agencies and their responses. First, which features are associated with the level of response in the form of activity? Second, which features are associated with the interval between an agency's Facebook post and the last activity it generates?[10] We address our goals by analyzing an almost comprehensive set of Facebook posts from 72 Facebook accounts of 24 Federal Health Agencies, 19 of which are also present on Twitter. We explore associations between factors with level of activity using hurdle models. We explore the temporal factors related to our second question using survival model. Factors we examine include standard features such as the number of page likes as well as less studied features relating to the semantic

[10]Since the time to first activity for Facebook posts was unavailable, we cannot study the factors associated with it as we did with Twitter

content of a post.

4.2.1 Data Collection

4.2.1.1 Agencies & Accounts

Similar to the Twitter dataset, we selected health agencies through the HHS Social Hub website[11] which lists all Facebook accounts affiliated to various Federal health agencies. A total of 72 Facebook accounts were identified which correspond to 24 different health agencies, 17 of which are NIH division such as NIH/NIDA, NIH/NIMH and NIH/NICHD. Some agencies have quite a few accounts such as NIH/NLM (6 accounts: Women's_Health_Resources, NLM_4_Caregivers, etc.), CDC (10 accounts: CDC_Tobacco_Free, Health_Hazard_Evaluation_Program, etc.), OS (16 accounts: HealthCare.gov, Medical_Reserve_Corps, etc.) and several others have just one account such as ACF, FDA, NIH/NCCAM, etc. Table 4.13 lists the various agencies, the number of accounts for each and few examples of accounts.

4.2.1.2 Posts & Activity

The Facebook Graph API[12] was used to collect all posts from an account's timeline as of late January 2013. All posts from an account's timeline, starting from the account creation date, can be retrieved using this method. In contrast, in Twitter, as stated earlier we could only retrieve the most recent 3200 tweets from a user's timeline. For each Facebook post, we recorded its unique identifier, number of

[11] http://www.hhs.gov/socialhub/

[12] https://developers.facebook.com/docs/graph-api

Table 4.13: Agencies and accounts on Facebook (agencies also present on Twitter are marked with an asterisks).

Agency	**Name**	**# accounts**	**Examples of accounts**
ACF*	Administration for Children & Families	1	Child_Welfare_Information_Gateway
AoA	Administration on Aging	2	Administration_on_Aging, etc.
CDC*	Center for Disease Control & Prevention	10	CDC_Tobacco_Free, etc.
FDA*	U.S. Food & Drug Administration	1	U.S._Food_and_Drug_Administration
HRSA*	Health Resources & Services Administration	2	Health_Resources_and_Service_Administration_(HRSA), etc.
NIH*	National Institutes of Health	8	Fogarty_International_Center, etc.
NIH/NCCAM*	National Center for Complementary & Alternative Medicine	1	National_Center_for_Complementary_and_Alternative_Medicine
NIH/NCI*	National Cancer Institute	3	National_Cancer_Institute, etc.
NIH/NEI*	National Eye Institute	1	National_Eye_Health_Education_Program_(NEHEP)
NIH/NHGRI*	National Human Genome Research Institute	1	National_DNA_Day
NIH/NHLBI*	National Heart, Blood & Lung Institute	4	National_Heart,_Lung,_and_Blood_Institute_(NHLBI), etc.
NIH/NIAID*	National Institute of Allergy & Infectious Diseases	1	National_Institute_of_Allergy_and_Infectious_Diseases_(NIAID)
NIH/NIAMS*	National Institute of Arthritis & Musculoskeletal & Skin Diseases	2	National_Institute_of_Arthritis_and_Musculoskeletal_and_Skin_Diseases_Labs, etc.
NIH/NICHD	National Institute of Child Health and Human Development	1	Eunice_Kennedy_Shriver_National_Institute_of_Child_Health_and_Human_Development
NIH/NIDA*	National Institute of Drug Abuse	2	National_Institute_on_Drug_Abuse_(NIDA), etc.
NIH/NIDDK	National Institute of Diabetes and Digestive and Kidney Diseases	3	National_Diabetes_Education_Program_(NDEP), etc.
NIH/NIEHS*	National Institute of Environmental Health Sciences	1	National_Institute_of_Environmental_Health_Sciences
NIH/NIGMS*	National Institute of General Medical Sciences	1	National_Institute_of_General_Medical_Sciences
NIH/NIMH*	National Institute of Mental Health	1	National_Institute_of_Mental_Health
NIH/NINDS	National Institute of Neurological Disorders and Stroke	1	Know_Stroke
NIH/NLM*	National Library of Medicine	6	Women's_Health_Resources, NLM_4_Caregivers, etc.
NIH/OBSSR	NIH Office of Behavioral and Social Sciences Research	1	The_Office_of_Behavioral_and_Social_Sciences_Research_(OBSSR)
OS*	Office of the Secretary	16	Best_Bones_Forever!, U.S._Public_Health_Service_Pharmacists, etc.
SAMHSA*	The Substance Abuse & Mental Health Services	2	Disaster_Distress_Helpline, SAMHSA
Grand Total		72	

Table 4.14: Posts and activities per agency on Facebook.

Agency	#posts	# posts with zero activity	# posts with at least 1 activity	# likes	# shares	# comments	# total activity	# activity per post	# activity per non-zero activity post
ACF	372	21 (5.65%)	351 (94.35%)	2235	647	265	3147	8.46	8.97
AoA	1878	320 (17.04%)	1558 (82.96%)	5138	3381	363	8882	4.73	5.70
CDC	7313	1149 (15.71%)	6164 (84.29%)	253607	118644	35659	407910	55.78	66.18
FDA	538	119 (22.12%)	419 (77.88%)	12008	6321	6085	24414	45.38	58.27
HRSA	2456	609 (24.8%)	1847 (75.2%)	8203	1306	2092	11601	4.72	6.28
NIH	2831	738 (26.07%)	2093 (73.93%)	27391	10012	1985	39388	13.91	18.82
NIH/NCCAM	659	79 (11.99%)	580 (88.01%)	5803	2338	510	8651	13.13	14.92
NIH/NCI	3455	585 (16.93%)	2870 (83.07%)	27685	4429	5475	37589	10.88	13.10
NIH/NEI	447	87 (19.46%)	360 (80.54%)	1799	1860	86	3745	8.38	10.40
NIH/NHGRI	417	25 (6%)	392 (94%)	5226	1613	409	7248	17.38	18.49
NIH/NHLBI	3510	524 (14.93%)	2986 (85.07%)	82420	26606	6078	115104	32.79	38.55
NIH/NIAID	632	114 (18.04%)	518 (81.96%)	2811	383	181	3375	5.34	6.52
NIH/NIAMS	414	44 (10.63%)	370 (89.37%)	1165	128	63	1356	3.28	3.66
NIH/NICHD	332	40 (12.05%)	292 (87.95%)	762	192	48	1002	3.02	3.43
NIH/NIDA	1657	177 (10.68%)	1480 (89.32%)	13772	11423	1232	26427	15.95	17.86
NIH/NIDDK	1720	451 (26.22%)	1269 (73.78%)	4702	1239	785	6726	3.91	5.30
NIH/NIEHS	148	47 (31.76%)	101 (68.24%)	287	90	41	418	2.82	4.14
NIH/NIGMS	236	53 (22.46%)	183 (77.54%)	1191	222	166	1579	6.69	8.63
NIH/NIMH	427	23 (5.39%)	404 (94.61%)	13130	6574	1752	21456	50.25	53.11
NIH/NINDS	83	17 (20.48%)	66 (79.52%)	427	121	86	634	7.64	9.61
NIH/NLM	4076	1695 (41.58%)	2381 (58.42%)	24280	5861	1903	32044	7.86	13.46
NIH/OBSSR	188	75 (39.89%)	113 (60.11%)	212	55	26	293	1.56	2.59
OS	9158	1233 (13.46%)	7925 (86.54%)	172550	57372	28281	258203	28.19	32.58
SAMHSA	2915	761 (26.11%)	2154 (73.89%)	25657	11059	3089	39805	13.66	18.48
Total	45862	8986 (19.59%)	36876 (80.41%)	692461	271876	96660	1060997	23.13	28.77
Mean (SD)	1910.92 (2327.30)	374.42 (459.42)	1536.50 (1947.75)	28852.54 (60566.59)	11328.17 (25970.74)	4027.50 (8879.44)	44208.21 (94905.29)	15.24 (15.67)	18.29 (18.17)

likes, shares, comments and other metadata as described below (Table 4.14).

A total of 45,867 posts were collected from the timelines of the 72 accounts. Around 20% of the posts (8,986) had no likes, shares or comments i.e. no activity[13]. Only 2245 posts (4.8%) had 100 or more total shares, likes and comments (total activity = 547,476, mean = 243.8)[14]. The remaining three-fourths (34,631) of posts fell between these ranges (total activity= 513,521, mean = 14.8). Compared to this, tweets, in general, were less likely to get retweeted and get fewer retweets with only 1% having more than 100 retweets.

In raw numbers we find that the Office of the Secretary (OS) had the highest number of posts (9,158) with most of its posts (7,925) being liked, shared or commented. On the other hand, the CDC (7,313) with the second highest number of posts, also gets the most activity on aggregate (407,910) as well as per post (55.78). The NLM had the highest number of posts with no activity (1,695). Table 4.15 shows the top 10 accounts ranked by activity per post. These are: Let's _Move, StopBullying.Gov, Million _Hearts, CDC _Tobacco _Free, CDC, The_Heart _Truth, National _Institutes _of _Health _(NIH), U.S._Food _and _Drug _Administration, National_Institute_of _Mental _Health, and NCBI _-_National _Center _for _Biotechnology _Information.

[13]9,889 (21.5%) posts had no likes, 31,699 (69.1%) had no shares and 30,160 (65.%) had no comments

[14]The highest number of likes, shares and comments for a post were 8436, 1070 and 7552, respectively

Table 4.15: Top 10 accounts with most activity per Facebook post.

Account	**# posts**	**# posts with non-zero activity**	**# posts with zero activity**	**# likes**	**# shares**	**# comments**	**# total activity**	**# activities per non-zero activity post**
Let's_Move	457	446	11	73144	23535	13117	109796	246.18
StopBullying.Gov	173	168	5	21882	9583	4788	36253	215.79
Million_Hearts	488	432	56	36041	13515	2204	51760	119.81
CDC_Tobacco_Free	457	317	140	15315	17355	1803	34473	108.75
CDC	2867	2667	200	177302	78890	29155	285347	106.99
The_Heart_Truth	1056	879	177	61843	21387	3733	86963	98.93
National_Institutes_of_Health_(NIH)	427	408	19	17522	8885	947	27354	67.04
U.S._Food_and_Drug_Administration	538	419	119	12008	6321	6085	24414	58.27
National_Institute_of_Mental_Health	427	404	23	13130	6574	1752	21456	53.11
NCBI_-_National_Center_for _Biotechnology_Information	298	260	38	9658	1930	619	12207	46.95

4.2.2 Facebook Features

Similar to the Twitter-based study, we first decided which features we would use to represent each post. We included those that are generally used in Facebook-based studies as well as those that are seldom considered. Table 4.16 lists 4 features that we consider in our study.

Table 4.16: Facebook features examined.

Features	**Description**
Page likes	# of Facebook users liking a page (log-transformed). Note that this is different from a post like which is considered as an activity.
Post type	Classification of posts into 6 categories such as link, photo, etc.
Sentiment	Two scores: one for positivity and the other for negativity
Content (Semantic Groups)	Classification of each posts into 15 semantic groups using MTI followed by post-processing. Multiple classes per post allowed.

We note that the number of features are considerably less than that considered

in the Twitter-based study (11). This is primarily because of the nature of Facebook posts as well as due to restrictions imposed by the Facebook Graph API. For example, the only account level feature that applies to Facebook is the number of page likes, while other features are either not applicable (e.g. # followers, # friends) or unavailable (e.g. betweenness-centrality). In contrast to tweet-based features such as hashtag, URL or user-mention, Facebook posts are classified into 6 types such as link, photo, video, etc. Other features that overlap with those considered in the Twitter-based study are sentiment and content (semantic groups).

4.2.2.1 Page Likes

The number of page likes shows the number of users that endorse a particular account. A page like is different from a post like which is considered as an engagement activity. Users liking a page receives all posts from an account in their news feeds. Table 4.17 shows the top 10 accounts with the most page likes. The CDC has the highest number of page likes (241,342) followed by Let's_Move (115,940).

4.2.2.2 Post Type

The Faceook Graph API provides information about the type of a particular post. Posts are classified into 6 self-explanatory categories, namely link, music, photo, question, status[15], and video/ SWF[16]. Table 4.18 shows the various types of post as well as their counts. Links are the most common (28,833) type of posts while questions

[15]A post is a status if it does not belong to any of the other categories. It is simply a text-based post.

[16]Adobe's ShockWave Flash format

Table 4.17: Facebook page likes.

Account	**Page likes**
CDC	241342
Let's_Move	115940
Million_Hearts	53728
StopBullying.Gov	49721
U.S._Food_and_Drug_Administration	43240
NCBI_-_National_Center_for_Biotechnology_Information	43201
National_Institutes_of_Health_(NIH)	35054
The_Heart_Truth	34012
National_Institute_of_Mental_Health	32484
CDC_en_Espaniol	20923

are the least common (74).

Table 4.18: Count of various post types.

Post type	**# of posts**
link	28833 (62.9%)
status	9121 (19.8%)
photo	6429 (14.1%)
video/swf	1334 (2.8%)
music	76 (0.2%)
question	74 (0.2%)

4.2.2.3 Facebook Post Sentiment

Similar to the Twitter-based study, we hypothesize that sentiment of Facebook posts may be associated with engagement. Using SentiStrength to measure the senti-

ment of a Facebook post[17] we find that more posts are sentiment laden on Facebook compared to Twitter (Table 4.19). This is in sharp contrast to our Twitter-based findings where significantly more tweets lacked sentiment and were categorized as neutral. We speculate that these differences may be because of the nature of Facebook posts compared to tweets. Facebook posts being more verbose tend to use more words that imply sentiment compared to tweets which are more succinct. Tweets may often lack sentiment because instead of elaborating on a topic (because of 140 character limit) they tend to refer to external resources such as webpages for such discussions. Facebook posts have the scope of elaborating on topics without referring to external resources. We would like to explore more in these directions in future research.

Facebook posts are also generally classified as positive (percentage of moderate to extreme positive is 61.89% while for negative this percentage is 47.04%).We speculate that positive posts generate greater readership and engagement compared to negative posts on Facebook. We verify this in our experiments using the statistical models later in this chapter.

4.2.2.4 Facebook Post Semantics

One aspect of Facebook analysis that is often overlooked is the content of the post. As with tweets, we hypothesize that some topics are more attractive that others. For example, a post about information dissemination of the outbreak of West

[17]5 posts in our dataset did not have any associated textual information and hence could not be analyzed using SentiStrength.

Table 4.19: Distribution of positive and negative sentiments for Facebook posts on a 5-point scale.

Sentiment level	**# of positive posts**	**# of negative posts**
neutral	17477 (38.11)	24281 (52.94)
moderate-medium	22846 (49.81)	10426 (22.73)
medium	4625 (10.08)	5267 (11.48)
medium-extreme	905 (1.97)	5673 (12.37)
extreme	9 (0.02)	215 (0.47)
Total	45862	45862

Nile virus (*"West Nile virus is a potentially serious illness. What you need to know: http://go.usa.gov/r9g4"*) generated far more activity compared to a job posting from U.S._Public_Health_Service_Nurses (*"National Park Service has a Registered Nurse Manager position open in Yosemite, CA. This position closes on November 19. If interested, please send a cover letter and CV to S**** C**** at email@nps.gov."*).

Similar to our Twitter-based study we use the NLM's MTI program for assigning MeSH recommendations. These were then mapped to semantic types and subsequently to semantic groups using the methods outlined in Section 4.1.2.2.4. Table 4.20 shows the 15 semantic groups and their prevalence in our Facebook dataset. Note that a particular post can be classified into multiple semantic groups. Similar to the Twitter analysis we find that "Concepts & Ideas" is the most prevalent semantic type in our dataset with 54.34% posts containing terms that correspond to this semantic group. "Devices" and "Genes & Molecular Sequences" are the rarest semantic group in both social media platforms. Less than 1% of posts mention terms corresponding to "Devices" and "Genes & Molecular Sequences", respectively.

Table 4.20: Semantic groups and their prevalence in the Facebook dataset.

Semantic Groups	**# of posts (%)**
Concepts & Ideas	24922 (54.34)
Living Beings	22733 (49.56)
Geographic Areas	19891 (43.37)
Disorders	19826 (43.22)
Organizations	19299 (42.08)
Activities & Behaviors	15072 (32.86)
Physiology	14158 (30.87)
Chemicals & Drugs	9549 (20.82)
Procedures	9223 (20.11)
Objects	9034 (19.7)
Phenomena	6784 (14.79)
Occupations	4367 (9.52)
Anatomy	3731 (8.13)
Genes & Molecular Sequences	406 (0.89)
Devices	364 (0.79)

4.2.3 Modeling Activity using Hurdle Model

4.2.3.1 Choice of Model

As mentioned before, around 20% of Facebook posts don't get any likes, shares or comments i.e. have zero activity. The variance (14053.43) is much greater than the mean (23.13) implying overdispersed data with zero-inflation. Similar to the Twitter-based analysis we use the hurdle regression model for this analysis.

Table 4.21 shows our comparisons of P, NB, HP and HNB for modeling activity counts. With the lowest AIC value, HNB is deemed to be the best model for fitting our data. The same is true while comparing HNB with other models using Vuong and LRT statistics.

Table 4.21: Comparison of count data regression models for Facebook data.

	P	**NB**	**HP**	**NBH**
AIC	1443334	304590.7	304590.7	297667.3
P	–	58.57083*** (1138746***)	35.64085***	59.10284***
NB		–	-52.45593***	35.85095***
HP			–	53.14036*** (995044***)
HNB				–

Using variance inflation factor (VIF) to check for the presence of multicollinearity in our experiments we found VIF scores for all independent variables in our regression analysis were within the range of zero to 5 indicating no multicollinearity issues.

Table 4.22 presents results from the hurdle regression model applied for Facebook. The regression coefficients in the zero-portion are exponentiated as odds ratios (OR) while the exponentiated regression coefficients in the count portion are treated as incident rate ratios (IRR) [25]. In the analysis we consider all other variables to remain constant while we interpret the results of a particular variable.

4.2.3.2 Analysis for Activity Presence

The coefficients of the logit regression in the zero portion of the model indicate how the features relate to crossing the 'hurdle' of obtaining at least 1 activity (i.e. either a like, share or comment).

A unit increase in the log-transformed page likes increase the odds of getting at least one activity by 201% (OR=3.010), all other variables remaining constant.

A unit increase in positive sentiment increases the odds of getting an activity

by 17.4% while a unit increase in negative sentiment decrease the odds of getting an activity by 11.4%.

Of the various post types, the presence of a question or status are both linked to a decrease in the odds of a post getting an activity by 99.6% and 91.8%, all other variables remaining constant. The other post types are not significantly associated with the presence of an activity.

Twelve of the 15 semantic groups increase the odds of getting an activity with the group "Activities & Behavior" showing the highest increase (90.3%). "Organizations" is the only semantic group that decrease the odds of getting an activity by 29.5%.

4.2.3.3 Analysis for Activity Abundance

We now analyze the coefficients of the Negative Binomial regression in the count portion of the hurdle model (Table 4.22). This allows us to study factors related to the rate of activity for posts that succeed in getting at least one activity.

Given a unit increase in the log-transformed count of page likes, the rate of activity is expected to increase by a factor of 6.033, while holding all other variable in the model constant.

For the sentiment, a unit increase in positive sentiment increases the rate of activity by a factor of 1.126 while a unit increase in negative sentiment decreases the rate of activity by a factor of 0.934, with all other variables remaining constant.

Of the various post types, the presence of a photo, link, status or video each

increase the expected rate of activity with photos giving the highest increase by a factor of 6.302 with all other variables remaining constant.

Of the 15 semantic groups only 5 have significant positive association with the rate of activity. The presence of the semantic group "Phenomena" increases the rate of activity by a factor of 1.155 (highest amongst the semantic groups) followed by "Chemicals & Drugs" which increases the rate of activity by a factor of 1.073. Of the 6 semantic groups having significant negative associations with the abundance of activity, "Occupations" has the largest decrease in the rate of activity with a factor of 0.793. Examples of other groups negatively associated are "Objects", "Geographic Areas" and "Organizations".

4.2.3.4 Analysis across Hurdle Components

Looking across both components of the hurdle model several features show consistent benefit for engagement. These include numbers of page likes as well as positive sentiment of a post. Emphasizing semantic groups such as Activities & Behaviour, Chemicals & Drugs, Phenomena and Physiology correlate with increased engagement. Negative sentiment in posts almost always correlates with lower engagement. So does the semantic group Organizations. Post types such as status or video are not important for crossing the initial hurdle of getting at least 1 activity but then their presence correlate with higher activity rate.

Table 4.22: Results of hurdle negative binomial model for Facebook data.

	Zero Portion				**Count Portion**			
	Estimate(SE)	**OR**	**z value**	**p**	**Estimate(SE)**	**IRR**	**z value**	**p**
(Intercept)	-2.71 (0.47)	0.067	-5.763	***	-5.631 (0.169)	0.004	-33.356	***
Log-transformed page likes	1.102 (0.025)	3.010	43.931	***	1.797 (0.01)	6.033	174.673	***
Link	-0.817 (0.462)	0.442	-1.77		0.554 (0.162)	1.741	3.421	***
Music	-0.48 (0.57)	0.619	-0.843		0.06 (0.223)	1.062	0.271	
Photo	-0.22 (0.464)	0.802	-0.475		1.833 (0.163)	6.253	11.267	***
Question	-5.62 (0.659)	0.004	-8.528	***	-0.456 (0.54)	0.634	-0.844	
Status	-2.499 (0.462)	0.082	-5.408	***	0.861 (0.163)	2.365	5.28	***
Video	-0.388 (0.473)	0.679	-0.82		1.041 (0.165)	2.833	6.302	***
Positive Sentiment	0.16 (0.023)	1.174	7.051	***	0.118 (0.009)	1.126	12.986	***
Negative Sentiment	-0.121 (0.015)	0.886	-7.857	***	-0.068 (0.006)	0.934	-10.692	***
Activities & Behaviors	0.644 (0.031)	1.903	20.605	***	0.06 (0.013)	1.061	4.741	***
Anatomy	0.088 (0.051)	1.092	1.743		0.048 (0.022)	1.049	2.191	*
Chemicals & Drugs	0.112 (0.035)	1.118	3.237	**	0.07 (0.015)	1.073	4.771	***
Concepts & Ideas	0.366 (0.027)	1.441	13.361	***	-0.013 (0.012)	0.987	-1.041	
Devices	0.321 (0.161)	1.378	1.998	*	-0.021 (0.066)	0.980	-0.312	
Disorders	0.329 (0.032)	1.390	10.369	***	-0.035 (0.014)	0.965	-2.514	*
Genes & Molecular Sequences	0.567 (0.199)	1.763	2.85	**	-0.084 (0.06)	0.920	-1.402	
Geographic Areas	0.091 (0.041)	1.095	2.232	*	-0.187 (0.017)	0.830	-10.776	***
Living Beings	0.242 (0.028)	1.274	8.675	***	0.01 (0.012)	1.010	0.787	
Objects	0.212 (0.036)	1.236	5.9	***	-0.117 (0.015)	0.889	-7.769	***
Occupations	0.055 (0.05)	1.057	1.108		-0.232 (0.02)	0.793	-11.472	***
Organizations	-0.35 (0.041)	0.705	-8.468	***	-0.078 (0.018)	0.925	-4.425	***
Phenomena	0.257 (0.041)	1.293	6.25	***	0.144 (0.017)	1.155	8.44	***
Physiology	0.284 (0.031)	1.328	9.13	***	0.034 (0.013)	1.035	2.614	**
Procedures	0.2 (0.036)	1.222	5.597	***	-0.034 (0.015)	0.966	-2.277	*
Log(theta)					-0.172 (0.011)	0.842	-15.005	***

Note: The estimate/coefficient (SE), exponent of coefficient (OR and IRR), z and p-values (*p<0.05, **p<0.01, ***p<0.001) are shown.

4.2.4 Modeling Activity Life Span

In our dataset almost 80% of posts have their last activity on the same day that it is posted while there are posts that garner attention for months or even years. Thus we see that the time to last activity can vary considerably. Therefore the characteristics of a post that influence such behavior are of great interest. Similar to the Twitter-based study, we use the Cox proportional hazards regression model [16] to predict how the different features (see Table 4.16) influence the time to last activity.

Table 4.23 shows the results. The regression coefficients are exponentiated as hazard ratios (HR) and used in the interpretation of the survival models. It is important to note here that a longer interval is desirable for the time to last activity. Thus the features with negative coefficients are the beneficial ones.

For continuous variables such as log-transformed counts of page likes, a unit increase in these values may change the time to last activity with all other variables remaining constant. For binary variables (each post type or each semantic group) the time to last activity may increase or decrease based on the presence of a feature compared to its absence in a post.

We find that a unit increase in the number of log-transformed page likes increases the time to last activity by 34.6% with all other variables remaining constant. A unit increase in positive sentiment increases the time to last activity by 2.1% while a unit increase in negative sentiment has no significant association with the time to last activity.

Of the various post types, the presence of photos or videos are both linked to an increase in the time to last activity. The other post types are not significantly associated with the time to last activity.

Amongst the 15 semantic groups, only eight have significant relation to the time to last activity. Posts containing semantic groups "Activities & Behavior", "Concepts & Ideas", "Genes & Molecular Sequences", "Phenomena" and "Procedures" are positively related to an increase in the time to last activity by 2.9%, 2.3%, 13.6%, 6.5% and 2.7% respectively. "Devices", "Organizations" and "Occupations" are the only ones that decrease the time to last activity by 14.7%, 4.3% and 5.6% respectively.

4.2.5 Discussion

Our results show that there is considerable difference between levels of Facebook use among organizations. We found that less than 5% of posts get more than 100 shares, likes or comments; a three-fourths majority of posts get on average 15 activities. We also found that an account's page likes have strong positive relationships with activity. While it is not an easy task for agencies to increase the number of users liking a page, it is still an easy metrics to follow.

Results also show that the Facebook users are typically not interested in dry textual posts from health agencies. Photos, info-graphics, videos or interactive links may increase the likelihood of posts to get more activities. This is partly consistent with our Twitter-based study where we found that the use of URLs was also associated with higher engagement. Quite surprisingly, question-related posts, which

Table 4.23: Results of Cox proportional hazards model for interval between a Facebook post and its last activity.

	Interval between FB Post & Last Activity			
	Coefficient (SE)	**HR**	**z**	**p**
Log-transformed page likes	-0.424(0.008)	0.654	-54.583	***
Link	-0.142(0.128)	0.868	-1.103	
Music	-0.211(0.172)	0.810	-1.228	
Photo	-0.435(0.129)	0.647	-3.377	***
Question	0.221(0.173)	1.248	1.28	
Status	-0.105(0.129)	0.900	-0.816	
Video	-0.291(0.131)	0.748	-2.214	*
Positive Sentiment	-0.022(0.007)	0.979	-2.989	**
Negative Sentiment	0.007(0.005)	1.007	1.437	
Activities & Behaviors	-0.03(0.01)	0.971	-2.925	**
Anatomy	-0.004(0.017)	0.996	-0.207	
Chemicals & Drugs	-0.011(0.012)	0.989	-0.935	
Concepts & Ideas	-0.023(0.01)	0.977	-2.376	*
Devices	0.137(0.053)	1.147	2.593	**
Disorders	0.012(0.011)	1.012	1.101	
Genes & Molecular Sequences	-0.146(0.051)	0.864	-2.876	**
Geographic Areas	-0.004(0.014)	0.996	-0.295	
Living Beings	0.0001(0.01)	1.000	0.02	
Objects	0.02(0.012)	1.020	1.66	
Occupations	0.042(0.016)	1.043	2.578	**
Organizations	0.054(0.014)	1.056	3.846	***
Phenomena	-0.068(0.014)	0.935	-4.988	***
Physiology	-0.005(0.011)	0.995	-0.434	
Procedures	-0.028(0.012)	0.973	-2.296	*

Note: The Coefficient (SE), hazard ratio (HR), z and p-values (*p<0.05, **p<0.01, ***p<0.001) for various independent variables are shown.

are typically posted to encourage public participation or interaction, are apparently not useful in engaging the public. Probably the organizations can look into more innovative ways to frame questions that would encourage user engagement.

We found that positive sentiment in Facebook posts from government agencies is associated with higher activity. This is quite interesting, especially because we found positive sentiment to have negative or no association with the level of engagement in Twitter. The reasons for this are not quite obvious and we would like to investigate more on this in future research.

As with Twitter, semantic groups have not been studied in the context of Facebook activities. We found that posts about activities and behaviors, and phenomenon are positively associated with engagement amount and duration. In contrast, posts about organizations and occupations tend to lower engagement. It may be that such posts are meant to be more informative than engaging.

4.2.6 Predictive Modeling of Activities

Similar to the Twitter-based study we built classifiers and regressors for predicting whether a post will generate an activity or not, as well as for predicting the activity count for posts that get at least one activity.

Using the same features as shown in Table 4.16 we built various classifiers, namely naïve Bayes (NB), decision tree (J48) and Support Vector Machine (SVM), for predicting whether a post gets any activity or not. The results in 10-fold cross-validation experiments are shown in Table 4.24. J48 outperforms the other methods

in terms of precision (0.849), recall (0.86) and F-score (0.849).

For modeling activity count we implemented various regressors, namely Support Vector Regression (SVR), M5P and ZeroR. As before, we use the same set of features as listed in Table 4.16. The results using 10-fold cross validation experiments are in Table 4.25. M5P outperforms the other methods in terms of higher correlation coefficient and lower root mean squared error (RMSE) while SVR has a lower mean absolute error (MAE).

It seems easier to distinguish posts receiving some activity from those that do not. It is easier to predict this for Facebook posts compared to Twitter (F-score of 0.849 vs 0.785). Predicting the number of actual activities is a much more challenging for both platforms. Even the best regression techniques in our Facebook experiment (M5P or SVR) perform quite poorly with low correlation and high MAE and RMSE.

Table 4.24: Results of classifiers for predicting activity presence.

Algorithms	**Precision**	**Recall**	**F-score**
NB	0.803	0.815	0.808
J48	0.849	0.86	0.849
SVM	0.815	0.834	0.808

4.2.7 Conclusions

Similar to the Twitter-based study, we present the first comprehensive analyses of engagement of health agencies on Facebook. The level of Facebook activity varies greatly by health agency. We find that the semantic content of the Facebook posts

Table 4.25: Results of regressors for predicting activity counts.

Algorithms	**Correlation coefficient**	**MAE**	**RMSE**
ZeroR	-0.014	28.882	118.548
M5P	0.337	21.028	111.613
SVR	0.292	20.645	129.032

(e.g., about activities and behaviors, chemicals and drugs) is as important as the number of page likes as both correlate with higher level of engagement (activity count). Positive sentiment of posts correlates with not just higher engagement but also longer duration of engagement. Predictive modeling of activities on Facebook works well for classifying posts that get any activity versus those that don't. Overall, we show the robustness of our methods for engagement analysis on Facebook.

4.3 Analysis across Twitter and Facebook Studies

4.3.1 Comparison of Agencies

Our study on engagement of health organizations on Twitter and Facebook reveal interesting insights. Table 4.26 shows 19 agencies that have one or more accounts across both Twitter and Facebook and relevant details. We note that across both platforms two agencies dominate the number of posts/tweets and activities/retweets. The Office of Secretary (OS) has the highest number of Facebook posts and the second highest number of tweets. The CDC generates the most total activity on Facebook and leads other agencies in number of tweets posted. The CDC also generates the most activity per Facebook post while NIH/NIMH gets most retweets per tweet. NIH/NIEHS has the fewest Facebook posts and activities. It also has a very

small footprint in Twitter. On Twitter, NIH/NEI has the fewest number of tweets, retweets and retweets per tweet. It seems to be more engaging on Facebook with an average of 8.38 activities per post compared to only 0.83 on Twitter. Some accounts are present in only one of the platforms such as AHRQ, AoA, NIA, CMS, etc.

Table 4.26: Comparison of agencies across Facebook and Twitter in terms of posts and responses.

Agency	**# FB posts**	**# total FB activity**	**# FB activity per post**	**# tweets**	**# retweets**	**# retweets per tweet**
ACF	372	3147	8.46	605	1924	3.18
CDC	7313	**407910**	**55.78**	**37136**	278885	7.51
FDA	538	24414	45.38	10574	75245	7.12
HRSA	2456	11601	4.72	1241	5391	4.34
NIH	2831	39388	13.91	15550	49666	3.19
NIH/NCCAM	659	8651	13.13	1489	4102	2.75
NIH/NCI	3455	37589	10.88	15679	46586	2.97
NIH/NEI	447	3745	8.38	401	331	0.83
NIH/NHGRI	417	7248	17.38	401	652	1.63
NIH/NHLBI	3510	115104	32.79	5135	29447	5.73
NIH/NIAID	632	3375	5.34	1725	2808	1.63
NIH/NIAMS	414	1356	3.28	822	1850	2.25
NIH/NIDA	1657	26427	15.95	2191	7484	3.42
NIH/NIEHS	148	418	2.82	682	858	1.26
NIH/NIGMS	236	1579	6.69	983	1791	1.82
NIH/NIMH	427	21456	50.25	959	16779	**17.5**
NIH/NLM	4076	32044	7.86	15058	48497	3.22
OS	**9158**	258203	28.19	36587	**376158**	10.28
SAMHSA	2915	39805	13.66	4971	21729	4.37

4.3.2 Comparison of Statistical Modeling

Here we compare the findings from the hurdle and Cox proportional hazards models across Twitter and Facebook studies. While the set of features considered for both platforms vary considerably, the set of common features can be used in analyzing observations across platforms. We find that while negative sentiment lowers

engagement in both Twitter and Facebook, positive sentiment enhances engagement only in Facebook but not in Twitter. Analyzing the 15 semantic groups we find that while "Activities & Behavior", "Chemicals & Drugs", "Living Beings", "Phenomena" and "Physiology" enhance engagement across both social media platforms, "Organizations" consistently decrease engagement.

Analysis of the time to last retweet or activity across the two platforms reveal that a unit increase in positive sentiment increases the time to last retweet/activity. Amongst the 15 semantic groups we find that "Activities & Behaviors" increases the time to last retweet/activity in both platforms while "Occupations" and "Organizations" consistently decrease the time to last retweet/activity.

4.4 Conclusions

In this chapter, we addressed the questions raised in Sections 4.1 and 4.2. While the level of engagement and its duration vary greatly by agencies and platforms, we find that various factors have significant influence on engagement. For Twitter, the number of followers or the use of hashtags are associated with enhanced engagement while for Facebook the type and sentiment of posts are important factors. We also find that semantic content of tweets or posts are important and tweets/posts about activities and behaviors or chemicals and drugs correlate with higher levels of engagement. Predictive models of analyzing engagement work well in broader categorizations but perform poorly in other scenarios. Overall, we show the feasibility of using computational approaches for analyzing factors influencing engagement of

health agencies.

CHAPTER 5
CONCLUSIONS

This thesis presents two projects one aimed at the individual level (belief surveillance) and one at the organization level (engagement). It contributes to the study of computational methods for assessing health communications in Web 2.0, more specifically, in two social media platforms namely Twitter and Facebook.

We proposed a novel framework for belief surveillance and tested our ideas in healthcare though our methods may apply more generally. We demonstrate that although factual statements garner a high degree of support, some are still being questioned. Most fictional statements also garner a high degree of support. These results potentially offer an informed basis for targeting educational strategies. Overall Twitter offers valuable signals for belief surveillance. The divided positioning on high controversy probes such as vaccines causing autism are also seen adding to the credibility of our methods. We also show that we may use off-the-shelf tools to build classifiers for belief surveillance. Though we focus almost exclusively on F-score, our experiments produce overall reasonable results in various measures. We intentionally kept our classifiers general so as to handle new probes in the future. We find we are able to handle new probe statements that fall within the same 'genre' of relationships. We also show that one may use off-the-shelf tools to mine Twitter conversations for beliefs discussed. This is exciting as it supports proactive belief surveillance.

We hypothesized that the levels of support, opposition and doubt for certain probes may change over time. We found using longitudinal data that this is true and

surprisingly the overall support for false probes increases over time for a specific set of probes. This is of concern and again emphasizes the importance of our surveillance framework in identifying areas that lack public awareness.

We also propose methods for *harvesting* probes automatically from twitter and thereby discovering naturally expressed health beliefs on Twitter. This is an exciting aspect as it points to being able to proactively conduct belief surveillance with tweet data. This also shows that our method is robust in the sense that it does not rely solely on the predefined probes to measure a population's beliefs, but it can automatically discover and monitor new probes from social media. We find a large number of probes related to not just therapeutic drugs in our study, but also recreational drugs which demonstrate the prevalence of discussions on such topics in Twittersphere. Beyond known effects or side-effects of drugs, were able to uncover several probes containing novel information (e.g. *Cialis causes anxiety*).

We also demonstrate that our belief surveillance framework may be used as input for hypothesis discovery in the sense of suggesting 'proto-ideas'. While for some harvested probes (e.g. Coconut oil treats psoriasis) we found explicit published evidence, a considerable number of probes (e.g. Neem treats psoriasis) lacked explicit scientific evidence. Our goal is not to verify the scientific validity of such hypothesis but to simply show the use of contemporary and ever-growing channel of information propagation, that is social media, in literature-based discovery.

For the engagement study, we present one of the first comprehensive analyses of Twitter and Facebook engagement via public health agencies. The level of Twitter

and Facebook activity varies greatly by health agency: some health accounts are very active and others are not as much. However, it seems that for Twitter, the content of the tweets (e.g., about activities and behaviors, disorders) and not the number of tweets alone that is associated with a higher level of engagement (number of re-tweets). Similarly for Facebook not only the content of the post but also the type of post (e.g. photo or video) and its sentiment are associated with higher engagement. Furthermore, although some of the factors associated with more engagement cannot be changed by the agency (e.g., the length of time they have been active on Twitter), several factors associated with higher engagement can be changed (e.g., use of hash-tags, URLs for tweets and type of post for Facebook). Our results will help provide a framework for future experiments designed to improve the publics engagement with health agencies via Twitter and Facebook.

Summarizing across the two studies we find that the social media population generally support various false health-related notions and health organizations that communicate via the social media can do better or worse depending on the way they communicate. In other words, we are able to identify health-related notions that needs attention from public-health educators or health organizations and we propose methods in which they can communicate more effectively to deliver the pertinent information.

There has been an unprecedented growth in digital communication in the past few years. Social media, blogs and online communities have become an integral part of our daily life and are increasingly being used for conveying or procuring

health information. The two projects discussed in this thesis show the usefulness of mining health communications for developing various insights. The application of computational approaches enhance the scalability of our methods. We conclude that the development of sophisticated computational approaches is critical for mining and analysis of health communications and holds the key to improve health and well-being of population.

REFERENCES

[1] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576. Association for Computational Linguistics.

[2] Cory L Armstrong and Fangfang Gao. Now tweet this how news organizations use twitter. *Electronic News*, 4(4):218–235, 2010.

[3] Alan R Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*, pages 17–21, 2001.

[4] Alan R Aronson, Olivier Bodenreider, Dina Demner-Fushman, Kin Wah Fung, Vivian K Lee, James G Mork, Aurlie Nvol, Lee Peters, and Willie J Rogers. From indexing the biomedical literature to coding clinical text: Experience with MTI and machine learning approaches. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 105–112. Association for Computational Linguistics.

[5] Alan R Aronson, James G Mork, Clifford W Gay, Susanne M Humphrey, and Willie J Rogers. The NLM indexing initiative's Medical Text Indexer. *Medinfo*, 11(Pt 1):268–72, 2004.

[6] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 492–499, Washington, DC, USA, 2010. IEEE Computer Society.

[7] Sanmitra Bhattacharya, Viet Ha-Thuc, and Padmini Srinivasan. MeSH: a window into full text for document summarization. *Bioinformatics*, 27(13):i120–128, Jul 2011.

[8] Sanmitra Bhattacharya and Padmini Srinivasan. A semantic approach to involve Twitter in LBD efforts. In *2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, pages 248–253. IEEE, 2012.

[9] Sanmitra Bhattacharya, Hung Tran, and Padmini Srinivasan. Discovering Health Beliefs in Twitter. In *AAAI-2012 Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text. Washington, DC*, 2012.

[10] Sanmitra Bhattacharya, Hung Tran, Padmini Srinivasan, and Jerry Suls. Belief Surveillance with Twitter. In *Proceedings of the Fourth ACM Web Science Conference (WebSci12)*, pages 55–58, Evanston, IL, USA, 2012.

[11] Olivier Bodenreider. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

[12] Axel Bruns and Jean E Burgess. # ausvotes: How Twitter covered the 2010 Australian Federal Election. *Communication, Politics and Culture*, 44(2):37–56, 2011.

[13] A Colin Cameron and Pravin Trivedi. *Regression Analysis of Count Data*, volume 53. Cambridge University Press, 2013.

[14] Brett Caraway. Online labour markets: An inquiry into oDesk providers. *Work Organisation, Labour and Globalisation*, 2010.

[15] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. *In Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM 2010)*, 10:10–17, 2010.

[16] Yin Bin Cheung. Zeroinflated models for regression analysis of count data: a study of growth and development. *Statistics in medicine*, 21(10):1461–1469, 2002.

[17] Cynthia Chew and Gunther Eysenbach. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PloS One*, 5(11):e14118, 2010.

[18] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[19] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. Political Polarization on Twitter. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[20] Aron Culotta. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the First Workshop on Social Media Analytics (SOMA 2010)*. ACM, 2010.

[21] K Michael Cummings, Andrew Hyland, Gary A Giovino, Janice L Hastrup, Joseph E Bauer, and Maansi A Bansal. Are smokers adequately informed about the health risks of smoking and medicinal nicotine? *Nicotine & Tobacco Research*, 6(Suppl 3):S333–S340, 2004.

[22] Batya B Davidovici and Ronni Wolf. The role of diet in acne: facts and controversies. *Clinics in Dermatology*, 28(1):12–16, 2010.

[23] Frank DeStefano. Vaccines and autism: evidence does not support a causal association. *Clinical Pharmacology & Therapeutics*, 82(6), 2007.

[24] Lorie Donelle and Richard G Booth. Health tweets: an exploration of health promotion on Twitter. *Online Journal of Issues in Nursing*, 17(3), 2012.

[25] Robert D Dvorak, Matthew R Pearson, and Nicholas J Kuvaas. The five-factor model of impulsivity-like traits and emotional lability in aggressive behavior. *Aggressive behavior*, 39(3):222–228, 2013.

[26] Regina Elandt and Norman Lloyd Johnson. *Survival models and data analysis*. John Wiley and Sons, 1980.

[27] Joseph L. Fleiss. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, 1981.

[28] King-wa Fu and Michael Chau. Reality check for the chinese microblog space: a random sampling approach. *PloS one*, 8(3):e58356, 2013.

[29] Ted Gansler, S Jane Henley, Kevin Stein, Eric J Nehl, Carol Smigal, and Edwin Slaughter. Sociodemographic determinants of cancer treatment health literacy. *Cancer*, 104(3):653–660, 2005.

[30] Dean Giustini and Mary-Doug Wright. Twitter: an introduction to microblogging for health librarians. *Journal of the Canadian Health Libraries Association JCHLA*, 17(1):11–17, 2009.

[31] Scott A. Golder and Michael W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 2011.

[32] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[33] Marion E Hambrick, Jason M Simmons, Greg P Greenhalgh, and T Christopher Greenwell. Understanding Professional Athletes' Use of Twitter: A Content Analysis of Athlete Tweets. *International Journal of Sport Communication*, 3(4), 2010.

[34] Lars Kai Hansen, Adam Arvidsson, Finn Årup Nielsen, Elanor Colleoni, and Michael Etter. Good friends, bad news-affect and virality in Twitter. In *Future Information Technology*, pages 34–43. Springer, 2011.

[35] Jenine K Harris, Nancy L Mueller, and Doneisha Snider. Social media adoption in local health departments nationwide. *American Journal of Public Health*, 103(9):1700–1707, 2013.

[36] Carleen Hawn. Take two Aspirin and tweet me in the morning: How Twitter, Facebook, and other social media are reshaping health care. *Health Affairs*, 28(2):361–368, 2009.

[37] N Heaivilin, B Gerbert, JE Page, and JL Gibbs. Public health surveillance of dental pain via Twitter. *Journal of Dental Research*, 90(9):1047–1051, 2011.

[38] Alfred Hermida. From TV to Twitter: How ambient news became ambient journalism. *Media/Culture Journal*, 13(2), 2010.

[39] Bernardo Huberman, Daniel Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1), 2008.

[40] Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11):2169–2188, November 2009.

[41] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why We Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA, 2007. ACM.

[42] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

[43] Punam Anand Keller and Donald R Lehmann. Designing effective health communications: a meta-analysis. *Journal of Public Policy & Marketing*, 27(2):117–130, 2008.

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

[45] Vasileios Lampos, Tijl Bie, and Nello Cristianini. Flu Detector - Tracking Epidemics on Twitter. In JosLuis Balczar, Francesco Bonchi, Aristides Gionis, and Michle Sebag, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 6323 of *Lecture Notes in Computer Science*, pages 599–602. Springer Berlin Heidelberg, 2010.

[46] David Laniado and Peter Mika. Making Sense of Twitter. In *Proceedings of the 9th International Semantic Web Conference on The Semantic Web - Volume Part I*, ISWC'10, pages 470–485, Berlin, Heidelberg, 2010. Springer-Verlag.

[47] Jerald F Lawless. *Statistical models and methods for lifetime data*, volume 362. John Wiley & Sons, 2011.

[48] Carolyn E Lipscomb. Medical subject headings (MeSH). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

[49] Kristen Lovejoy, Richard D Waters, and Gregory D Saxton. Engaging stakeholders through Twitter: How nonprofit organizations are getting more out of 140 characters or less. *Public Relations Review*, 38(2):313–318, 2012.

[50] Alexa T McCray, Anita Burgun, and Olivier Bodenreider. Aggregating UMLS semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, (1):216–220, 2001.

[51] K McNeil, P M Brna, and K E Gordon. Epilepsy in the Twitter era: a need to re-tweet the way we think about seizures. *Epilepsy & Behavior*, 23(2):127–130, Feb 2012.

[52] Marcus Messner, Maureen Linke, and Asriel Eford. Shoveling tweets: An analysis of the microblogging engagement of traditional news organizations. In *12th International Symposium for Online Journalism, Austin*.

[53] Margaret E Morris, Sunny Consolvo, Sean Munson, Kevin Patrick, Janice Tsai, and Adam D I Kramer. Facebook for health: opportunities and challenges for driving behavior change. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, pages 443–446, New York, NY, USA, 2011. ACM.

[54] Maghboeba Mosavel and Nadia El-Shaarawi. "I have never heard that one": Young girls' knowledge and perception of cervical cancer. *Journal of Health Communication*, 12(8):707–719, 2007.

[55] John Mullahy. Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365, 1986.

[56] Dhiraj Murthy, Alexander Gross, and Daniela Oliveira. Understanding Cancer-Based Networks in Twitter Using Social Network analysis. In *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing*, ICSC '11, pages 559–566, Washington, DC, USA, 2011. IEEE Computer Society.

[57] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. Bad news travel fast: A content-based analysis of interestingness on Twitter. In *Proceedings of the 3rd International Web Science Conference (WebSci)*, page 8. ACM, 2011.

[58] Brad L Neiger, Rosemary Thackeray, Scott H Burton, Christophe G Giraud-Carrier, and Michael C Fagen. Evaluating social media's capacity to develop engaged audiences in health promotion settings use of Twitter metrics as a case study. *Health Promotion Practice*, 14(2):157–162, 2013.

[59] Brad L Neiger, Rosemary Thackeray, Scott H Burton, Callie R Thackeray, and Jennifer H Reese. Use of Twitter among local health departments: An analysis of information sharing, engagement, and action. *Journal of Medical Internet Research*, 15(8), 2013.

[60] Finn rup Nielsen. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*, 2011.

[61] Julie R Palmer, Lauren A Wise, Elizabeth E Hatch, Rebecca Troisi, Linda Titus-Ernstoff, William Strohsnitter, Raymond Kaufman, Arthur L Herbst, Kenneth L Noller, Marianne Hyer, et al. Prenatal diethylstilbestrol exposure and risk of breast cancer. *Cancer Epidemiology Biomarkers & Prevention*, 15(8):1509–1514, 2006.

[62] Michael J Paul and Mark Dredze. You are what you Tweet: Analyzing Twitter for public health. In *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM), Barcelona, Catalonia, Spain, July 17-21*, 2011.

[63] Kyle W Prier, Matthew S Smith, Christophe Giraud-Carrier, and Carl L Hanson. Identifying health-related topics on Twitter: An exploration of tobacco-related tweets as a test topic. In *Proceedings of the 4th international conference on Social computing, behavioral-cultural modeling and prediction*, SBP'11, pages 18–25, Berlin, Heidelberg, 2011. Springer-Verlag.

[64] Judith J Prochaska, Cornelia Pechmann, Romina Kim, and James M Leonhardt. Twitter= quitter? An analysis of Twitter quit smoking social networks. *Tobacco Control*, 21(4):447–449, 2012.

[65] John R Quinlan. Learning with continuous classes. In *Proceedings of the 5th Australian joint Conference on Artificial Intelligence*, volume 92, pages 343–348. Singapore, 1992.

[66] Svetlana Rybalko and Trent Seltzer. Dialogic communication in 140 characters or less: How Fortune 500 companies engage stakeholders using Twitter. *Public Relations Review*, 36(4):336–341, 2010.

[67] Daniel Scanfeld, Vanessa Scanfeld, and Elaine L Larson. Dissemination of health information through social networks: Twitter and antibiotics. *American Journal of Infection Control*, 38(3):182–188, 2010.

[68] Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PloS One*, 6(5):e19467, 2011.

[69] Sarabjot Singh, Nishchol Mishra, and Sanjeev Sharma. Survey of various techniques for determining influential users in social networks. In *Emerging Trends in Computing, Communication and Nanotechnology (ICE-CCN), 2013 International Conference on*, pages 398–403. IEEE, 2013.

[70] Uglješa Stankov, Lazar Lazić, and Vanja Dragićević. The extent of use of basic Facebook user-generated content by the national tourism organizations in Europe. *European Journal of Tourism Research*, 3(2), 2010.

[71] Stefan Stieglitz and Linh Dang-Xuan. Political communication and influence through microblogging–An empirical analysis of sentiment in Twitter messages and retweet behavior. In *45th Hawaii International Conference on System Science (HICSS 2012)*, pages 3500–3509. IEEE.

[72] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. Want to be retweeted? large scale analytics on factors impacting retweet in Twitter network. In *2010 IEEE Second International Conference on Social Computing (SocialCom),*, pages 177–184. IEEE.

[73] S John Sullivan, Anthony G Schneiders, Choon-Wi Cheang, Emma Kitto, Hopin Lee, Jason Redhead, Sarah Ward, Osman H Ahmed, and Paul R McCrory. "What's happening?" A content analysis of concussion-related traffic on Twitter. *British Journal of Sports Medicine*, 46(4):258–263, 2012.

[74] Don R Swanson. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18, 1986.

[75] Rosemary Thackeray, Brad L Neiger, Amanda K Smith, and Sarah B Van Wagenen. Adoption and use of social media among public health departments. *BMC Public Health*, 12(1):242, 2012.

[76] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418, 2011.

[77] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

[78] Luca Toldo, Sanmitra Bhattacharya, and Harsha Gurulingappa. Automated identification of adverse events from case reports using machine learning. In *Proceedings XXIV Conference of the European Federation for Medical Informatics. Workshop on Computational Methods in Pharmacovigilance, Pisa, Italy*, pages 26–29, 2012.

[79] Jay M Ver Hoef and Peter L Boveng. Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–2772, 2007.

[80] Richard D Waters, Emily Burnett, Anna Lamm, and Jessica Lucas. Engaging stakeholders through social networking: How nonprofit organizations are using Facebook. *Public Relations Review*, 35(2):102–106, 2009.

[81] Richard D Waters and Jia Y Jamal. Tweet, tweet, tweet: A content analysis of nonprofit organizations Twitter updates. *Public Relations Review*, 37(3):321–324, 2011.

[82] Richard D Waters and Jensen M Williams. Squawking, tweeting, cooing, and hooting: Analyzing the communication patterns of government agencies on Twitter. *Journal of Public Affairs*, 11(4):353–363, 2011.

[83] Leila Weitzel, Paulo Quaresma, and Jos Palazzo M de Oliveira. Evaluating quality of health information sources. In *Advanced Information Networking and Applications (AINA), 2012 IEEE 26th International Conference on*, pages 655–662. IEEE.

[84] Joshua H West, Parley C Hall, Carl L Hanson, Kyle Prier, Christophe Giraud-Carrier, E Shannon Neeley, and Michael D Barnes. Temporal variability of problem drinking on Twitter. *Open Journal of Preventive Medicine*, 2:43, 2012.

[85] Rainer Winkelmann and Klaus F Zimmermann. Recent developments in count data modelling: theory and application. *Journal of economic surveys*, 9(1):1–24, 1995.

[86] Shanchan Wu, Leanna Gong, William Rand, and Louiqa Raschid. Making recommendations in a microblog to improve the impact of a focal user. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 265–268. ACM.

[87] Chao Yang, Sanmitra Bhattacharya, and Padmini Srinivasan. Lexical and machine learning approaches toward online reputation management. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[88] Dan Zarrella. Science of retweets. *Retrieved December*, 15:2009, 2009.

[89] Qing T Zeng, Tony Tse, Guy Divita, Alla Keselman, Jon Crowell, Allen C Browne, Sergey Goryachev, and Long Ngo. Term identification methods for consumer health vocabulary development. *Journal of Medical Internet Research*, 9(1):e4, 2007.

[90] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing Twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer, 2011.