UNDERSTANDING MISINFORMATION ON TWITTER IN THE CONTEXT OF CONTROVERSIAL ISSUES

BY

ASEEL ADDAWOOD

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Informatics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Professor Umberto Ravaioli, Chair
Professor David Tewksbury
Adjunct Associate Professor Nahil Sobh
Assistant Professor Chadly Stern

ABSTRACT

Social media is slowly supplementing, or even replacing, traditional media outlets such as television, newspapers, and radio. However, social media presents some drawbacks when it comes to circulating information. These drawbacks include spreading false information, rumors, and fake news. At least three main factors create these drawbacks: The filter bubble effect, misinformation, and information overload. These factors make gathering accurate and credible information online very challenging, which in turn may affect public trust in online information. These issues are even more challenging when the issue under discussion is a controversial topic. In this thesis, four main controversial topics are studied, each of which comes from a different domain. This variation of domains can give a broad view of how misinformation is manifested in social media, and how it is manifested differently in different domains.

This thesis aims to understand misinformation in the context of controversial issue discussions. This can be done through understanding how misinformation is manifested in social media as well as by understanding people's opinions towards these controversial issues. In this thesis, three different aspects of a tweet are studied. These aspects are 1) the user sharing the information, 2) the information source shared, and 3) whether specific linguistic cues can help in assessing the credibility of information on social media. Finally, the web application tool *TweetChecker* is used to allow online users to have a more in-depth understanding of the discussions about five different controversial health issues. The results and recommendations of this study can be used to build solutions for the problem of trustworthiness of user-generated content on different social media platforms, especially for controversial issues.

To My Family

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my advisor, Prof. Umberto Ravaioli, for his continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. His guidance helped me through the entire process of researching and writing this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Nahil Sobh, Prof. David Tewksbury, and Dr. Chadly Stern, for their insightful comments and encouragement, but also for their hard questions, which gave me the incentive to widen my research from various perspectives.

My sincere thanks also go to Dr. Jana Diesner, Dr. Masooda Bashir, Dr. Catherine Blake, and Dr. Jodi Schneider, who provided me with advice and support throughout my Ph.D. Without their precious support it would not have been possible to accomplish this degree. I am also grateful to the following university staff: Karin Readel and Meg Edwards for their unfailing emotional and academic support and assistance.

I gratefully acknowledge the funding received towards my Ph.D. from the Imam Muhammad Bin Saud University. A strong personal acknowledgment also goes to my country Saudi Arabia, for providing such support, without which I would not have been able to accomplish this achievement.

This work is dedicated to everyone who supported me through my journey to accomplish this degree. My beloved parents, Abdulmohsen Addawood and Hassah Alzaid, who have always been there for me and provided moral and emotional support throughout my life. My brothers and sister, Eyad, Zeyad, Maram, Faisal, and Mishal, whose unbelievable support was invaluable to me. Also, this work is dedicated to my son, Mohammad, who has been the light of my life. Looking at him every day gave me the inspiration that made getting this degree possible. I also give special thanks to my best friend Hassah Alswiulim for being there for me throughout the entire process. Also, my thanks extend to all the people who have been part of this journey as friends or colleagues.

All of you have been my biggest cheerleaders!

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Background

Online information sources are gradually supplementing, or even replacing, traditional media outlets such as television, newspapers, and radio (Gaskins & Jerit, 2012). This is particularly the case for social media, which serves as a major source of information in the everyday lives of individuals (Gil de Zúñiga, Jung, & Valenzuela, 2012). Real-time information shared via social media from the actual locations where events are unfolding usually spreads faster, and to a wider audience, than information from traditional news media sources (Gayo-Avello, Castillo, Mendoza, & Poblete, 2013).

1.2 Statement of the Problem

Since social media has become an integral part of our life and society and increasing amounts of information is being spread on the internet and through our mobile phones, it is important to investigate the effects of social media on our community. However, social media as a source of information presents some drawbacks. These drawbacks include the spread of false, misleading, or unsubstantiated information, rumors, and fake news (Nyhan & Reifler, 2010), or specify the spread of Misinformation.

Examples of misinformation include messages that are isolated from their original contexts or that contain facts mixed with opinions or fiction. For example, a study that showed that when the Ebola crisis broke out in 2014, lies, half-truths, and rumors spread as quickly as accurate news on social media, specifically Twitter (Jin et al., 2014). In other words, conspiracy theories, innuendo, and rumors about the disease propagated on social media just as readily as factual news reports. Moreover, another past study revealed that being asked misleading questions about an experience led individuals to forget targeted details and remember false information instead (Ayers & Reder, 1998). This study shows that information presentation affects public perceptions of information truthfulness.

One of the main factors of the spread of misinformation, ‘the filter bubble effect’, refers to the means by which people can become isolated from a diversity of viewpoints or content given to the rise of online personalization tools (Nguyen, Hui, Harper, Terveen, & Konstan, 2014).

Moreover, the information they are exposed to is selected through recommendation algorithms (Liao & Fu, 2013) separating users from information (and news) that disagrees with their viewpoints (Pariser, 2011).

One study demonstrated that online users learn about a topic more efficiently when presented with information from contrasting viewpoints and with information about the credibility of the sources for these viewpoints (Galland, Abiteboul, Marian, & Senellart, 2010). The filter bubble effect limits the extent to which users are exposed to this information. The filter bubble effect facilitates the creation of "echo chambers," in which individuals are largely exposed to information from like-minded individuals. Previous studies have shown that this effect increased after the emergence of the Internet (Sunstein, 2009).

Misinformation plays a key role in creating polarized groups (Zollo et al., 2015). These polarized "echo chamber" communities are more susceptible to the dissemination of misinformation (Michela Del Vicario et al., 2016). Another way these phenomena may interact is when users incorrectly classify news as sources of misinformation simply due to disagreement, not because it reports actual false or imprecise facts (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012). This behavior makes identifying accurate and credible information online very challenging, which in turn affects public trust in the information people encounter online (Corritore, Wiedenbeck, Kracher, & Marble, 2007; Metzger & Flanagin, 2013). One example of this challenge is that inaccurate tweets propagated on Twitter in 2009 about swine flu, which caused a large-scale panic among the general public (Coursey, 2009). The same year, the Fox News Twitter account was hacked to falsely report that former President Barack Obama had been shot dead, which caused panic amongst the public and made the verification of this information more difficult (Gabbatt, 2009). Gathering accurate and credible information is more challenging when the issue under discussion is a controversial topic, such as the debate about a possible relationship between MMR vaccinations and Autism.

Controversial issues are more effected by misinformation because of the subnational polarization of opinions towards controversial issues. Searching for a piece of information regarding a controversial issue online increases this polarization. As search engines feed into confirmation biases and encourage users to remain in their echo chambers (Carmel, Yom-Tov, Darlow, & Pelleg, 2006; Novin & Meyers, 2016), which in turn prevent users from being exposed to other viewpoints. For example, on social media platforms such as Facebook, a large portion of users are exposed to news shared by their friends (Bakshy, Messing, & Adamic, 2015; Matsa & Mitchell, 2014). Moreover, people may express strong emotions about controversial topics, i.e., they are either for or against the substance of controversial issues (Popescu & Pennacchiotti, 2010). Also, when presented with multiple contradictory viewpoints that contain no clear evidence, users may not be able to make the right decision about the information they should rely on.

An example of this is the vaccine controversy, in which a variety of sources disseminate information from different perspectives. The spectrum of sources ranges from traditional sources, such as public health officials and physicians, to celebrities and parent/child advocacy groups (Freed, Clark, Butchart, Singer, & Davis, 2011), which makes evaluating the credibility of conflicting viewpoints more challenging. Furthermore, some online discussions regarding controversial topics contain less accurate information. Myths propagated online surrounding vaccinations, for instance, have prompted some parents to withhold immunizations from their children (Lewandowsky et al., 2012). Together, these reasons make it difficult for online users to distinguish between accurate and inaccurate information in the case of controversial topics. So, gathering accurate and credible information online is very challenging and in turn may affect public trust in online information. Such issues are even more challenging when the issue under discussion is a controversial topic.

1.3 Objectives of the Study

From the information presented above it is increasingly important to understand misinformation in the context of controversial issue discussions. This will allow us to accomplish several objectives:

1.3.1 The ability to mitigate misinformation spread in social media.

One of the main objectives is to understand how misinformation is manifested in social media. This understanding can help Twitter users identify misinformation in social media by applying the characteristics of misinformation identified in this research. Moreover, it can help Twitter, the company, by identifying the characteristics that can spot misinformation easily in the tweets.

1.3.2 The ability to understand and detect people’s opinions towards controversial issues.

As explained before, people have polarized opinions toward different controversial issues. Detecting polarized opinions in social media can help with identifying false information since these topics are prone to such kind of information, where each side of the debate is trying to convince the other side even with the usage of fabricated information. Moreover, understanding public opinions and attitudes towards controversial topics may help scholars, law enforcement officials, and policy-makers develop better policies and guidelines.

1.3.3 Improve assessment literacy

Another objective is to help online users identify misinformation in social media by providing them with another layer into the tweets they encounter. An online tool was developed to evaluate controversial health issues discussion in real time. The tool is discussed in more details in chapter 7.

1.4 Study Components

In this thesis, four main controversial topics are studied; each of these topics comes from a different domain. This variation of domains can give a broad view of how misinformation is manifested in social media, and how it is manifested differently in different domains. The four domains are:

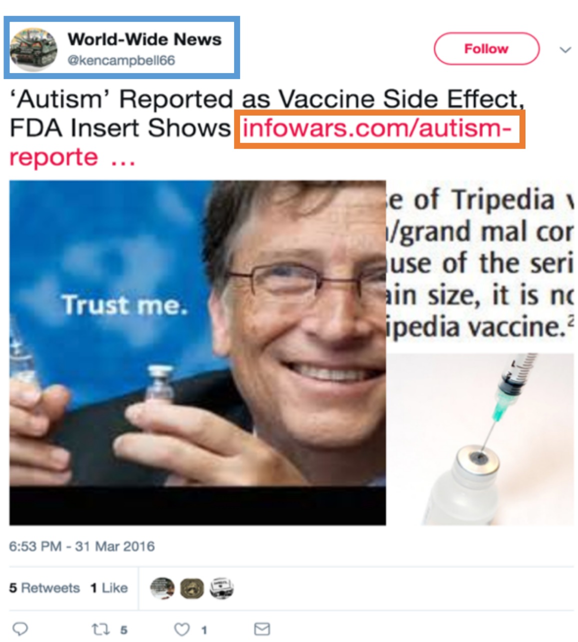
- **Technical domain:** The encryption debate
- **Health domain:** MMR vaccine debate
- **Social domain:** Women's driving in Saudi Arabia
- **Politics domain:** 2016 US presidential election

A tweet is the component a user read in Twitter, were each tweet is comprised of the following:

- **A URL:** allows the user to retain characters, and often provides analytic measurements. Links are optional as the overuse of them can create the appearance of a news feed rather than a humanized personality.

- **Tweet Text (content):** or the message the user wants to deliver.
- **Message writer:** The user who wrote the message.

An example of a tweet with these three components highlighted is shown in figure 1.



World-Wide News
@kencampbell66

'Autism' Reported as Vaccine Side Effect, FDA Insert Shows infowars.com/autism-reporte ...

6:53 PM - 31 Mar 2016

5 Retweets 1 Like

Figure 1: Tweet example

Figure 2 shows the overall workflow of this study. For each of the tweet components, there are two phases 1) identification and understanding phase of that component, 2) identify the features that can help with identifying and predicting that component. The figure also shows for each phase and component, what is the controversial topic that was used as a case study.

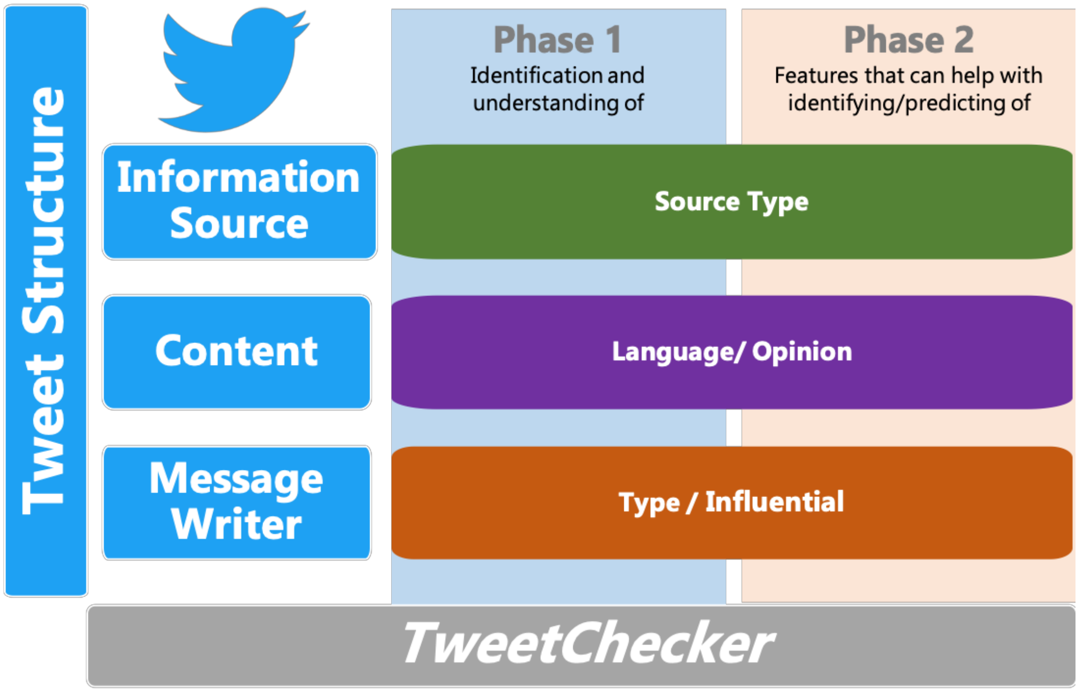


Figure 2: Thesis component

1.5 Definition of Terms

1.5.1 Social media:

Internet-based channels of mass personal communication that facilitate the perception of interaction among users, who derive value primarily from user-generated content (Carr & Hayes, 2015). Boyd (2014) uses the term social media to refer to the sites and services that emerged during the early 2000s, including social network sites, video sharing sites, blogging and micro blogging platforms and related tools that allow participants to create and share their own content.

1.5.2 A Controversial Topic:

A controversial topic in this study refers to a topic that generates disagreement or different opinions among large groups of people (Dori-Hacohen, Yom-Tov, & Allan, 2015).

1.6 Study Structure

The following chapter (chapter 2) presents a thorough review of the different methodologies used in this thesis. This chapter introduces the different types of natural language processing techniques used in this thesis and the approaches that can be applied to perform these analyses. Moreover, this chapter introduces the technologies used in this thesis to scale up these types of analyses. Chapter 3 outlines the studies that used the encryption debate as a use case of a controversial topic where online users' opinions are polarized towards an event. This chapter tries to understand who frequently participates in controversial discussions on social media. Moreover, correlating users' stance with their sentiments and demographics may help further describe users' behavior online.

Health controversial issues such as the debate towards MMR vaccines are prevalent in social media. In chapter 4, we use this issue as the case study for our analysis. This chapter discusses how scientific sources are used in Twitter when discussing such a controversial issue.

In chapter 5, the social movement generated in regard to women driving in Saudi Arabia is used as an example of a controversial issue. In this chapter we discuss how the identification of users' opinions toward such issues can help with not just having a better understanding of users' opinions towards the issues, but also, to understand the effects of other factors such as location and gender on user stance.

To combine all of the previous social media features, chapter 7 introduces a web application tool *TweetChecker* that allows online users to have a more in-depth understanding of the discussions towards five different health controversial issues. Results, recommendations and limitations of the research are discussed in Chapter 8, alongside with identification of future directions for this research.

CHAPTER 2: LITERATURE REVIEW AND METHODOLOGY

2.1 Introduction

As the term suggests, the methodology chapter underpins an integral section of a dissertation, which explains the approach or method that the researcher uses in research. The chapter has the primary purpose of presenting the philosophical assumptions of the study by introducing empirical techniques and research strategy used in the research.

By defining the scope of the research, this chapter helps in situating the researcher within the underpinnings of the existing research traditions so as to increase the reliability and validity of the information. In this view, the methodology chapter plays a key role in guiding the researcher through the entire research process (Young, Hazarika, Poria, & Cambria, 2018). The method used in this research also helps in increasing the reliability of the study findings to the extent that another researcher can follow the same process and get similar results. For the purpose of this dissertation, this chapter will focus on discussing different analysis methodologies that can help with gaining a better understanding of how people discuss controversial issues in social media, specifically Twitter, and how people's opinions are polarized towards these issues.

2.2 Analysis Methodologies

Social media analysis has become an important area of study with the growth of social networks. The analysis of social media is done through different methods including social network analysis and natural language processing (NLP). NLP is very effective since it helps in extracting structured data from the unstructured information on social networks to utilize the valuable information (Wilson, Wiebe, & Cardie, 2017). NLP comprises a subfield of artificial intelligence, information engineering, and computer science focused on understanding the interaction between natural human languages and computer systems. Most specifically, NLP focuses on how computers are programmed to analyze and process natural language data in large amounts (Kumar et al., 2016). Computers work with structured and standardized data such as financial records and database tables and often process the data at an amazing speed. In this section, different analysis methodologies will be discussed to show how programs give computers standardized techniques and a set of rules for processing natural language data.

2.2.1 Natural Language Processing

Natural language processing (NLP) has different applications which relate to artificial intelligence and the ability of computers to process, analyze, and understand human languages, as well as how they develop to the human-level of understanding languages (Mukhtar, Khan, & Chiragh, 2018). Some of the applications within the scope of NLP include sentiment analysis, opinion mining or stance analysis, and content analysis or text analysis, which are used to connect people's language to their behaviors, as their online behaviors.

2.2.1.1 Sentiment Analysis

This is a computational, automated process of mining a text contextually to identify, categorize, and extract subjective information and opinions in text. The primary purpose of undertaking sentiment analysis of a text is to classify its polarity at the feature, sentence, or document level (Kumar et al., 2016). The importance of sentiment analysis is that it helps in identifying and classifying neutral, negative, and positive opinions expressed in a feature, sentence, or document.

2.2.1.2 Stance Analysis

Stance analysis is different from sentiment analysis in that stance analysis classifies the speaker's position towards a topic while sentiment analysis categorizes the speaker's opinion towards a topic. Stance analysis involves determining the neutrality, support, or opposition of the author towards a proposition (Kucher, Schamp-Bjerede, Kerren, Paradis, & Sahlgren, 2016). Stance analysis, therefore, involves the process of textual entailment, text summarization, and information retrieval to identify the favorability of a speaker or text towards a certain target (Mukhtar et al., 2018). As such, it plays an instrumental role in determining if the author of a text is neutral, against, or in favor towards a given target by undertaking sentiment classification, subjectivity analysis, and argument mining.

2.2.1.3 Text Analysis / Content Analysis

Text analysis entails the process of analyzing the rhetorical concepts and features of a text. The technique is used to undertake a systemic evaluation of text whether in graphic, oral, or written form to interpret and code the text and make valid, replicable inferences of the material (Kumar et al., 2016). Text analysis is, therefore a method used to analyze the features and artifacts of text using rhetorical concepts to understand the larger conversation through a non-invasive way and without imitating social experiences.

2.2.2 Approaches to Natural Learning Processing

The above tasks can be applied through three main approaches which are commonly used to undertake any Natural Learning Processing (NLP) task including the lexicon-based approach, machine learning based approach, and the combined approach.

2.2.2.1 Lexicon Based

In the lexicon-based approach, a practical, viable, and simple technique is used to analyze human language data. The approach uses computational methods to identify and analyze the structure of phrases and words in a text for the purposes of opinion mining and sentiment analysis (Mukhtar et al., 2018). good example on how this approach has been used before include AltaVista, Word Net database to evaluate the sematic gaps between words. In this view, the lexicon-based approach for NLP takes advantage of the linguistic facts established in the paradigm of lexicon grammar theories to analyze a language. The greatest importance of this approach is that it is able to identify and gather meaningful information from a text and categorize the information thoroughly using semantic descriptions.

Based on the precepts of the semantic predicates theory, lexicon approach makes it possible to match the syntactic structures of a verb with the semantic information and attach the same to the lexical entry of a database (Gitari, Zuping, Damien, & Long, 2015). The lexicon-based approach can either be dictionary-based which involves the collection and annotation of terms manually or corpus-based which provides domain-related dictionaries that are created from a collection of opinion terms (Kucher et al., 2016). Some opinion terms include negating, swearing, social, and affection terms established in Linguistic Inquiry and Word Counts (LIWC) and expression-level, private states, sentiments, and multi-attribute of opinions (Wilson et al., 2017).

2.2.2.2 Machine Learning-Based Approach

NLP helps in developing computational algorithms which analyze and categorize human language automatically and is, therefore, very useful in teaching machines how to undertake tasks related to natural language like dialogue generation and machine translation (Jordan & Mitchell, 2015). Machine learning (ML) is, in this regard, a useful technique that enables the self-learning of computers by taking advantage of expert or rule-based systems that use coded rules to perform text analytics. A good example of the machine learning includes Naïve Byes algorithm and Support Vector Machines (Wilson et al., 2017). These techniques use different statistical techniques, which make it possible to identify and analyze sentiments, entities, and other features of text either through supervised or unsupervised techniques.

As a data analysis method, machine leaning automates the building of an analytical, classification model and then trains it using pre-labeled datasets of neutral, negative, or positive content (Kucher et al., 2016). The built model can either be a supervised machine learning which is used to other texts or it can be in the form of algorithms which operate across a wide range of datasets to extract meaningful information as unsupervised machine learning (Jordan & Mitchell, 2015). There are different types of supervised machine learning algorithms including random forest (RF), Neural Networks (Perceptron), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Trees (DT).

2.2.2.3 Combined Approach

The combined approach uses a combination of concepts borrowed from the machine learning model and lexicon approach (see figure 3).

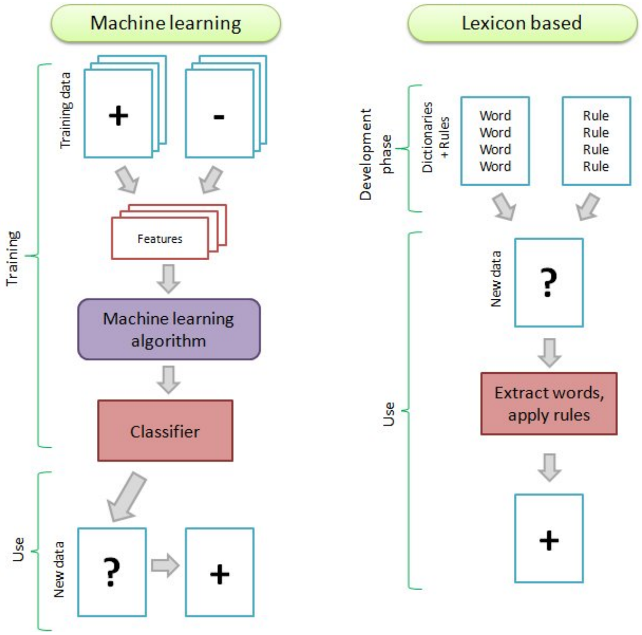


Figure 3: Combined approach (Taboada, 2016)

As seen in Figure 3 the combined approach uses a blend of pre-labeled datasets with lexicon dictionary to develop a classification model. Example of the hybrid approach is the multi-nominal naïve Bayes that is also known as the polling multinomial classifier (PMC).

2.3 Technologies Implemented

2.3.1 Cloud Computing

Cloud computing involves the use of internet-hosted remote servers in storing, managing, and processing data instead of a personal computer or a local server. Cloud computing, therefore, uses a pool of shared computer systems which are configured over the internet to manage data with minimal efforts (Botta, De Donato, Persico, & Pescapé, 2016). It is very useful as it helps in scaling how data is implemented by ensuring the delivery of computing resources on demand. On the other hand, cloud computing enables the storage of huge amounts of data using powerful computing resources which then makes analysis more efficient, timely, and quicker.

A good example of cloud computing is the Amazon Web Services (AWS) (Amazon, 2015) which provides cloud computing platforms on-demand to governments, companies, and individuals. AWS is provided by Amazon.com as a comprehensive platform for cloud computing. The comprehensive nature of the platform means that it has overarching capabilities which help it to offer a variety of cloud computing services like packaged software, platform, and infrastructure. AWS is also used in streaming of data to obtain information and real-time insights into business trends and customers' activities due to its capability to collect information from different data sources (Amazon, 2015). Based on its capability, AWS is preferred as it provides a wide array of remote computing services by enabling the hosting of various servers on its platform with heightened efficiency and in a timely, fast, reliable, and cost-effective manner.

2.4 Conclusion

This chapter discussed different analysis methodologies that can help with gaining a better understanding of how people discuss controversial issues in social media, specifically Twitter, and how people's opinions are polarized towards these issues. The research makes use of the Natural language processing (NLP) which includes sentiment analysis, opinion mining or stance analysis, and content analysis or text analysis, that are used to connect people's language to their behaviors, as their online behaviors. Approaches to Natural Learning Processing task include the lexicon-based approach, machine learning based approach, and the combined approach which uses a blend of pre-labeled datasets with lexicon dictionary to develop a classification model.

Technologies Implemented Cloud computing involves the use of internet-hosted remote servers in storing, managing, and processing data instead of a personal computer or a local server. Based on its capability, Amazon Web Services AWS is preferred as it provides a wide array of remote computing services which enable the hosting of various servers on its platform with heightened efficiency and in a timely, fast, reliable, and cost-effective manner.

CHAPTER 3: THE ENCRYPTION DEBATE IN TWITTER

3.1 Introduction

In recent years, social media has revolutionized how people communicate and share information. One function of social media, besides connecting with friends, is sharing opinions with others. Micro blogging sites, like Twitter, have often provided an online forum for social activism. When users debate about controversial topics on social media, they typically share different types of evidence to support their claims. Classifying these types of evidence can provide an estimate for how adequately the arguments have been supported.

To better understand online users' attitudes and opinions, we use stance classification in the second part of this chapter. Stance classification is a relatively new and challenging approach to deepen opinion mining by classifying a user's stance in a debate. Our stance classification use case is tweets that were related to the spring 2016 debate over the FBI's request that Apple decrypt a user's iPhone. In this "encryption debate," public opinion was polarized between advocates for individual privacy and advocates for national security. We propose a machine learning approach to classify stance in the debate, and a topic classification that uses lexical, syntactic, Twitter-specific, and argumentative features as a predictor for classifications.

Most of the contents in this chapter are published in two papers, the first is titled *"What is Your Evidence? A Study of Controversial Topics on Social Media"* at the proceedings of the third workshop on argumentation mining(Addawood & Bashir, 2016). The second is published in the Proceedings of the International Conference on Social Media & Society (#SMSociety17) under the title *"Stance Classification of Twitter Debates: The Encryption Debate as A Use Case"* (Addawood, Schneider, & Bashir, 2017). This is joint work with Masooda Bashir and Jodi Schneider.

3.2 Recognizing Evidence

Social media has grown dramatically over the last decade. Researchers have now turned to social media, via online posts, as a source of information to explain many aspects of the human experience (Gruzd & Goertzen, 2013). Due to the textual nature of online users' self-disclosure of their opinions and views, social media platforms present a unique opportunity for further analysis of shared content and how controversial topics are argued.

On social media sites, especially on Twitter, user text contains arguments with inappropriate or missing justifications—a rhetorical habit we do not usually encounter in professional writing. One way to handle such faulty arguments is to simply disregard them and focus on extracting arguments containing proper support (Cabrio & Villata, 2012; Villalba & Saint-Dizier, 2012). However, sometimes what seems like missing evidence is actually just an unfamiliar or different type of evidence. Thus, recognizing the appropriate type of evidence can be useful in assessing the viability of users' supporting information, and in turn, the strength of their whole argument.

One difficulty of processing social media text is the fact that it is written in an informal format. It does not follow any guidelines or rules for the expression of opinions. This has led to many messages containing improper syntax or spelling, which presents a significant challenge to attempts at extracting meaning from social media content. Nonetheless, we believe processing such corpora is of great importance to the argumentation-mining field of study. Therefore, the motivation for this study is to facilitate online users' search for information concerning controversial topics. Social media users are often faced with information overload about any given topic and understanding positions and arguments in online debates can potentially help users formulate stronger opinions on controversial issues and foster personal and group decision-making (Freeley & Steinberg, 2013).

Continuous growth of online data has led to large amounts of information becoming available for others to explore and understand. Several automatic techniques have allowed us to determine different viewpoints expressed in social media text, e.g., sentiment analysis and opinion mining. However, these techniques struggle to identify complex relationships between concepts in the text. Analyzing argumentation from a computational linguistics point of view has led very recently to a new field called argumentation mining (Green, Ashley, Litman, Reed, & Walker, 2014). It formulates how humans disagree, debate, and form a consensus. This new field focuses on identifying and extracting argumentative structures in documents. This type of approach and the reasoning it supports is used widely in the fields of logic, AI, and text processing (Mochales & Ieven, 2009). The general consensus among researchers is that an argument is defined as containing a claim, which is a statement of the position for which the claimant is arguing. The claim is supported with premises that function as evidence to support the claim, which then appears as a conclusion or a proposition (Toulmin, 2003; Walton, Reed, & Macagno).

One of the major obstacles in developing argumentation mining techniques is the shortage of high-quality annotated data. An important source of data for applying argumentation techniques is the web, particularly social media. Online newspapers, blogs, product reviews, etc. provide a heterogeneous and growing flow of information where arguments can be analyzed. To date, much of the argumentation mining research has been limited and has focused on specific domains such as news articles, parliamentary records, journal articles, and legal documents (Ashley & Walker, 2013; Hachey & Grover, 2005; Reed & Rowe, 2004) . Only a few studies have explored arguments on social media, a relatively under-investigated domain. Some examples of social media platforms that have been subjected to argumentation mining include Amazon online product reviews (Wyner, Schneider, Atkinson, & Bench-Capon, 2012) and tweets related to local riot events (Llewellyn, Grover, Oberlander, & Klein, 2014).

In our study, the researcher describes a novel and unique benchmark data set achieved through a simple argument model and elaborates on the associated annotation process. Unlike the classical Toulmin model (Toulmin, 2003), we search for a simple and robust argument structure comprising only two components: a claim and associated supporting evidence. Previous research has shown that a claim can be supported using different types of evidence (Rieke & Sillars, 1984). The annotation that is proposed is based on the type of evidence one uses to support a particular position on a given debate. We identify six types, which are detailed in the methods section (Section 3). To demonstrate these types, we collected data regarding the recent Apple/FBI encryption debate on Twitter between January 1 and March 31, 2016. We believe that understanding online users' views on this topic will help scholars, law enforcement officials, technologists, and policy makers gain a better understanding of online users' views about encryption.

3.3 Argumentation Mining

Argumentation mining is the study of identifying the argument structure of a given text. Argumentation mining has two phases. The first consists of argument annotations and the second consists of argumentation analysis. Many studies have focused on the first phase of annotating argumentative discourse. Reed and Rowe (Reed & Rowe, 2004) presented Araucaria, a tool for argumentation diagramming that supports both convergent and linked arguments, missing premises (enthymemes), and refutations. They also released the AracuariaDB corpus, which has been used for experiments in the argumentation-mining field. Similarly, Schneider et al. (Schneider, Samp, Passant, & Decker, 2013) annotated Wikipedia talk pages about deletion using Walton's 17 schemes (Walton et al.). Rosenthal and McKeown (2012) annotated opinionated claims, in which others should adopt the author expresses a belief they think. Two annotators labeled sentences as claims without any context. Habernal, Eckle-Kohler & Gurevych (2014) developed another well-annotated corpus, to model arguments following a variant of the Toulmin model. This dataset includes 990 instances of web documents collected from blogs, forums, and news outlets, 524 of which are labeled as argumentative. A final smaller corpus of 345 examples was annotated with finer-grained tags. No experimental results were reported on this corpus.

As far as the second phase, Stab and Gurevych (2014) classified argumentative sentences into four categories (none, major claim, claim, premise) using their previously annotated corpus and reached a 0.72 macro-F1 score. Park and Cardie (2014) classified propositions into three classes (unverifiable, verifiable non-experimental, and verifiable experimental) and ignored non-argumentative text. Using multi-class SVM and a wide range of features (n-grams, POS, sentiment clue words, tense, person) they achieved a 0.69 Macro F1. The IBM Haifa Research Group (Rinott et al., 2015) developed something similar to our research; they developed a data set using plain text in Wikipedia pages. The purpose of this corpus was to collect context-dependent claims and evidence, where the latter refers to facts (i.e., premises) that are relevant to a given topic. They classified evidence into three types (study, expert, anecdotal). Our work is different in that it includes more diverse types of evidence that reflect social media trends while the IBM Group's study was limited to looking into plain text in Wikipedia pages.

3.4 Social Media as a Data Source for Argumentation Mining

As stated previously there are only a few studies that have used social media data as a source for argumentation mining (Llewellyn et al., 2014) experimented with classifying tweets into several argumentative categories, specifically claims and counter-claims (with and without evidence), and used verification inquiries previously annotated by Procter, Vis, and Voss (2013). They used unigrams, punctuations, and POS as features in three classifiers. Schneider and Wyner (2012) focused on online product reviews and developed a number of argumentation schemes - inspired by Walton et al. (Walton et al.) - based on manual inspection of their corpus.

By identifying the most popular types of evidence used in social media, specifically on Twitter, our research differs from the previously mentioned studies because we are providing a social media annotated corpus. Moreover, the annotation is based on the different types of premises and evidence used frequently in social media settings.

3.5 Data Collection

This study uses Twitter as its main source of data. Crimson Hexagon (S Etlinger & W Amand, 2012), a public social media analytics company, was used to collect every public post from January 1, 2016 through March 31, 2016. Crimson Hexagon houses all public Twitter data going back to 2009. The search criterion for this study was searching for a tweet that contains the word "encryption" anywhere in its text. The sample only included tweets from accounts that set English as their language; this was filtered in when requesting the data. However, some users set their account language to English, but constructed some tweets in a different language. Thus, forty accounts were removed manually, leaving 531,593 tweets in our dataset.

Although most Twitter accounts are managed by humans, there are other accounts managed by automated agents called social bots or Sybil accounts. These accounts do not represent real human opinions. In order to ensure that tweets from such accounts did not enter our data set, in the annotation procedure, we ran each Twitter user through the Truthy BotOrNot algorithm (Davis, Varol, Ferrara, Flammini, & Menczer, 2016). This cleaned the data further and excluded any user with a 50% or greater probability of being a bot. Overall, 946 (24%) bot accounts were removed.

3.6 Coding Scheme

In order to perform argument extraction from a social media platform, we followed a two-step approach. The first step was to identify sentences containing an argument. The second step was to identify the evidence-type found in the tweets classified as argumentative. These two steps were performed in conjunction with each other. Annotators were asked to annotate each tweet as either having an argument or not having an argument. Then they were instructed to annotate a tweet based on the type of evidence used in the tweet. Figure 4 shows the flow of annotation.

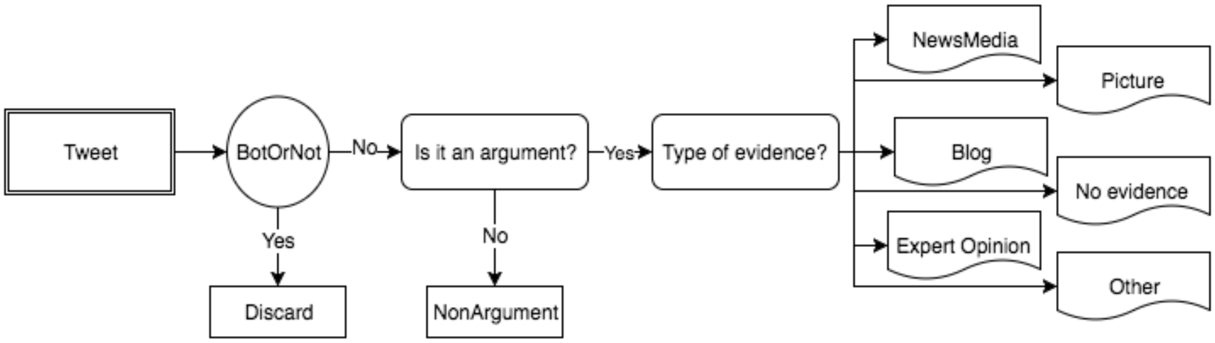


Figure 4: flow chart for annotation

After considerable observation of the data, a draft-coding scheme was developed for the most used types of evidence. In order to verify the applicability and accuracy of the draft-coding scheme, two annotators conducted an initial trial on 50 randomized tweets to test the coding scheme. After some adjustments were made to the scheme, a second trial was conducted consisting of 25 randomized tweets that two different annotators annotated. The resulting analysis and discussion led to a final revision of the coding scheme and modification of the associated documentation (annotation guideline). After finalizing the annotation scheme, two annotators annotated a new set of 3000 tweets. The tweets were coded into one of the following evidence types.

- **News media account (NEWS)** refers to sharing a story from any news media account. Since Twitter does not allow tweets to have more than 140 characters, users tend to communicate their opinions by sharing links to other resources. Twitter users will post links from official news accounts to share breaking news or stories posted online and add their own opinions. For example:

Please who don't understand encryption or technology should not be allowed to legislate it. There should be a test... https://t.co/I5zkvK9sZf

- **Expert opinion (EXPERT)** refers to sharing someone else’s opinion about the debate, specifically someone who has more experience and knowledge of the topic than the user. The example below shows a tweet that shares a quotation from a security expert.

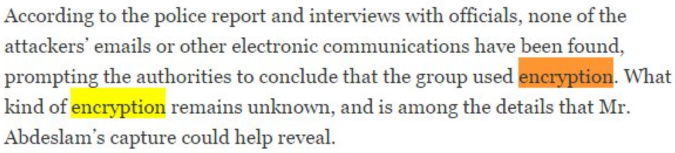
RT @ItIsAMovement "Without strong encryption, you will be spied on systematically by lots of people" - Whitfield Diffie

- **Blog post (BLOG)** refers to the use of a link to a blog post reacting to the debate. The example below shows a tweet with a link to a blog post. In this tweet, the user is sharing sharing a link to her own blog post.

  I care about #encryption and you should too. Learn more about how it works from @ Mozilla at https://t.co/RTFiuTQXyQ

- **Picture (PICTURE)** refers to a user sharing a picture related to the debate that may or may not support his/her point of view. For example, the tweet below shows a post containing the picture shown in figure 5.

RT @ ErrataRob No, morons, if encryption were being used, you'd find the messages, but you wouldn't be able to read them.



According to the police report and interviews with officials, none of the attackers' emails or other electronic communications have been found, prompting the authorities to conclude that the group used encryption. What kind of encryption remains unknown, and is among the details that Mr. Abdeslam's capture could help reveal.

Figure 5: an example of sharing a picture as evidence

- **Other (OTHER)** refers to other types of evidence that do not fall under the previous annotation categories. Even though we observed Twitter data in order to categorize different, discrete types of evidence, we were also expecting to discover new types while annotating. Some new types we found while annotating include audio, books, campaigns, petitions, codes, slides, other social media references, and text files.
- **No evidence (NO EVIDENCE)** refers to users sharing their opinions about the debate without having any evidence to support their claim. The example below shows an argumentative tweet from a user who is in favor of encryption. However, he/she does not provide any evidence for his/her stance.

I hope people ban encryption. Then all their money and CC's can be stolen and they'll feel better knowing terrorists can't keep secrets.

- **Non-Argument (NONARG)** refers to a tweet that does not contain an argument. For example, the following tweet asks a question instead of presenting an argument.

  RT @cissp_googling what does encryption look like.

Another NONARG situation is when a user shares a link to a news article without posting any opinions about it. For example, the following tweet does not present an argument or share an opinion about the debate; it only shares the title of the news article, “Tech giants back Apple against FBI's 'dangerous' encryption demand,” and a link to the article.

Tech giants back Apple against FBI's 'dangerous' encryption demand #encryption https://t.co/4CUushsVmW

Retweets are also considered NONARG because simply selecting “retweet” does not take enough effort to be considered an argument. Moreover, just because a user retweets something does not mean we know exactly how they feel about it; they could agree with it, or they could just think it was interesting and want to share it with their followers. The only exception would be if a user retweeted something that was very clearly an opinion or argument. For example, someone retweeting Edward Snowden speaking out against encryption backdoors would be marked as an argument. By contrast, a user retweeting a CNN news story about Apple and the FBI would be marked as NONARG.

Annotation discussion. While annotating the data, we observed other types of evidence that did not appear in the last section. We assumed users would use these types of evidence in argumentation. However, we found that users mostly use these types in a non-argumentative manner, namely as a means forwarding information. The first such evidence type was “scientific paper,” which refers to sharing a link to scientific research that was published in a conference or a journal. Here is an example:

A Worldwide Survey of Encryption Products. By Bruce Schneier, Kathleen Seidel & Saranya Vijayakumar #Cryptography https://t.co/wmAuvu6oUb.

The second such evidence type was “video,” which refers to a user sharing a link to a video related to the debate. For example, the tweet below is a post with a link to a video explaining encryption.

An explanation of how a 2048-bit RSA encryption key is created https://t.co/JjBWym3poh.

3.7 Annotation results

The results of the annotation are shown in Table 1 and Table 2.

Table 1: Argumentation classification distribution over tweets

Argumentation classification	Class distribution
Argument (ARG)	1,271
Non-argument (NONARG)	1,729
Total	3000

Table 2: Evidence type distribution over tweets

Evidence type	Class distribution
No evidence	630
News media accounts	318
Blog post	293
Picture	12
Expert opinion	11
Other	7
Total	1,271

Table 1 and Table 2 show that the inter-coder reliability was 18% and 26% for the two tasks, respectively, yielding a 70% inter-annotator observed agreement for both tasks. The un-weighted Cohen's Kappa score was 0.67 and 0.79, respectively, for the two tasks.

3.8 Experimental Evaluation

We developed an approach to classify tweets into each of the six major types of evidence used in Twitter arguments.

3.8.1 Preprocessing

Due to the character limit, Twitter users tend to use colloquialisms, slang, and abbreviations in their tweets. They also often make spelling and grammar errors in their posts. Before discussing feature selection, we will briefly discuss how we compensated for these issues in data preprocessing. We first replaced all abbreviations with their proper word or phrase counterparts (e.g., 2night => tonight) and replaced repeated characters with a single character (e.g., haaaapy => happy). In addition, we lowercased all letters (e.g., ENCRYPTION => encryption), and removed all URLs and mentions to other users after initially recording these features.

3.8.2 Features

We propose a set of features to characterize each type of evidence in our collection. Some of these features are specific to the Twitter platform. However, others are more generic and could be applied to other forums of argumentation. Many features follow previous work (Agichtein, Castillo, Donato, Gionis, & Mishne, 2008; Castillo, Mendoza, & Poblete, 2011) . The full list of features appears in appendix A. In table 31, we identify four types of features based on their scope: Basic, Psychometric, Linguistic, and Twitter-specific.

- **Basic Features** refer to N-gram features, which rely on the word count (TF) for each given unigram or bigram that appears in the tweet.
- **Psychometric Features** refer to dictionary-based features. They are derived from the linguistic enquiry and word count (LIWC). LIWC is a text analysis software originally developed within the context of Pennebaker's work on emotional writing (Pennebaker, 1997; Pennebaker & Francis, 1996). LIWC produces statistics on eighty-one different text features in five categories. These include psychological processes such as emotional and social cognition, and personal concerns such as occupational, financial, or medical worries. In addition, they include personal core drives and needs such as power and achievement.
- **Linguistic Features** encompass four types of features. The first is grammatical features, which refer to percentages of words that are pronouns, articles, prepositions, verbs, adverbs, and other parts of speech or punctuation. The second type is LIWC summary variables. The newest version of LIWC includes four new summary variables (analytical thinking, clout, authenticity, and emotional tone), which resemble "person-type" or personality measures. The LIWC webpage[1] describes the four summary variables as follows. Analytical thinking "captures the degree to which people use words that suggest formal, logical, and hierarchical thinking patterns." Clout "refers to the relative social status, confidence, or leadership that people display through their writing or talking." Authenticity "is when people reveal themselves in an authentic or honest way," usually by becoming "more personal, humble, and vulnerable." Lastly, with emotional tone, "although LIWC includes both positive emotion and negative emotion dimensions, the tone variable puts the two dimensions into a single summary

[1] http://liwc.wpengine.com/

variable." The third type is sentiment features. We first experimented with the (Wilson, Wiebe, & Hoffmann, 2005) subjectivity clue lexicon to identify sentiment features. However, we decided to use the sentiment labels provided by the LIWC sentiment lexicon. We found that it provides more accurate results than we would have had otherwise. For the final type, subjectivity features, we did use the Wilson et al. (2005) subjectivity clue lexicon to identify the subjectivity type of tweets.

- **Twitter-Specific Features** refer to characteristics unique to the Twitter platform, such as the length of a message and whether the text contains exclamation points or question marks. In addition, these features encompass the number of followers, number of people followed ("friends" on Twitter), and the number of tweets the user has authored in the past. Also included is the presence or not of URLs, mentions of other users, hashtags, and official account verification. We also considered a binary feature for tweets that share a URL as well as the title of the URL shared (i.e., the article title).

3.9 Experimental Results

Our first goal was to determine whether a tweet contains an argument. We used a binary classification task in which each tweet was classified as either argumentative or not argumentative. Some previous research skipped this step (V. W. Feng & Hirst, 2011), while others used different types of classifiers to achieve a high level of accuracy (Palau & Moens, 2009).

In this study, we chose to classify tweets as either containing an argument or not. Our results confirm previous research showing that users do not frequently utilize Twitter as a debating platform (Smith, Zhu, Lerman, & Kozareva, 2013). Most individuals use Twitter as a venue to spread information instead of using it as a platform through which to have conversations about controversial issues. People seem to be more interested in spreading information and links to webpages than in debating issues.

As a first step, we compared classifiers that have frequently been used in related work: Naïve Bayes (NB) approaches as used in Teufel and Moens (Teufel & Moens, 2002), Support Vector Machines (SVM) as used in Liakata et al. (Liakata, Saha, Dobnik, Batchelor, & Rebholz-Schuhmann, 2012), and Decision Trees (J48) as used in Castillo, Mendoza, & Poblete (Castillo et al., 2011). We used the Weka data mining software as used in Hall et al. (Hall et al., 2009) for all approaches.

Before training, all features were ranked according to their information gain observed in the training set. Features with information gain less than zero were excluded. All results were subject to 10-fold cross-validation. Since, for the most part, our data sets were unbalanced, we used the "Synthetic Minority Oversampling TEchnique" (SMOTE) approach (Nitesh V. Chawla, Bowyer, Hall, & Kegelmeyer, 2002). SMOTE is one of the most renowned approaches to solve the problem of unbalanced data. Its main function is to create new minority class examples by interpolating several minority class instances that lie together. After that, we randomized the data to overcome the problem of over-fitting the training data.

- **Argument classification.** Regarding our first goal of classifying tweets as argumentative or non-argumentative, Table 3 shows a summary of the classification results.

Table 3: Summary of the argument classification results in percentage

Feature Set	Decision tree			SVM			NB		
	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
UNI (Base)	72.5	69.4	66.3	81	78.5	77.3	69.7	67.3	63.9
All features	87.3	87.3	87.2	89.2	89.2	**89.2**	79.3	79.3	84.7

The summary of the argument classification in Table 3 shows that the best overall performance was achieved using SVM, which resulted in a 89.2% F_1 score for all features compared to basic features, unigram model. We can see there is a significant improvement from just using the baseline model.

- **Evidence type classification.** Our second goal was for evidence type classification, results across the training techniques were comparable. Table 4 is a summary of the evidence type classification results in percentage.

Table 4: Summary of the evidence type classification results in %

Feature Set	Decision tree			SVM			NB		
	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
UNI (Base)	59.1	61.1	56.3	63.7	62.1	56.5	27.8	31.6	19.4
All features	76.8	77	76.9	78.5	79.5	78.6	62.4	59.4	52.5

Table 4 shows a summary of the classification results. It appears that the best results were again achieved by using SVM, which resulted in a 78.6% F1 score. The best overall performance was achieved by combining all features.

In table 5, we computed Precision, Recall, and F1 scores with respect to the top-used three evidence types, employing one-vs-all classification problems for evaluation purposes. We chose the top-used evidence types since other types were too small and could have led to biased sample data.

Table 5: Summary of evidence type classification results using one-vs-all in %

Feature Set	NEWS vs. All			BLOG vs. All			NO EVIDENCE vs. All			Macro Average F1
	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	
UNI (Base)	76.8	74	73.9	67.3	64.4	63.5	78.5	68.7	65.6	67.6
Basic Features	842	81.3	81.3	85.2	83	82.9	80.1	75.5	74.4	79.5
Psychometric Features	62	61.7	57.9	64.6	63.7	63.5	59.2	58.9	58.6	60
Linguistic Features	65	65.3	64.2	69.1	69	69	63.1	62.6	62.4	65.2
Twitter-Specific Features	65.7	65.2	65	63.7	63.6	63.6	68.7	68.1	67.9	65.5
All features	84.4	84	84.1	86	85.2	85.2	79.3	79.3	79.3	82.8

Table 6: Most informative features for combined features for evidence type classification

Feature set	Features
Unigram	I'm, surveillance, love, I've, I'd, privacy, I'll, hope, wait, obama
All	1st person singular, RT, personal pronouns, URL, function words, user mention, followers, auxiliary verbs, verb, analytic

The results in Table 5 show that the SVM classifier achieved a F_1 macro-averaged score of 82.8%. The baseline outperformed Linguistic and Psychometric features. This was not expected. However, Basic features (N-gram) had very comparable results to those from combining all features. In other words, the combined features captured the characteristics of each class. This shows that we can distinguish between classes using a concise set of features with equal performances.

3.10 Feature Analysis

The most informative features for the evidence type classification are shown in Table 6. There are different features that work for each class. For example, Twitter-specific features such as title, word count, and WPS are good indicators of the NEWS evidence type. One explanation for this is that people often include the title of a news article in the tweet with the URL, thereby engaging the aforementioned Twitter-specific features more fully.

Table 7: Most informative features argumentation classification

Feature Set	All Features
NEWS vs. All	Word count, title, personal pronoun, common adverbs, WPS, “iphone”, “nsa director”
BLOG vs. All	Emotional Tone, 1st person singular, negation, colon, conjunction, “wrote”, negative emotions, “blog”
NO EVIDENCE vs. All	Title,1st person singular, colon, Impersonal pronouns, discrepancies, insight, differentiation (cognitive processes), period, adverb, positive emotion

Another example is that linguistic features like grammar and sentiments are essential for using the BLOG evidence type. The word “wrote,” especially, appears often to refer to someone else’s writing, as in the case of a blog. The use of the BLOG evidence types also seemed to correlate with emotional tone and negative emotions, which is a combination of positive and negative sentiment. This may suggest that users have strong negative opinions toward blog posts.

Concerning the NO EVIDENCE type, a combination of linguistic features and psychometric features best describe the classification type. Furthermore, in contrast with blogs, users not using any evidence tend to express more positive emotions. That may imply that they are more confident about their opinions. There are, however, mutual features used in both BLOG and NO EVIDENCE types as 1st person singular and colon. One explanation for this is that since blog posts are often written in a less formal, less evidence-based manner than news articles, they are comparable to tweets that lack sufficient argumentative support. One further shared feature is that “title” appears frequently in both NEWS and NO EVIDENCE types. One explanation for this is that “title” has a high positive value in NEWS, which often involves highlighting the title of an article, while it has a high negative value in NO EVIDENCE since this type does not contain any titles of articles.

As Table 5 shows, “all features” outperforms other stand-alone features and “basic features,” although “basic features” have a better performance than the other features. Table 7 shows the most informative feature for the argumentation classification task using the combined features and unigram features. We can see that first-person singular is the strongest indication of arguments on Twitter, since the easiest way for users to express their opinions is by saying “I …”.

3.11 Conclusion

In this study, we have presented a novel task for automatically classifying argumentation on social media for users discussing controversial topics like the recent FBI and Apple encryption debate. We classified six types of evidence people use in their tweets to support their arguments. This classification can help predict how arguments are supported. We have built a gold standard data set of 3000 tweets from the recent encryption debate. We find that Support Vector Machines (SVM) classifiers trained with n-grams and other features capture the different types of evidence used in social media and demonstrate significant improvement over the unigram baseline, achieving a macro-averaged F1 score of 82.8 %. One consideration for future work is classifying the stance of tweets by using machine learning techniques to understand a user’s viewpoint and opinions about a debate. Another consideration for future work is to explore other evidence types that may not be presented in our data.

3.12 Stance Classification of Twitter Debates: The Encryption Debate as A Use Case

Researchers have turned to user-generated content in social media as a source of information to explain many aspects of human experience (Gruzd & Goertzen, 2013). Due to the often-textual nature of online users’ self-disclosure of their opinions and views, social media platforms present a unique opportunity to analyze shared content and, in particular, how controversial topics are argued. Continuous growth of online data has led to large amounts of information becoming available for others to explore and understand. For instance, Twitter has grown dramatically since its introduction over a decade ago to become one of the world’s most popular social media platforms. Today, more than 288 million people actively use the site on a monthly basis (W. Wang, Hernandez, Newman, He, & Bian, 2016).

Automatic techniques such as sentiment analysis and opinion mining have allowed researchers and business people to determine the different viewpoints expressed in social media text (e.g. (Pang, Lee, & Vaithyanathan, 2002)). As their main task, these approaches assign a polarity score to an opinion that is presented in an online format. Although it is important to determine whether a user's opinion is positive or negative, it is even more essential to determine the user's position toward a specific topic (Mohammad, Kiritchenko, Sobhani, Zhu, & Cherry, 2016).

Stance classification offers complementary information to sentiment analysis. Given a collection of debate-style discussions on a controversial topic, stance classification seeks to identify a user's attitudes toward the topic. This can support the identification of the user's affiliation with social or political groups, help develop better user-targeted recommendation systems, or tailor a user's information preferences to match his or her ideologies and beliefs (Abu-Jbara, Diab, Dasigi, & Radev, 2012; Anand et al., 2011; Gawron et al., 2012; Hasan & Ng, 2013; Qiu, Yang, & Jiang, 2013). Automatic stance classification can be used in applications such as information retrieval, text summarization, opinion summarization, and textual entailment. Over the last decade, there has been active research in modeling stance.

However, most of the work has focused on congressional debates (Thomas, Pang, & Lee, 2006) or debates in online forums (Anand et al., 2011; Hasan & Ng, 2013; Teufel & Moens, 2002; Walker, Anand, Abbott, & Grant, 2012). Compared to these domains, Twitter is a much more challenging domain for stance prediction. Tweets are written in an informal format; they do not follow any guidelines or rules for the expression of opinions. Many messages contain unconventional syntax and spelling, which present a significant challenge to attempts at extracting meaning (Reyes, Rosso, & Buscaldi, 2012; Riloff et al., 2013). In this work, we investigate whether two argumentative features are beneficial for ideological stance classification and detect stance in one ideological debate—encryption in the United States, as discussed on Twitter following a high-profile event.

This particular online debate was kindled after the San Bernardino, California terrorist attack, which occurred in December 2015 (Lee, 2016, February 18). For weeks following the attack, Apple Inc., one of the most well-known technology companies in the U.S., refused to create a "backdoor" that would give the Federal Bureau of Investigation (FBI) access to the encrypted iPhone of the alleged terrorist. Apple's refusal to comply with the FBI request gave rise to what we call the "encryption debate". This debate found its way into the mainstream media and became a popular topic of social media debate for months. It provoked reactions from IT experts, politicians, and technologists as well as the public. Although this debate continues both offline and online, in this study we focus on the online encryption debate that occurred on Twitter from January 1 through March 31, 2016. We selected this date range since it included tweets from the debate before and after a federal judge ordered that Apple unlock the iPhone on February 16, 2016 (Lee, 2016, February 18).

We were motivated to choose this use case because the tension between individual right to privacy and national security has long been of interest to philosophical, political, and technological debates. Those who favor national security argue that good citizens who have "nothing to hide" should not fear government surveillance and that law enforcement should have access to their information whenever necessary. Those who favor individual right to privacy argue for limiting government surveillance and access to personal information. As our mobile devices contain increasingly sensitive information and intricate details about our lives, the debate over whether information from these devices should be made available to law enforcement has become heated. Thanks to technological advances, many mechanisms have been developed to secure information to prevent unauthorized access. One of the most robust mechanisms, cryptography (i.e., encryption), allows messages to be sent confidentially. It is the use of the Advanced Encryption Standard that makes the iPhone such a formidable device to crack**.**

In this study, we explore whether classifying stance in an ideological debate can determine how frequently each position is expressed in Twitter and what attitudes users express. We also explore which features can enhance the stance classification task. We describe a novel benchmark dataset of tweets that we labeled by both the topic of discussion and the user's stance towards that topic. The annotation is based on the stance that a user has expressed toward one of two topics: individual right to privacy and national security. Compared to our earlier work on argumentation mining of tweets (Addawood & Bashir, 2016), we use an additional layer of manual annotation to indicate the stance expressed about the main topics of discussion and perform a detailed analysis on the annotation results from both human annotators and automatic prediction. As we discuss below, we found that the argumentativeness of the tweet and its tone are suitable features for predicting the stance of the tweet. Figure 6 summarizes the workflow of this study.

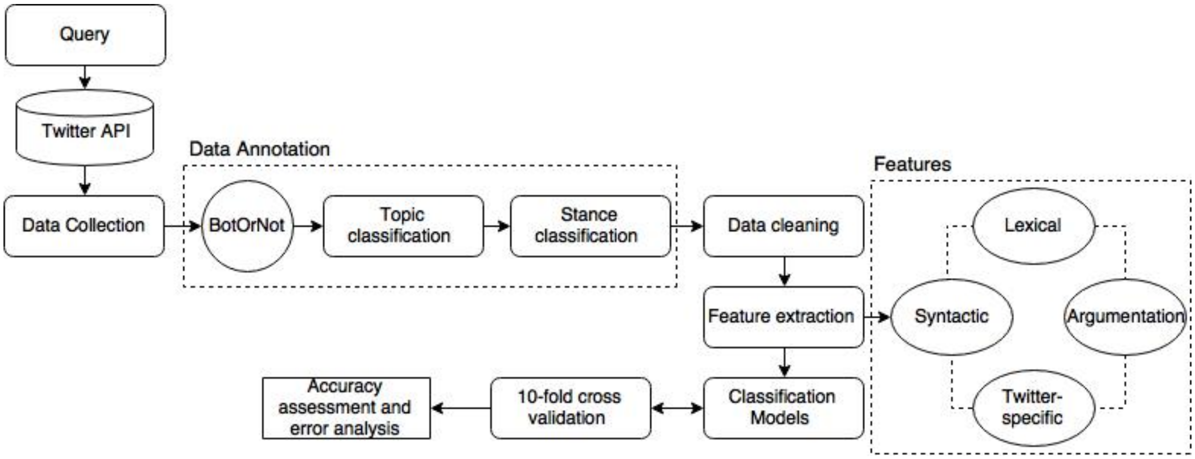


Figure 6: Project workflow

3.13 Stance Classification Approaches

Supervised machine learning has been used in almost all of the current approaches to stance classification. One of the first studies related to stance classification dealt with perspective identification. Lin, Wilson, and Hauptmann (2006) used articles from the Bitter-Lemons website, which discusses the Palestinian-Israeli conflict from each side's point of view, to train a system to perform automatic perspective detection on sentence and document levels. Later, Anand et al. (2011) deployed a rule-based classifier with several features such as unigrams, bigrams, punctuation marks, syntactic dependencies, and the dialogic structure of posts from a competitive debating site. Their results ranged from 54% to 69% accuracy. Somasundaran and Wiebe (2010) created a lexicon for detecting argument trigger expressions and subsequently leveraged it to identify arguments. These extracted arguments, together with sentiment expressions and their targets, were used in a supervised learner as features for stance classification. This experimental work included both argument and sentiment features from four datasets—abortion, creationism, gun rights, and gay rights—each containing news articles from a wide variety of sources. Their overall accuracy result was 63.93%. Murakami and Raymond (2010) identified general user opinions in online debates, distinguishing between global positions (opinions on a topic) and local positions (opinions on previous remarks). By calculating the degree of disagreement between any two users from the link structure and the text of each pair of their adjacent replies. Faulkner (2014) investigated the problem of detecting document-level stance in student essays; their key features are (1) stance-taking clauses (in a generalized format that tracks long-distance dependencies, which they call part-of-speech-generalized stance proposition subtrees); and (2) reuse of words from the essay prompt. Sobhani, Inkpen, and Matwin (2015) detected and classified stance starting by extracting online news comments using topic modeling.

To date, stance classification research has mainly focused on specific domains and mediums. Only a few studies have explored stance classification on social media. For example, Rajadesingan and Liu's study (2014) used Twitter-based stance classification. The authors proposed a retweet-based label propagation method which starts from a set of known opinionated users and labels the tweets posted by the people in their retweet network. By contrast, in this work, we focus on detecting stance, as well as possible, from a single tweet starting from a set of labeled tweets. Mohammad, Kiritchenko, Sobhani, Zhu, and Cherry's (2015) study also used Twitter as a dataset for stance classification. Their aim was to determine user stance (favor, against, or no position) in tweets on five selected topics: abortion, atheism, climate change, feminism, and Hillary Clinton. This dataset was made available for SemEval 2016, with two tasks. Task A was a traditional supervised classification task where 70% of the annotated data for a target is used as training and the rest for testing. The highest classification F-score for Task A was 67.82, with 19 teams participating. For Task B, test data was all of the instances for a new target (not used in Task A) and no training data was provided. The highest F-score for Task B was 56.28 with 9 teams participating. The dataset was offered to task participants without any context such as conversational structure or tweet metadata, which made classification challenging. In contrast, our approach for determining stance in this study takes into consideration tweet metadata (e.g., number of followers) as well as tweet labels that indicate a specific topic identified for the encryption debate.

3.14 Data Acquisition

In this research, we use publicly available social media data from Twitter. The initial dataset was originally gathered to investigate the classification of argumentative tweets (Addawood & Bashir, 2016). This dataset was composed of 3000 tweets from the encryption debate which we collected and then hand-annotated as we describe below. First we collected every public post on Twitter from January 1, 2016 through March 31, 2016 sent from accounts that set English as their language: 531,633 tweets in total, which we collected using Crimson Hexagon (S. Etlinger & W. Amand, 2012), a social media analytics platform that provides paid firehose access. We then filtered this data in several ways. We manually removed forty tweets that were in another language even though the accounts language was set to English. This left 531,593 tweets in our data set. Since we were only interested in real human opinions (not social bots or Sybil accounts), we excluded any user with a 50% or greater probability of being a bot based on the Truthy BotOrNot algorithm (Davis et al., 2016). Overall, 946 tweets by bot accounts were removed. The total number of tweets after all adjustments was 530,647 tweets.

3.15 Data Annotation

3.15.1 Codebook Development and Annotation Schema

We used a data-driven and theoretically grounded approach to develop a practical solution to stance classification. We randomly selected a small sample of our corpus, 30 tweets, for close reading done by the first author. Our annotation outline consisted of two segments: topic classification and stance classification for each tweet. These two tasks were performed manually in conjunction with each other. We used an iterative process for developing the codebook. Initially, we developed three stance classifications and three topic classes (from the three most frequently discussed topics relevant to the debate) which are information privacy, national security and right to encryption. Two human annotators were trained through discussions with the first author to label 100 tweets in each of three iterations which created a total of 300 tweets as the development set. After each iteration, we had an extensive discussion of the challenges and limitations of the codebook. The resulting analysis led to a final revision of the coding scheme and modification of the associated codebook.

Table 8 contains a short overview of the codebook, showing specific definitions and example tweets. For topic classification, the final codebook had two main topics: information privacy and national security. We excluded the topic ‘right to encryption’ from our codebook since we realized, after discussions with annotators, that it was too generic and could cover both information privacy and national security. Moreover, users’ attitudes toward the encryption debate seemed polarized into those who valued either individual privacy or national security more highly. We added two additional categories to incorporate other types of tweets: those that shared news without expressing opinions about the two main topics (‘other’); and those that contained jokes or nonsense (‘irrelevant’). The final category scheme thus had four topic classifications:

‘individual privacy’, ‘national security’, ‘other’, and ‘irrelevant’. For stance classification, three possible positions toward each topic were considered: ‘favor,’ ‘against,’ and ‘neutral.’

Table 8: Excerpt from codebook

	Class	Description	Example
Topic	National security	Government should protect the state and its citizens against all kinds of "national" crises related to the public’s/the whole nation’s interests.	“I'm against backdoors, not against trying to hack a murderous terrorists encrypted phone #encryption”
	Individual privacy	“The right to be let alone” [40].	“I'm oddly paranoid of people reading my phone over my shoulder. Some day I will need to design personally language for encryption.”
	Other	Tweets that don’t talk about national security or individual privacy, but are somewhat related to encryption. OR Tweets that are copies of news article titles without any comments.	“End to end encryption: when will it be universal as a safe communication mode?” “Tim Cook Wants a Government Commission to Settle the War Over iPhone Encryption https://t.co/NshUf43f9b #TechNews”
	Irrelevant	Tweets that are completely unrelated to encryption: jokes, nonsense.	“Apple: ‘Okay, here's the deal. We'll give you backdoor encryption, but you have to go through iTunes.’”
Stance	Favor	Tweets that support one of topics by reacting positively or showing positive sentiments toward the topic or expressing their agreement.	“I'm oddly paranoid of people reading my phone over my shoulder. Some day I will need to design personally language for encryption.”
	Against	Tweets that oppose one of topics by reacting negatively or showing negative sentiments toward the topic or expressing their disagreement.	“@QuadPiece But why encryption in the first place? It's not realistically more secure, it's just slower.”
	Neutral	Tweets that ask questions OR Tweets that neither support nor oppose any of the topics or that do not show any positive or negative sentiments toward the topic.	“I'm really torn on this phone encryption issue. #justsaying”

To start the annotation process of the 3000 tweets, we instructed our two annotators to first annotate each tweet based on the topic to which it was most related (topic classification), and to then annotate the posting user’s overall position toward the topic (stance classification). By the end of the three iterations, approximately 33% (990) of the 3000 tweets was labeled by both coders. We used Cohen’s Kappa (Cohen, 1960) to measure inter-annotator agreement. Our annotation consisted of two separate tasks, and the inter-coder reliability was 81.30% for topic classification and 87% on stance classification. The unweighted Cohen's Kappa score was 70% for topic classification and 64% for stance classification.

3.15.2 Annotation Challenges

We faced many challenges while annotating the tweets; some tweets needed special handling. We categorized tweets placed into the 'other' category as neutral since they did not provide an opinion, opposing or favoring, any of the topics analyzed. In this classification, we did not consider the stance of the article that was linked. For example, the following tweet does not represent a stance or share an opinion about the debate; it only shares the title of a news article and a link to it:

Amazon backtracks, decides to bring encryption back to Fire OS https://t.co/gK0I4tXn9l #tech.

We cannot be certain how users feel about items they choose to retweet and we generally classified most retweets as neutral. For instance, a user's retweeting of a CNN news story about an exchange between Apple and the FBI was marked as neutral. However, when a user retweeted something that very clearly expressed a stance, we counted the tweet as having that stance. For example, someone retweeting Edward Snowden speaking out against encryption backdoors would be marked as having a stance in 'favor' of the topic 'individual privacy'. We classified tweets that were completely unrelated to the debate as irrelevant; we did not consider it necessary to evaluate a user's stance in an irrelevant tweet. Tweets categorized as irrelevant were later excluded from the dataset, because they had no impact on the classification.

3.16 Corpus Analysis

We manually labeled 3,000 tweets in total. The distribution of topics over the three stance labels is illustrated in Table 9. Additionally, this table shows the number of occurrences of each topic in the corpus. We can see that 'individual privacy' had a higher number of tweets than 'national security'. Table 10 provides an overview of the stance labels in this corpus.

Table 9: Distribution of topics labels in the corpus

Topic classification	Class distribution	Percentage
Individual privacy	329	10.96%
National security	25	0.83%
Other	2,505	83.5%
Irrelevant	141	4.7%

From this table, we can see that the neutral stance classification has the highest value. This echoes previous research that found that users do not frequently use Twitter as a debating platform (Smith et al., 2013). Rather, most individuals use Twitter as a venue to spread information and share links to web pages instead of using it as a platform through which to have conversations about controversial issues. The results also illustrate that very few tweets were classified as being ‘against’ one of the topics.

Table 10: Distribution of stance labels in the corpus

Stance classification	Class distribution	Percentage
Favor	345	11.5%
Against	8	0.27%
Neutral	2,647	88.2%

3.17 Experimental Setup

3.17.1 Preprocessing

Due to character limits, Twitter users tend to use colloquialisms, slang, and abbreviations. They also often make spelling and grammar errors. Before discussing feature selection, we will briefly discuss how we compensated for these issues in data preprocessing. First, we tokenized tweets using the ARK Tweet NLP tokenizer (Gimpel et al., 2011). This Twitter-specific tokenizer segments tweet features such as emoticons, hashtags, and mentions. We replaced emoticons with their sentiment polarity. Next, we replaced abbreviations with their whole word or phrase counterparts (e.g., 2night => tonight). We then removed duplicated vowels in the middle of words (e.g., haaaapy). Any letter occurring more than two times in a row was replaced with exactly two occurrences. Inspired by (Addawood & Bashir, 2016), this modification significantly reduced feature space. Finally, we lowercased all letters (e.g., ENCRYPTION => encryption) and removed URLs and mentions to other users, after first recording these features.

3.17.2 Features

Based on prior work (Agichtein et al., 2008; Castillo et al., 2011), we chose four types of features: lexical, syntactic, Twitter-specific, and argumentation. Table 11 provides a summary of the features we extracted for each tweet. Below, we describe and explain the motivation for these feature sets.

Table 11: Feature types used in our model

Type	Feature	Description
Lexical	Unigram	Word count for each single word that appears in the tweet
	Bigram	Word count for every two words that appear in the tweet
Syntactic	Sentiment	Positive, negative, or neutral sentiment
	Subjectivity	Strong, weak, or neutral subjectivity
	Grammatical	Number of occurrences of noun, verb, adjective, preposition, adverb, and pronoun
Twitter-specific	Retweet	1.0 if the tweet is a retweet
	Title	1.0 if the tweet contains the title to an article
	Mention	1.0 if the tweet contains a mention to another user "@"
	Verified account	1.0 if the author has a "verified" account
	URL	1.0 if the tweet contains a link to a URL
	Followers	Number of people this user is following at posting time
	Following	Number of people following this user at posting time
	Posts	Total number of user's posts
	Hashtag	1.0 if the tweet contains a hashtag "#"
Argumentation	Argumentativeness	1.0 if the tweet is argumentative
	Source type	Type of source used in the tweet

In the following sections, we propose a set of features to characterize stance in tweets. Much of our work uses lexical features, which can help find words that are both highly salient and highly informative in a text or text set. This process also entails the removal of a) non-content-bearing words that dominate with respect to the cumulative power-law distribution of word frequencies and b) highly rare words in a collection. After preprocessing the data, we considered salient unigrams and bigrams, removing stop words and removing any word with fewer than five occurrences. Previous work suggests that the unigram baseline can be difficult to beat for certain types of debates (Somasundaran & Wiebe, 2010). Thus, we used both unigrams and bigrams as features. We kept the top 500 unigrams and the top 300 bigrams according to the TF-IDF metric as shown in Equation 3.

$$TF(t) = tf(t, d) \qquad (1)$$

$$IDF(t) = \log\left(\frac{|D|}{1+|\{d: t \in d\}|}\right) \qquad (2)$$

$$tf - IDF(t) = TF * IDF \qquad (3)$$

In these formulas, t is a term, d is the document in which t occurs, and D is the document space (collection of documents). Equation 1 shows the term frequency of word t, Equation 2 the inverse document frequency, and Equation 3 the TF-IDF score calculation for term t.

3.17.2.1 Syntactic Features

Syntactic features describe the relationship between words and their roles in a sentence, as the subjectively connoted adjectives and other modifiers, sentiment, and the ratio of different parts of speech in a sentence. In natural language processing, these characteristics are standard features for machine learning.

- **Sentiment.** After experimenting with other sentiment analysis dictionaries such as the Subjectivity Lexicon (Wilson et al., 2005), we selected the sentiment labels provided by Crimson Hexagon (S. Etlinger & W. Amand, 2012), since it seemed to provide more accurate results than other sentiment analysis dictionaries.
- **Subjectivity.** We used the MPQA Subjectivity Lexicon (Wilson et al., 2005) to identify the subjectivity or objectivity of tweets.
- **Grammatical features.** We used the NLTK part-of-speech tagger (Bird, Klein, & Loper, 2009) to assign a single best-fitting part of speech (POS) to every token. We calculated POS diversity by finding the number of occurrences of each POS tag.

3.17.2.2 Twitter-Specific Stylistic Features

Twitter-specific features refer to characteristics unique to the Twitter platform that are associated with user accounts and the tweets sent from them, such as the number of followers, number of people followed, and the number of tweets the user has posted in the past. Twitter-specific features also include the presence or lack of URLs, mentions of other users, hashtags, and official account verification. These features were acquired using the Twitter API, and we treated them as part of the structure of the tweets and thus necessary for our analysis. Therefore, before preprocessing the data, we first calculated the number of occurrences of each of these features in a tweet and added them to the set of attributes.

3.17.2.3 Argumentation Features

We used the dataset provided by (Addawood & Bashir, 2016), which is labeled with argumentation and source type. We used these two labels as part of our feature set.

- **Argumentativeness.** We used a simple argument model that an argument is comprised of only two components: a claim and associated supporting evidence. If the tweet presented an argument or shared an opinion about the debate, it was marked as argumentative, and otherwise, as not argumentative.
- **Source type.** Source type refers to the type of evidence a user has given to support a particular position in a given debate. Six types of evidence were identified: ‘news media accounts’; ‘blog post’; ‘picture’; ‘expert opinion’; ‘other types of evidence’, and ‘no evidence,’ which referred to not having presented any evidence.

3.18 Imbalanced Class Distributions

It was not possible to control the class distribution by controlling the Twitter query, because determining the topic class and the stance class had to be manually determined as described in Section 3.3. However, the imbalances shown in Tables 9 and 10 above could bias the classifier, i.e. the classes with fewer instances could be predicted incorrectly and with lower accuracy than classes with more instances. Previous studies have proposed various balancing strategies, including oversampling, undersampling, cost-sensitive learning, and a combination of these methods (Nitesh V Chawla, 2005 ; Kotsiantis, Kanellopoulos, & Pintelas, 2006). Previous work has shown that the combination of oversampling and undersampling techniques performs better than plain undersampling (Nitesh V. Chawla et al., 2002; Rezapour & Diesner, 2017) and has a better outcome than cost-sensitive learning (Nitesh V Chawla, Japkowicz, & Kotcz, 2004). Therefore, to resolve imbalanced class distributions, we used a combination of two techniques: oversampling for classes with a small number of instances and undersampling for classes with a large number of instances. For oversampling, we used the Synthetic Minority Oversampling Technique (SMOTE) (Nitesh V. Chawla et al., 2002). SMOTE is one of the most accepted approaches for solving the problem of imbalanced data, and has better performance compared to oversampling with replacement (Nitesh V. Chawla et al., 2002). Its main function is to create new minority class examples by interpolating several minority class instances that occur together. In this method, new instances are synthetically created using k-nearest neighborhoods. Based on the number of cases in each class, a range from 500% to 900% was chosen using k=5 to minimize the risk of overfitting the classifier. After that we used random undersampling to reduce the size of large classes with a ratio of 5:1. Finally, we randomized the data to reduce the likelihood of overfitting the training data. Table 12 shows the new class distributions after balancing the dataset. The size differences between classes have been minimized.

Table 12: Number of instances for each class after balancing

	Class	**Class distribution after balancing**
Topic	Individual privacy	329
	National security	150
	Other	750
Stance	Favor	345
	Against	80
	Neutral	480

3.19 Experimental Results

The primary aim of our study was to determine the stance of tweets towards a certain topic. We used a multi-classification task to classify each tweet as having a stance in 'favor,' 'against,' or 'neutral'. As a first step, we compared classifiers that have frequently been used in related work: Naïve Bayes (NB) as used in (Teufel & Moens, 2002); Support Vector Machines (SVM) as used in (Liakata et al., 2012); and Decision Trees (DT, J48) as used in (Castillo et al., 2011). For all approaches, we used WEKA data mining software (Hall et al., 2009). Before training, all features were ranked by their information gain (Roobaert, Karakoulas, & Chawla, 2006). Information gain is presented in Equation 4.

$$InfoGain(Class, Attribute) = P(Class) - P(Class|Attribute) \quad (4)$$

Features with information gain of less than 0 were excluded. All results were subjected to 10-fold cross validation. For assessing prediction accuracy, we used the standard metrics of precision, recall, and F-measure. The results for each feature set and classifier are listed in Tables 13 and 14.

Table 13: Topic classification results of three classifiers using 10-fold cross validation

Feature set	DT			SVM			NB		
	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
Unigram (Baseline)	90.3	90.3	90.2	88.3	88.4	88.3	84.6	83.4	83.8
All features	93.7	93.7	93.7	93.2	93.2	93.2	85.9	84.5	84.9

Table 14: Stance classification results of three classifiers using 10-fold cross validation

Feature set		DT			SVM			NB		
		Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
Lexical	Unigram (Baseline)	76.3	76.1	76.2	81	81	81	79.1	78.7	78.8
	Unigram + Bigram	76.5	76.6	76.5	81.7	81.7	81.6	78.8	78.2	78.4
Lexical + Syntactic		75.7	75.7	75.7	82.9	82.9	82.8	81.4	81.0	81.1
Lexical + Argumentation		77.8	77.8	77.8	90.4	90.4	90.4	83.4	82.8	82.9
All features		77.6	77.6	77.6	83.8	83.8	83.8	79.4	79.3	79.2

3.19.1 Classification

Our first goal was to classify the topics related to encryption i.e. national security and individual privacy. Table 13 shows a summary of the classification results. The best results were achieved by using DT, which resulted in an F1 score of 93.7%. Our second goal was to classify tweets based on their stance toward a predefined topic. Adding a bigram feature to the baseline did not increase performance; however, adding argumentation features to the combination increased the performance by 10% for SVM. Table 14 shows a summary of our classification results. To achieve them, we created a baseline model by using the top salient unigrams. A baseline needed to be established so that we could assess the influence of added features on the models. The best overall performance was achieved by using SVM, which resulted in a 90.4% F1 score with lexical and argumentation-mining features added to the baseline. Moreover, combining all the features slightly decreased SVM and NB performance but did not change DT results substantively.

3.19.2 Feature Analysis

To identify and rank the most informative attributes of each feature, we calculated information gain (Eq. 4). The top 10 features with the largest weight (magnitude) with respect to each class are listed in Table 15. As shown in the table, the most informative feature in all classes was argumentativeness. Among Twitter-specific features, retweets appeared in both ‘favor’ and ‘neutral’ stances as well as in the ‘individual privacy’ and ‘other’ topics. Among syntactic features, Crimson Hexagon sentiment features were informative for the ‘favor’ stance as well as for the ‘individual privacy’ and ‘other’ topics. Moreover, among syntactic features we found that the most informative grammatical features for the topic of national security were preposition, adjective, verb, and adverb.

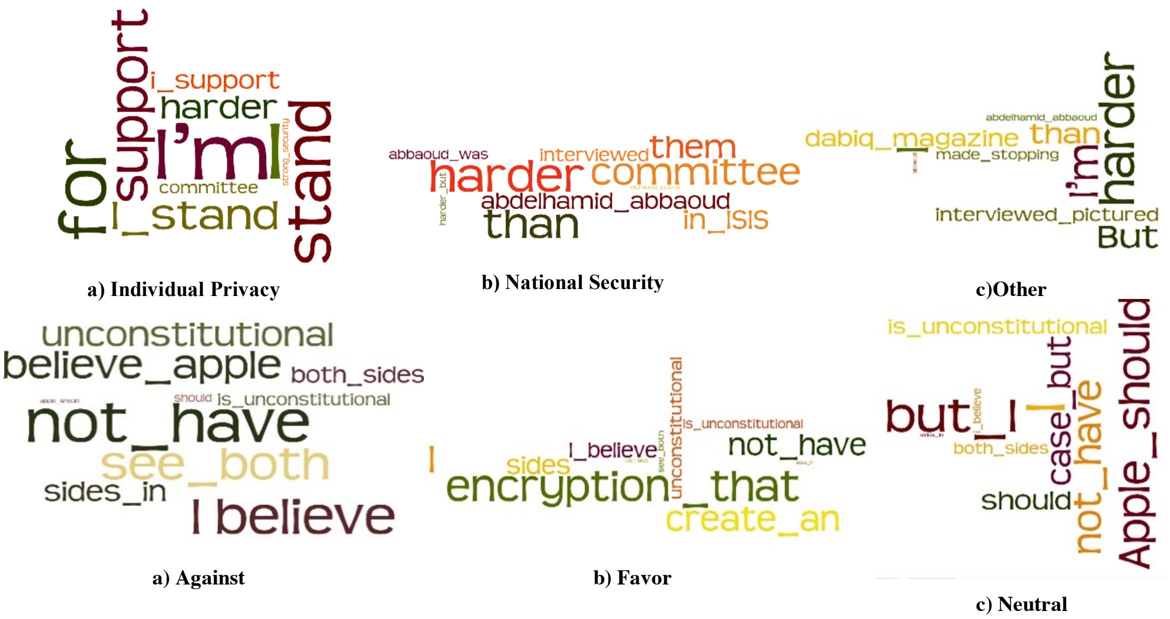
Table 15: Most and least informative features (non-lexical features in italic)

	Class	Most informative Features	Least informative Features
Topic	Individual privacy	argumentativeness, retweet, I, sentiment, I’m, stand, support, I stand, harder, I support	encryption from, encryption fight, encryption engineers, encryption security, encryption for, encryption debate, encryption I, encryption so, encryption technology, encryption support
	National security	preposition, adjective, verb, adverb, harder, than, committee, them, Argumentativeness, in ISIS	requiring encryption, really hope, powerful encryption, protect us, protest against, phone mass, privacy apples, one don’t, one consider, outta luck
	Other	argumentativeness, retweet, sentiment, verb, preposition, adjective, harder, I, adverb, I’m	internet commerce, internet like, i trust, iphone might, iphone encryption, layer encryption, law enforcement, key encryption, keys secure, keep getting
Stance	Favor	argumentativeness, retweet, sentiment, encryption that, I, create an, not have, sides, I believe, unconstitutional	sense privacy, so called, shocked that, should get, should too, sick of, side encryption, side of, side w, cloud storage

Table 15 (cont).

	Class	Most informative Features	Least informative Features
	Against	not have, noun, I believe, see both, believe apple, unconstitutional, sides in, both sides, is unconstitutional, create an	standup privacymatters, spying on, stand behind, stance how, stance in, standing up, stand by, stand up, stand with, cloud storage
	Neutral	argumentativeness, retweet, I, verb, but I, Apple should, not have, case but, should, is unconstitutional	retweet to, secure because, right don't, right to, secure at, right on, san Bernardino, rock solid, same encryption, cloud storage

Figure 7 shows the top 10 lexical features for each class. Among these features, we found that all classes had a combination of unigram and bigram features. The unigram '*I*' was one of the top informative features in all classes except the topic of national security. From the table, we can see that features as '*privacymatters'*, '*spying on*' and words related to standing up for encryption are negatively associated with the '*against*' stance. Moreover, sentiment bearing words, i.e. '*should*' are a good indicator of 'neutral' stance where it is negatively associated with the 'favor' stance. For topic classification, we can see from the table that the word 'encryption' is negatively correlated with the topic of individual privacy. Based on these findings, we conclude that using both lexical and argumentation features was beneficial for this task. As our analysis of the top informative attributes shows, the structure of sentences, grammatical indices, subjective words, and argumentativeness of tweets were useful for predicting the stance and topic.



a) Individual Privacy   b) National Security   c)Other

a) Against   b) Favor   c) Neutral

Figure 7: Word cloud of the most informative lexical features for each class. *Topic classes (top row) and stance classes (bottom row).*

3.20 Error Analysis

In addition to analyzing contributions within and among features' classes, we also studied each classifier's confusion matrix to find patterns in misclassifications. For stance classification, we chose the SVM's confusion matrix because of its comparatively higher accuracy with lexical and argumentation feature sets. Table 17 shows the number of classified instances per stance class, rendered in percentages. As the table shows, 'favor' and 'neutral' classes were the most misclassified classes. This result was consistent with our human annotators' feedback. They found it difficult to distinguish between tweets that favored a topic and tweets that did not take a stance toward the debate.

Table 16: Topic classification confusion matrix of DT classifier (by percentage)

	Individual privacy (%)	National security (%)	Other (%)
Individual privacy	87.5	0.60	11.85
National security	1.33	94.6	4
Other	3.2	0.53	96.26

Table 17: Stance classification confusion matrix of SVM classifier (by percentage)

	Against (%)	Favor (%)	Neutral (%)
Against	95	2.5	2.5
Favor	0	85.8	14.2
Neutral	0	7.1	92.91

In particular, it was challenging to distinguish between those who had a clear opinion about the topic versus those who were just making a joke about it. For topic classification, we chose the Decision Tree's confusion matrix. Table 16 shows the classified instances per topic class, rendered in percentages. As the table shows, the 'individual privacy' and 'other' topic classes were the most misclassified. To further analyze these prediction errors, we randomly selected 30 tweets from different classes, removed the labels, and asked the same two human annotators to label them again. Their unweighted Cohen's Kappa scores were 81.37% for stance classification and 70.4% for topic classification. This finding shows that some tweets were hard to categorize and suggests that understanding the intended meaning of the tweets might be needed to solve this problem. Based on our discussion with the human annotators, we believe that being able to see the whole conversation preceding the tweet and being familiar with the content of the shared URLs could lower these errors.

3.21 Discussion

In this study, we developed a theoretically grounded and data-driven classification schema, related codebook, corpus annotation, and prediction model for detecting stance in tweets from the "encryption debate." Our data annotation and analysis procedure showed that most individuals use Twitter as a venue for spreading information and links to webpages rather than as a platform through which to take clear positions about controversial issues. In Table 9, the distribution of topic label results shows that 'individual privacy' had a higher number of tweets compared to 'national security.' We think this bias toward 'individual privacy' may have happened because people are more confident tweeting about their personal right to privacy rather than the more public responsibility to maintain national security. It may also indicate that people are more willing to share their opinions if they thought that their audience agreed with them. These results can be compared to a recent Pew research study (Hampton et al., 2014) about

Edward Snowden's 2013 revelations of widespread government surveillance of Americans' phone and email records. The survey showed that 86% of Americans were willing to have an in-person conversation about the surveillance program, but just 42% of Facebook and Twitter users were willing to post about it on those platforms. The distribution of stance label results in Table 10 confirm previous research which showed that users do not frequently use Twitter as a debating platform (Smith et al., 2013). The results also illustrate that very few tweets were classified as being 'against' one of the topics. This may indicate that Twitter users do not take 'against' stances as frequently as stances in 'favor' of controversial topics, especially if those topics are morally and not scientifically based.

To build classifiers, we worked with four sets of features: lexical, syntactic, Twitter-specific, and argumentation. We trained three commonly used types of classifiers: Support Vector Machine, Decision Tree, and Naïve Bayes. We built a baseline model using top unigrams, gradually added other feature types, and measured the incremental contribution of each type. For topic classification, we only compared the baseline to the combination of all features because of the need to limit our research scope for this paper. The classification results (Table 6) showed that the combination of all four sets of features was most beneficial for the DT classifier, with which the results improved from 90.2% for the baseline to 93.7%. The Naïve Bayes scores for F1, recall, and precision were lower than those for the other two classifiers. Our results indicate a 20% improvement in F-measure score compared with previous research (Qiu et al., 2013; Rajadesingan & Liu, 2014; Somasundaran & Wiebe, 2010). We believe that the unique combination of features used in the classification as Twitter-specific features, sentiment, and argumentation facilitated these improvements. The stance classification results (Table 14) show that the SVM classifier outperformed the other two training algorithms and achieved the best overall performance. It did so by using a combination of lexical and argumentation features, which led to a performance that improved from the 81% baseline F1 score to the 90.4% final model F1 score.

The comparison of the top attributes of each class revealed that one argumentation feature, which indicated whether or not a tweet is argumentative, is the best indicator for stance classification. This may indicate that when a tweet is argumentative it denotes that the user expresses a stance toward the topic. Moreover, the retweeting behavior was observed in tweets that have an in 'favor' or 'neutral' stance only. Also, the same retweeting behavior was observed in tweets discussing 'individual privacy' and 'other' topics. This may indicate that users on

Twitter are more comfortable sharing information in ‘favor’ or ‘neutral’ towards ‘individual privacy’ and ‘other’ topics but not toward ‘national security.’ One limitation of our study is that we understand that our dataset may not be representative of the overall opinions of Twitter users online. As our sample shows, only 1% of the annotated dataset was about national security while few tweets had an ‘against’ stance. However, we believe that these results still provide some information about Twitter users’ attitudes towards the encryption debate. We found that some lexical features are very indicative of the topic class. In the case of the individual privacy topic, for example, the top lexical features were ‘for,’ ‘stand,’ ‘support,’ and ‘I stand.’ These features indicate a very strong position toward the topic, in contrast to the national security topic where the first personal pronoun did not appear as one of the top features. This result may indicate that users are less comfortable expressing their own opinions when the topic involves national security or that they are more comfortable discussing a personal matter such as their privacy rather than a collective issue as the national security of the whole country. We also conducted an error analysis of misclassified instances, finding that tweets related to ‘favor’ and ‘neutral’ stances as well as ‘individual privacy’ and ‘other’ topics were the most challenging to classify. This occurred because of the challenge of classifying short texts that do not follow any guidelines or rules for the expression of opinions and the challenge of distinguishing sarcasm from earnest opinions.

3.22 Conclusion and Future Work

The analysis of social media content has been studied extensively. There are many challenges to opinion-mining social media content, because online users’ expressions are written informally, and so may include sarcasm, spelling mistakes, unconventional grammar, and slang words and expressions (Reyes et al., 2012; Riloff et al., 2013). Several works have begun to develop tools and computational models for tweet-level opinion and sentiment analysis. Although opinion mining and sentiment analysis can identify whether a user expresses a positive or negative emotion regarding a topic, these techniques may not capture a user’s stance (in favor or against a given position) on the topic. Stance classification has been introduced to address this gap. Although as yet under-investigated, stance classification has seen growing interest in recent years, as this technique can be advantageous, particularly in support of decision-making. In order to detect online users’ attitudes and stances on a given issue, we used Twitter data related to the recent Apple and FBI encryption debate. In this study, we presented the task of automatically classifying stance on social media for users discussing controversial topics like the recent FBI and Apple encryption debate utilizing unique feature sets. We classified two predefined topics related to the debate and built a dataset of 3,000 manually annotated tweets related to these topics. Our subsequent analysis, motivated by the research presented in (Addawood & Bashir, 2016), found that SVM classifiers trained with lexical and argumentative features were best at capturing stances taken toward different topics expressed on social media. While previous work has considered classifying stance without any tweet context, we show that using various features such as the sentiment and the argumentativeness of the tweet support the identification of the stance of the tweet and can lead to significant improvements in stance classification.

As stated previously, working with social media data has some challenges and limitations. Annotating tweets related to a controversial topic such as the encryption debate requires annotators who not only understand the English language used and its informing cultures, but who also understand the encryption debate as a whole. Another challenge of annotating the data was related to the language and structure of tweets, in which users tend to use informal and incoherent text. In addition, it is important to note that although our classification achieved a high score in our selected debate topic, these results may not be generalizable to other domains without further investigation. Understanding public opinions and attitudes towards controversial topics may help scholars, law enforcement officials, and policy-makers develop better policies and guidelines. People's attitudes and behaviors related to privacy are highly contextualized in the digital age. While many scholars have conceptualized information privacy in various disciplines, investigations of individual users' attitudes and behaviors towards information privacy and national security remain limited. The dataset developed in this study will be used in future research to develop a better understanding of users' attitudes towards the encryption debate: ultimately that may help enhance current privacy policies and guidelines. Given the growing significance of the role social media is playing in our world, studying stance classification can be beneficial for instance, in identifying electoral issues and understanding how public stance is shaped (Mohammad et al., 2015). One implication of our research is that it suggests that it is possible to understand who frequently participates in controversial discussions on social media. Moreover, correlating users' stance with their sentiments and demographics may help further describe users' behavior online. Also, predicting a user's stance toward a given issue can support the identification of social or political groups, help develop better recommendation systems, or tailor users' information preferences to their ideologies and beliefs. Additionally, it may provide engineers and designers with new ways of improving the design and users' acceptability of current privacy-enhancing technologies.

In future work, we hope to improve our results with more intelligent features for representing context, discourse, rhetorical structure, and dialogic structure, such as capturing irony and sarcasm. Another area to explore in future work is analyzing tweets based on the whole conversation, instead of just a single tweet, to get a better understanding of users' different opinions. Another line of research to pursue in the future is to develop a system that can detect the different stances users have regarding a controversial topic, i.e. explore how people decide what the sides (two, three, more) are in a given debate. A controversial topic may generate many different and nuanced stances, even on the same general side of a debate.

CHAPTER 4: HEALTH DEBATE ON TWITTER

4.1 Introduction

This chapter investigates information sharing behavior on Twitter for health discussions. Tweets related to the controversy over the supposed linkage between the MMR vaccine and autism were collected for analysis. The first part of this chapter concerns the analysis of scientific information sharing behaviors on Twitter. The usage pattern of scientific information resources by both sides of the ongoing debate were examined. Then, how each side uses scientific evidence in the vaccine debate was explored. To achieve this goal, the usage of scientific and non- scientific URLs by both polarized opinions was analyzed. A domain network, which connects domains shared by the same user, was generated based on the URLs "tweeted" by users engaging in the debate in order to understand the nature of different domains and how they relate to each other. Most of the contents of this part is published in (Addawood, 2018). The second part investigates the usage of the different types of information sources in social media and who shares them. In this part a classification schema for influential users who tweet about the Measles-Mumps-Rubella (MMR) vaccine is develop. Moreover, the key information sources cited by these users in the form of URLs is examined. This work is a collaboration with Umberto Ravaioli, Nahil Sobh, Peg Burnette, Amanda Avery and Anuja Majmundar.

4.2 Scientific Credibility Behind MMR Vaccination Debates

Social media has revolutionized how people disclose personal health concerns and discuss public health issues. Social media provide unique platforms without time and location constraints for sharing health-related information (Dunlap & Lowenthal, 2009; Lachlan, Spence, & Lin, 2014). Social media have been found to be important tools for facilitating discussions on health information, especially in health crisis situations (Simon, Goldberg, Aharonson-Daniel, Leykin, & Adini, 2014; Sullivan et al., 2012) in which users share insights, opinions, and apprehensions while disseminating interpretations of health events outside of a public health context (Khan, Fleischauer, Casani, & Groseclose, 2010; Sullivan et al., 2012).

When online users discuss topics on Twitter, they often include evidence to support their claims, including links to online sources, such as newspapers or blogs (Addawood & Bashir, 2016). However, these sources may include unverified or even false information, which may amplify the perceived risks of these health issues (Fung, Tse, Cheung, Miu, & Fu, 2014; Kasperson et al., 1988). From an audience perspective, online health information offers a quick and useful reference, but its accuracy and credibility often falls into question (Metzger & Flanagin, 2011).

Users of social media generally regard scientific sources, such as journal articles, to be credible. In this study, scientific sources are defined as sources that link to scholarly articles. Public opinion surveys from Europe and the US show that scientific institutions are trusted and are generally considered to be more credible than non-scientific sources (Bucchi & Trench, 2014). However, it is not clear from the previous literature how scientific evidence is deployed in discussions among Twitter users regarding health information. The problem of scientific research use in online, socially mediated discussions on health information is complicated by the controversies that surround certain health issues. These controversies can arise even when there is little to no credible evidence to support them. One significant controversy is the supposed relationship between the Measles, Mumps, and Rubella (MMR) vaccinations and autism.

During the 2014 holiday season, an outbreak of measles originated at the Disney theme parks in California. The outbreak generated extensive public discussion on some parents' resistance to childhood vaccinations. One reason for this outbreak is that, for some parents, concerns about the potential side effects of vaccines have overtaken concerns about the dangers of potentially deadly, vaccine-preventable diseases. "The Antivaccine Movement," a social movement composed of "antivaccine groups" and "antivaccine activists," is designated by scientists as the main cause of vaccine hesitancy or refusal (Ackermann, Chapman, & Leask, 2004; C Betsch, 2011; Cornelia Betsch et al., 2012; Gangarosa et al., 1998; Poland & Jacobson, 2011).

Another challenge regarding health information controversies that presents itself in social media discourse is selective exposure to online information (Frey, 1986). This phenomenon, which is due personalized web algorithms, happens when users find information that primarily supports their preconceptions and shields them from exposure to different ideas. Instances of selective exposure, including when like-minded people share their views with one another to reinforce their pre-conceived biases, are known as "echo chambers" (Sunstein, 2009). Social media users' attitudes were confirmed through observations regarding the use of hashtags related to the vaccine controversy. One of the main uses of hashtags is to highlight users' sentiments towards the topic under discussion (X. Wang, Wei, Liu, Zhou, & Zhang, 2011). In order to conduct the observation, specific instances in the ongoing MMR vaccine debate from January 1, 2016 to November 28, 2016 served as a case study. These instances demonstrated how narrative elements are extracted for public debates regarding the vaccine issue.

While previous research has examined how scholars use social media, mainly Twitter, to request and offer assistance to others (Veletsianos, 2012), to critique the work of other scholars (Mandavilli, 2011), to contribute to conferences via hashtags (J. Li & Greenhow, 2015; Ross, Terras, Warwick, & Welsh, 2011), to implement engaging pedagogies (Junco, Heiberger, & Loken, 2011), and to share and comment upon preprint and published articles (Eysenbach, 2011), no previous study found that scientific publications are referenced by online users to support their claims regarding vaccines. Moreover, previous research that has examined the credibility of information shared via social media has not considered the use of scientific evidence by users of social media. Several studies have shown that the incorporation of URLs into social media posts is a means by which users attempt to confer credibility (Castillo et al., 2011; Kinsella, Wang, Breslin, & Hayes, 2011). However, no studies have been found that investigate the content or types of these URLs.

The goal of this study is to analyze scientific information sharing behaviors on Twitter regarding the controversy over the supposed linkage between MMR vaccine and autism. We examine the usage pattern of scientific information resources by both sides of the ongoing debate. Then, we explore how each side uses scientific evidence in the vaccine debate. To achieve this goal, we analyzed the usage of scientific and non-scientific URLs by both side of the debate. A domain network, which connects domains shared by the same user, was generated based on the URLs "tweeted" by other users in order to understand the nature of different domains and how they relate to each other. This study has the potential to improve understanding about the ways in which health information is disseminated via social media.

4.3 Social Media for Public Health

Studying the patterns and mechanisms of health-related communication via social media has the potential to give valuable insights into how health information shapes users' beliefs and attitudes. (Salathé & Khandelwal, 2011) studied Twitter content to assess the levels of polarization between supporters and opponents of swine flu (H1N1) vaccination in the broader context of digital epidemiology (Salathe et al., 2012). They explored users' sentiments toward information shared via social media and users' following patterns. Their results show that people tend to follow other users who share the same sentiments about a topic. (Radzikowski et al., 2016) analyzed Twitter narratives regarding MMR vaccination to identify key terms, connections among such terms, retweet patterns, the structure of the narrative, and connections to geographical space.

Social media data has also been used in outbreak detection. For example, (Odlum & Yoon, 2015) studied the use of social media in the 2014 Ebola outbreak. They used a set of 42,236 tweets mentioning the word Ebola to assess the potential benefits of using social media as a real-time outbreak-tracking tool. Similarly, (Lampos & Cristianini, 2010) and (Culotta, 2010) correlated tweets mentioning influenza and related symptoms with historical data. Their results showed high correlations between Twitter statistics and Centers for Disease Control and Prevention (CDC) statistics in cases of influenza.

A closely related study was done by (Love, Himelboim, Holton, & Stewart, 2013), who conducted a content analysis of 2,580 reposted/shared vaccination Twitter posts to determine what vaccination information people share promote. Other researchers focused on one side of the debate: the anti-vaccine movement (Kata, 2012; Tomeny, Vargo, & El-Toukhy, 2017). This study's goal is to examine the use of scientific information sources from both sides of the debate with in-depth analysis of 6,112 tweets.

4.4 Information Credibility in Social Media

Some studies on information credibility on Twitter focus on identifying sets of features that are indicative of credibility (Castillo et al., 2011; Gupta & Kumaraguru, 2012; Mendoza, Poblete, & Castillo, 2010). One of these features is the presence of links in the tweet text. Castillo et al.(2011) use a complex set of features over tweets, re-tweets, the text of the posts, references to external sources, and users to predict the credibility of an event. Their results showed that having a URL tends to indicate that a tweet is credible. This was also confirmed by other studies (Morris, Counts, Roseway, Hoff, & Schwarz, 2012; Sikdar, Kang, O'Donovan, Hollerer, & Adal, 2013). In this study, we extend this research to investigating the use of a specific type of evidence, scientific evidence, in Twitter discussions about a controversial health issue.

4.5 Data Collection

We collected a corpus that contained ground-truth, or gold standard, data, i.e., tweets that contain scientific versus non-scientific evidence on the topic of vaccines. Our corpus contained two main datasets. One dataset contained tweets that discussed the topic of MMR vaccines and their relation to autism, providing different types of supplementary evidence. The second dataset contained tweets that talked about vaccines and provided supplementary scientific evidence in the form of URLs linked to a scientific paper about vaccination. These two datasets are referred to as non-scientific and scientific, respectively.

For the non-scientific dataset, we collected data using Crimson Hexagon (S. Etlinger & W. Amand, 2012), a public social media analytics platform. We collected a sample of public posts made from January 1, 2016 to November 28, 2016. The sample only included tweets from accounts that set English as their language. The search criteria were: ("vaccinations" OR "vaccination" OR "vaccines" OR "vaccine" OR "measles-mumps-rubella" OR "MMR" OR "mmr" OR "#MMR") AND ("autism" OR "autistic disorder") AND NOT "RT:"

The total number of tweets retrieved was 45,320. To have more concise results, we removed all duplicate tweets (e.g. tweets repeated more than once in the dataset), which we believe is going to affect our final results. The total number of remaining tweets was 28,848. To collect more features related to each tweet, we ran these tweets' IDs through the Twitter API. Even though this process gave us more meta-data for each tweet, it reduced the usable number of tweets in the dataset to 27,816, since some tweets were deleted by the users or not found.

For the scientific dataset, we used PubMed to collect research articles related to vaccination and autism. PubMed is a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. MEDLINE journals are selected by a technical advisory committee run by the U.S. National Institutes of Health (NLM, 2013). The search query used in PubMed was: ("Vaccinations" OR "vaccination" OR "vaccines" OR "vaccine" OR "measles-mumps-rubella" OR "MMR" OR "mmr") AND ("autism" OR "autistic disorder"). This method resulted in a collection of 794 research papers.

For the next step, we chose Altmetric.com as the data source for social media and mainstream media counts, as it is the most comprehensive source of social media data associated with scientific papers (Robinson-García, Torres-Salinas, Zahedi, & Costas, 2014). Altmetric.com links an identifier for each article that is provided by PubMed (i.e., its PMID). The Altmetric.com

API then returns the social media reaction to a specific article that has been associated with a given PMID. Not all PMID queries produced results. For the 794 articles in our collection, Altmetric.com returned 346 results, i.e., the number of papers that had been referenced on social media. Since we chose Twitter as the social media platform for this study, we needed to further narrow these results to those that were Twitter-specific. Altmetric.com only provides the ID of a tweet, so once we eliminated non-Twitter results, we used the Twitter API to capture the actual tweet text by matching tweet IDs. The Twitter API returned 25,751 tweets. However, the Twitter API returned tweets in all languages, so languages other than English also had to be removed. To make our dataset consistent, all tweets that were retweets were removed. The final dataset contained 8,612 tweets, which we will refer to as the scientific dataset. This number is very small compared to the nonscientific dataset; this was expected since few online users refer to scientific sources.

We acknowledge the possibility of having scientific evidence in the 27,816 non-scientific dataset and vice versa. However, to mitigate this we crosschecked the scientific and non-scientific datasets to make sure no tweet appeared in both of them. We found 94 tweets that appeared in both datasets and removed them. The final combined datasets contained 36,428 tweets.

4.6 Data Annotation

The next step was to annotate tweets for their stance towards vaccines. To accomplish this task, we utilized the hashtags present in each tweet. We followed previous work on hashtags use as indicators of users' common interests and opinions toward a health issue (Fang, Ounis, Habel, Macdonald, & Limsopatham, 2015; Xu, Chiu, Chen, & Mukherjee, 2015). As a first step, we identified all hashtags in the dataset; only 35% of the tweets contained hashtags (13,089 tweets). After that, two annotators hand-labeled all hashtags as either having a pro- or an anti-vaccine opinion. The inter coder reliability is 91.3%, with a 90.1% Cohen's Kappa. In total, there was 45 pro-vaccine hashtags and 94 anti-vaccine hashtags. The top hashtags identified in each category are shown in Table 18. To validate the selection of these hashtags a sample of 40 random tweets were chosen, two different annotators annotated the tweets for being either having a pro or an anti-vaccine opinion. Both annotators agreed on 39 cases out of 40 which also matched the hashtag opinion annotation.

Table 18: Distribution for anti and pro-vaccine attitudes hashtags

Anti-Vaccine	Count	Pro-Vaccine	Count
Vaxxed	3408	**vaccineswork**	264
Cdcwhistleblower	3132	**vaccinesNOVA**	98
sb277	322	**vaxwithme**	86
Cdcfraud	195	**vfvcall**	61
b1less	182	**whyivax**	26
Vaccineinjury	119	**vaccinessavelives**	17
Bigpharma	100	**antivaxxers**	16
Coverup	45	**vaccinateyourkids**	13
Learntherisk	33	**antivaxx**	10
vaccineskill	30	**teamvax**	9

4.7 Hashtag Selection Discussion

To accurately identify a hashtag as either presenting an anti- or pro-vaccine attitude, we investigated the hashtag's usage on Twitter. For some tweets, it was easier to identify the opinion of the poster. For example, #vaccineswork clearly implies that the person believes in vaccines and their effectiveness. Similarly, #killingusslowly noticeably identifies that the person believes that vaccines can result in death. However, some hashtags are harder to identify since they require some understanding of the subject matter. For example, the hashtag #sb277 refers to the California Senate Bill 277, which is a law that removes personal belief exemptions to vaccination requirements for entry to schools in California, a state with relatively low vaccination levels in some schools (Aftab, 2015). This hashtag is used mostly by users who are against this bill and hold anti-vaccine attitudes.

Another popular hashtag use is to reference people who hold the opposite opinion. For example, the hashtag #antivaxxers is used to refer to people who hold anti-vaccine attitudes.

Similarly, #provaxxer is used to refer to people who hold pro-vaccine beliefs. Our assumption was people who hold anti-vaccine beliefs do not identify themselves with the #antivaxxers hashtag, and the same applies for people with pro-vaccine attitudes. This assumption was validated after closely reading a sample of tweets that use these hashtags. This investigation confirmed that users use these hashtags to refer to people who have the opposite belief than the one they have.

Some hashtags may be connected to the issue of vaccines but only provide information with no clear opinion on the issue, as in the hashtag #vaxxfacts. This hashtag clearly presents facts about vaccines without taking a side, but even it can be used by either side to claim that they are presenting facts. Moreover, some of the hashtags in the dataset did not demonstrate any clear opinion toward the issue. For example, the most used hashtag was #autism, which had 5447 occurrences. However, this hashtag does not show an opinion toward the issue. Other hashtags were out of the scope of the issue, such as #jewish, #dating and #sports.

Table 19: The distribution of vaccine attitudes and the usage of scientific references

	Pro-vaccine	**Anti-vaccine**
Scientific	139 (2.3%)	945 (15.5%)
Non-Scientific	291 (4.8%)	4,737 (77.5%)

Table 20: The distribution of vaccine attitudes and the inclusion of URLs

	Pro-vaccine	**Anti-vaccine**
Contains a URL (one or more)	335 (5.5%)	4,782 (78.2%)
Does not contain a URL	95 (1.6%)	900 4.7%)

4.8 Findings of Vaccine Attitudes

After identifying hashtags representing online users' vaccine attitudes on Twitter, we investigated the distribution of these attitudes in the data. To do so, we applied the selected hashtags to the data to identify tweets that signify pro- or anti- vaccine attitudes. We identified 6,112 tweets as having an opinion: 430 tweets with a pro-vaccine opinion and 5,682 tweets with an anti-vaccine opinion. We also found 215 tweets containing both anti- and pro-vaccine hashtags, which were removed from further analysis. These results show that there is a much higher number of tweets discussing anti-vaccination than pro-vaccination attitudes. This may indicate that people who hold anti-vaccine attitudes utilize Twitter as a venue for disseminating their opinions more than people who hold pro-vaccination beliefs. This may happen because social media is intensifying the reach and power of anti-vaccination messages, which may lead to negative reactions to vaccines being increasingly shared across online platforms (UNICEF, 2013).

4.9 Usage of Scientific References

Another goal was to identify the use of scientific references in the discussion of vaccines on social media. To accomplish this, the number of pro- and anti-vaccine tweets with references to scientific and nonscientific evidence was recorded. Table 19 shows the distribution of vaccine attitudes and the use of scientific references. The ratio of pro-vaccine tweets containing links to non-scientific evidence compared to scientific evidence is 1:2.09, while the ratio of anti-vaccine tweets that contain a links to non-scientific evidence compared to scientific evidence is 1:5.01. These results show that people with both attitudes reference more non-scientific evidence compared to scientific evidence.

4.10 Scientific Reference Analysis

To better understand the usage of scientific references in discussions of vaccines via social media, we did a thorough analysis of the URLs shared. Since many top domains are shortened URLs (e.g., bit.ly), we expanded them and extracted domain names. All tweets in our scientific dataset contained URLs. However, only 80% of the non-scientific dataset contained URLs. Table 20 shows the distribution of vaccine attitudes and the inclusion of URLs. The ratio of anti-vaccine tweets that contain URLs compared to tweets that do not is 5.31:1, while the ratio of pro-vaccine tweets with URLs compared with non-URLs is 3.53:1. This shows that users with anti-vaccine attitudes refer to external sources more often.

Online users can share more than one URL in their tweets. In our dataset, we found that users shared up to five URLs in their tweets. People with pro-vaccine attitudes shared up to two URLs, while people holding anti-vaccination views shared more. This result may indicate that people with anti-vaccine attitudes are trying to strengthen their arguments by sharing more links to external references.

At first, we investigated the top 15 URLs in pro-vaccine discussions. For users sharing non-scientific references, we found that these discussions mostly contained evidence showing that there is no link between the MMR vaccine and autism. This evidence mostly came from blogs or news websites such as npr.com. There were three references in the list that showed evidence of a positive linkage between the MMR vaccine and autism; however, these websites seemed untrustworthy based on a quick online search, such as “naturalnews.com” and “vaccines.news”. Social media references, such as Twitter and YouTube, also appeared.

For users sharing scientific references, the top 15 URLs shared in pro-vaccine discussions show that this group shares references containing scientific evidence against the supposed link between the MMR vaccine and autism. The most shared URL linked to a paper published in the Vaccine journal with the title, “Vaccines are not associated with autism: An evidence-based meta-analysis of case-control and cohort studies,” which confirmed that there is no link between MMR vaccine and autism.

People who are pro-vaccination frequently share scientific references citing articles published by Brian Deer,[2] a journalist. Deer did a series between 2004 and 2010 investigating the concerns over the MMR vaccine that arose after the 1998 publication of a research paper in the medical journal, The Lancet, written by Andrew Wakefield and his colleagues (Wakefield et al., 1998), which was later retracted because of invalid research results. Two of Deer’s articles appeared in our list: “How the case against the MMR vaccine was fixed” (Deer, 2011a) and “How the vaccine crisis was meant to make money” (Deer, 2011b). In these articles, Deer shows how the results of this research were fraudulent.

The top 15 URLs shared in anti-vaccine discussions all claimed to show a link between the MMR vaccine and autism. We found that the domain “truthinmedia.com” was referenced the most in the list. We found that the website was no longer live. This website/project belongs to Ben Swann,[3] a journalist. He is best known for his investigation of the linkage between MMR vaccines and autism. In his fact-checking series, he argued that there is a link between MMR vaccines and autism. That same domain appeared in our top shared URLs four times. In our list, we also noticed many references to Ben Swann’s Facebook page or tweets. All of these materials have been deleted. Other websites that appeared in the top shared URLs were “vaxxed.com” and “vaxxedthemovie.com,” which are a movement and a movie with a conspiracy theory orientation that investigate the CDC's supposed destruction of a study linking autism to the MMR vaccine. The movie was directed by Andrew Wakefield, whose medical license was revoked after his paper was retracted.

[2] http://briandeer.com/

[3] https://en.wikipedia.org/wiki/Ben_Swann

We also found that these references linked to papers that claim to show a connection between MMR vaccines and autism. The most highly cited source was by Brian Hooker, which was titled “Measles-mumps-rubella vaccination timing and autism among young African American boys: a reanalysis of CDC data,” which provides seemingly strong evidence of the linkage between MMR vaccine and autism. This paper was published in Translational Neurodegeneration journal in 2014 but was later retracted. Another highly referenced paper in our list was “Hepatitis B vaccination of male neonates and autism diagnosis, NHIS 1997-2002,” (Gallagher & Goodman, 2010) which showed evidence that Hepatitis B vaccination causes autism.

There were two articles that were shared in both the anti and pro-vaccine conversations. The first article was published in the journal of Immunologic Research in 2013 with the title “Aluminum in the central nervous system (CNS): toxicity in humans and animals, vaccine adjuvants, and autoimmunity” (Shaw & Tomljenovic, 2013). This paper linked aluminum used in vaccines to autism; both pro- and anti-vaccine posters linked to this article. This finding may indicate that both groups of people have concerns regarding the content of vaccines and what they could do to young children, even though they have different attitudes toward the issue. The second article was published in the American Academy of Pediatrics in 2014 with the title “Safety of Vaccines Used for Routine Immunization of US Children: A Systematic Review” (Maglione et al., 2014). This paper had two main conclusions: “There is strong evidence that MMR vaccine is not associated with autism” and “We found evidence that some vaccines are associated with serious Adverse Events” (p. 334). While this paper confirms that there is strong evidence that MMR vaccines do not cause autism, it shows that in rare occasions rotavirus vaccines may be associated with intussusception, a different medical disorder. The citation of this paper by posters from both sides may indicate that people did not fully comprehend the results of the paper. All of these results show that on both sides there are influential people who people trust and reference as evidence for their beliefs.

When sharing non-scientific evidence, people mostly share links from social media websites such as Facebook, Twitter and YouTube. Sources such as news sites and personal blog posts are the next most commonly shared links. This may indicate that users on Twitter share other opinions manifested in tweets to support their own attitudes. Another observation was that the pro-vaccine community shared more diverse sources on Twitter than the anti-vaccine users. In the list of the top 15 URLs shared by the pro-vaccine community, all source domains were unique with no duplicates. The anti-vaccine top-shared URLs contained four pointers to the truthinmedia.com website and two references to Ben Swann, Truth in Media creator, including his social media webpages. Moreover, there were six different references to the movie Vaxxed: From Cover-Up to Catastroph", such as either a straight link to the website, a link to a website for a theatre for movie or a link to Periscope,[4] which shared a showing of the movie. This may indicate that the anti-vaccine community has few sources to support their opinions.

The same results occurred in the sharing of scientific evidence on both sides of the controversy. Users holding anti-vaccine attitudes referenced the retracted paper by Brian Hooker five times on the list of top 15 URLs. Also, they referenced the paper linking Hepatitis B to vaccines twice. This may indicate that users holding anti-vaccine attitudes have little solid evidence to support their opinions. Meanwhile, users with pro-vaccine attitudes shared meta-analysis (Taylor, Swerdfeger, & Eslick, 2014) and systemic review (Maglione et al., 2014) papers, which are based on the evidence-based medicine pyramid considered to be as the highest and most trusted type of evidence (Greenhalgh, 2014). These results indicate that it is mostly users holding anti-vaccine attitudes who rely on weak and redundant evidence compared with users holding pro-vaccine views.

4.11 Words Tell All: Unigram Analysis

Unigram analysis shows keywords for the overall narrative. Keywords reflect the topics that are considered relevant and important by the general public (Radzikowski et al., 2016). Given the design of the data collection process, all of the tweets in the data corpus for this analysis included one or more of the words in the search query used for the collection. All words that were used in the data collection were excluded from the analysis because their very high frequencies would make all other data points smaller. We also excluded stop words (i.e., articles, prepositions, and common verbs), as such words lack semantic significance.

[4] https://www.pscp.tv/

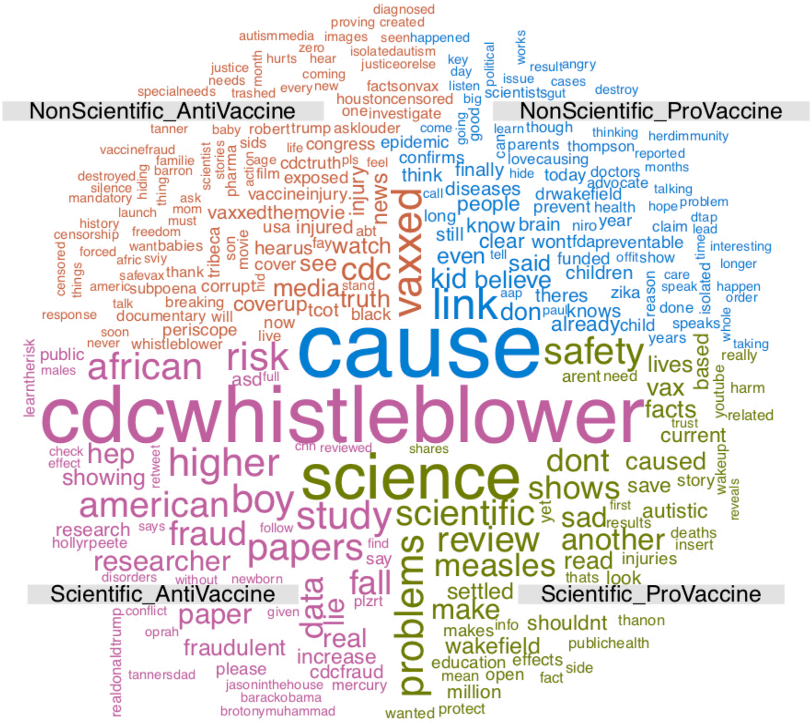


Figure 8: Comparative word cloud. Orange= Non Scientific_AntiVaccine; Blue= Non Scientific_ProVaccine; Pink= Scientific_AntiVccine; Green= Scientific_ProVccine.

In order to provide a general overview of the dominant narrative terms, Figure 8 shows a comparative word cloud visualization of the 100 most frequently encountered terms in each of the four data corpuses. The relative size of each word is proportional to its term frequency, where words in larger fonts are the ones more often encountered in the data corpus. Going from right to left, with the first being the non-scientific, pro-vaccine group’s top used words, we can see that the most common term encountered in these tweets is “cause.” Here is an example tweet from this dataset that shows how this word was used:

> *@username I've said it before and I'll say it again...VACCINES. DO. NOT. CAUSE. AUTISM. They do more good than harm. #vaccinateyourchildren*

This reflects the fact that pro-vaccine advocates' tweets demonstrate strong beliefs in a lack of linkage between vaccines and autism. The second most used word in this dataset is “link,” wherein users identify no link between vaccines and autism. The second word cloud set is for the dataset with anti-vaccine attitudes and the usage of non-scientific references.

We can see that the most encountered term is “vaxxed,” followed up with “cdc” and “vaxxedthemovie.” These words are related to the prevalence of references to the movie in tweets.

This result correlates with the previous results, which showed that users holding anti-vaccine attitudes share this movie widely using different outlets. The following tweet links to a podcast that discusses the importance of the movie and encourage others to watch it.

#Vaxxed shows the fraud of #CDC and #BigPharma by revealing true link between #MMR #vaccine and #autism <link5>

For the dataset discussing pro-vaccine attitudes and referencing scientific evidence, the most common terms encountered in these tweets are “science” and “scientific.”

The following tweets shows how these terms appeared in our dataset.

@username science: vaccines aren't linked to autism. <link6> <link7>#vaccinesNOVA #hearthiswell.

Reminder: An overwhelming body of scientific evidence shows that #VaccinesWork and don’t cause #autism <link8>

Finally, exploring the top word for anti-vaccine attitudes and the usage of scientific references, we can see the term “cdcwhistleblower,” which is a hashtag widely used by the antivaccination community in messages aligned with their views. This term did not originate from a formal organization, but instead is one that has emerged from an online advocacy community as a means to consolidate its views and promote its perspectives. Users using this hashtag claim that the CDC conceals evidence of the linkage between vaccine and autism, as shown in this tweet:

@username Dr. Hooker's abstract removed with bogus reasoning. #CDCwhistleblower 's truth censored! <link9>

In the same dataset of anti-vaccine attitudes and the usage of scientific references, we were puzzled by the word “fraud” in the top terms. When investigating the list of tweets including this term, we found that it was used to point to a study (DeStefano, Bhasin, Thompson, Yeargin-Allsopp, & Boyle, 2004) showing that the child’s age when getting vaccines does not affect getting autism. The following tweet shows how the term was used.

[5] https://t.co/aOroYavCkr

[6] http://www.sciencedirect.com/science/article/pii/S0264410X14006367

[7] https://academic.oup.com/cid/article/48/4/456/284219/Vaccines-and-Autism-A-Tale-of-Shifting-Hypotheses

[8] http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0003140

[9] https://translationalneurodegeneration.biomedcentral.com/articles/10.1186/2047-9158-3-16

@TIME No integrity. The lead story should be this CDC study was called FRAUD by one of its authors. <link10> #CDCwhistleblower

This tweet and some other similar ones claim that this study presented inaccurate results on the linkage between children ages when first administered with vaccines, and autism. Similar tweets claim the study was funded by the CDC to prove that there is no association between the required vaccines and autism. Taken together, these observations show how different opinion-holders express their attitudes toward the issue. These results show that users mostly discuss scientific studies to support their opinions. Others share these studies to reject them by accusing them of circulating fraudulent results.

4.12 Domain Network

To understand how different internet domains are used as supports in online users' discussions regarding the linkage between the MMR vaccine and autism, we created a network graph of domains connected through user activity, specifically the URLs shared in their tweets. Building on previous work on online rumors (Maddock et al., 2015; Starbird, 2017) to create the graphs, we first identified every distinct domain that is linked to by a tweet in the set. 83.72% of the tweets contain one or more URLs (5,117 tweets), and together they reference 495 distinct domains. These domains became the initial nodes for the graph. We created the edges between the nodes by observing the tweet patterns of each user, connecting two nodes if the same user posted tweets referencing both domains (Starbird, 2017).

Some domains were removed from the graphs for their high rates of connectivity to other sites as well as the different meanings encoded in those connections, i.e. they are used as tools, not for their content. These domains include social media services (e.g. Twitter.com, facebook.com, reddit.com) and all general link shortener services (e.g. bit.ly, t.co) which did not resolve to a URL. Finally, we trimmed the graph by removing domains that appeared fewer than five times in the set. In the graphs, nodes are sized proportionally to the total number of tweets that linked to the domain, and they are connected when an individual user wrote different tweets citing each domain. Furthermore, the strength of the edge grows proportionally to the number of users who shared tweets referencing both domains.

[10] https://www.ncbi.nlm.nih.gov/pubmed/14754936

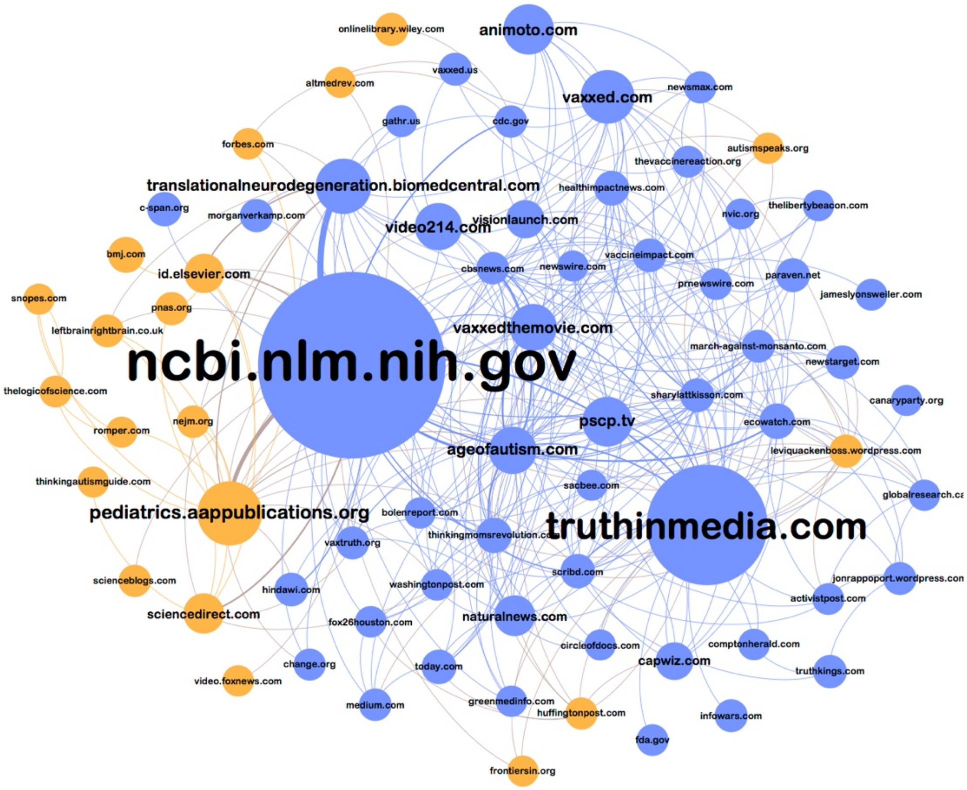


Figure 9: Domain Network Graph, colored by vaccine attitudes. Purple = Anti-vaccine; Orange = Pro-vaccine.

The resulting network graph represents how different domains are connected through the posting activity of Twitter users who contributed to the scientific evidence discourse surrounding the linkage of MMR vaccines to autism. We limited this analysis to the 74 nodes that are connected to the central graph. Figure 9 shows the domain network graph. In this graph, we distinguish domains by vaccine attitude, with anti-vaccine attitudes in Purple and pro-vaccine attitudes in Orange. To identify the attitude of each domain, we first identified all tweet attitudes where each domain was used. After that, we assigned each domain the attitude that was more strongly represented after normalization. 44 of 74 domains in our graph were classified as being used as a support for a tweet showing an anti-vaccine attitude, while 30 domains were classed as pro-vaccine. The network graph shows a tightly connected cluster of anti-vaccine domains, suggesting that many users cite multiple anti-vaccine sites as a support for their beliefs. Within that cluster, the three most-highly tweeted and most connected domains are ncbi.nlm.nih.gov, ageofautism.com, cbsnews.com, and truthinmedia.com.

Ncbi.nlm.nih.gov (the National Center for Biotechnology Information (NCB) is a branch of the National Institutes of Health. It houses a series of databases relevant to biotechnology and biomedicine. Online users use this domain as a reference for scientifically written and validated articles. However, some of these papers, which are frequently cited by the anti-vaccine community, have been retracted. Ageofautism.com is a site devoted to proving that autism is induced by the environment and that MMR vaccination is the main cause of autism. This domain is highly connected with domains that advocate for the same agenda, such as truthinmedia.com, vaxxedthemovie.com and vaxxed.com. CBSnews.com is a known source of daily news. This domain is the most highly connected domain (after ncbi.nlm.nih.gov and ageofautsim.com) that provides articles verifying and implying a causal link between vaccine and autism. Truthinmedia.com was the second most tweeted domain. Online users with anti-vaccine attitudes share different types of evidence. As an example, one user with apparently strong anti-vaccine attitudes used an article from cbsnews.com:

(DOCTORS - FIRST! Do No HARM: UnSAFE Vax = Autism: COMPELLING 2011 Scientific Review!! CDC=UNTRUSTWORTHY!! #vaxxed <link11>

And another from ncbi.nlm.nih:

2009 PubMedC: Regressive Autism Due to Overuse of Vaccines? (7 YEARS AGO - 1 of SEVERAL STUDIES...) WTW? #Vaxxed <link12>

The use of diverse types of sources to support users' claims is intended to prove that their points of view are accurate and sound. Figure 10 shows the domain network graph distinguished by the evidence type, with scientific evidence in Green and non-scientific evidence in Pink.

To identify if a domain is classified as scientific or non-scientific, we first mapped the number of scientific and non-scientific tweets that used each domain and then assigned the class that was more strongly represented for the domain after normalizing based on class distribution. Only 15 out of the 74 domains were considered to provide scientific evidence; these domains were highly connected to each other and made their own cluster. Some domains were identified as scientific domains even though they did not actually represent a scientific source. This may have happened because these websites were co-cited with scientific references, such as cbsnews.com, change.org, vaxtruth.org and morganverkamp.com.

[11] https://www.cbsnews.com/news/vaccines-and-autism-a-new-scientific-review/

[12] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3364648/

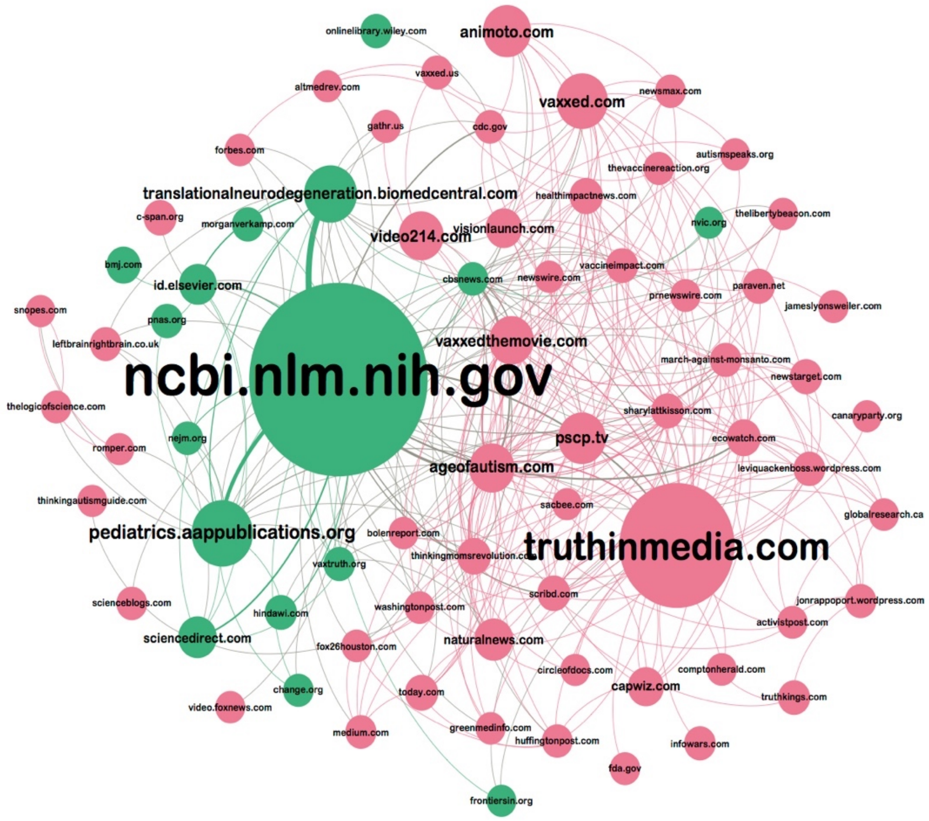


Figure 10: Domain Network Graph, colored by evidence type. Green = Scientific; Pink = Non-Scientific.

These domains include many references to scholarly articles, which may indicate that online users use non-scientific sources that cite scientific sources and regard them to be as sound as scientific references.

The domain of scientific evidence that was most heavily connected to non-scientific evidence domains was the National Center for Biotechnology Information (NCBI) website. This website, cdc.gov and fda.gov were the only governmental websites used as references in our dataset, with the CDC and FDA referenced very minimally. This result is similar to previous research, which indicates that a small number of people have trust in government vaccine experts/officials (Freed et al., 2011). The non-scientific evidence domain most connected to scientific evidence was ageofautism.com (described above). This domain is connected to 7 of the domains with scientific evidence. This may indicate that people consider this domain comparable with other higher credibility websites.

4.13 Conclusion

In this study, our goal was to investigate the use of scientific and non-scientific evidence in social media when discussing a controversial health issue, such as the MMR vaccine debate. This study showcased emerging data analysis approaches. These approaches are inherently interdisciplinary, bringing together principles and practices from health informatics, data analytics, and network analysis. Our results show that online users with anti-vaccine attitudes share more content via Twitter then users with pro-vaccine attitudes, which correlate with previous research (Love et al., 2013). Moreover, they share more tweets linking external references and, specifically, non-scientific evidence. Furthermore, our results show that people with anti-vaccine attitudes share many sources but with low diversity, while people with pro-vaccine attitudes share a smaller but more diverse number sources. Additionally, our results show that vocal journalists have a huge impact on users' opinions. Journalists often report on controversy by presenting claims both for and against an issue in a relatively 'balanced' fashion, which leads to more uncertainty on the part of their readers (Clarke, 2008; Lewis & Speers, 2003).

The overall results of this study can help us make more accurate interpretations of people's attitudes and opinions regarding controversial health topics, such as the debate over vaccines. However, our work is limited in many ways. First, tweets gathered for the "non-scientific" dataset may contain references to scientific papers not found through PubMed. Second, even though previous work has indicated that hashtags can be an indicator of users' opinions (Fang et al., 2015; Xu et al., 2015), some users use hashtags that indicate the opposite attitude to express the opinion that the other side of the debate is wrong or to voice sarcastic opinions regarding the other party. Moreover, given that we only examined one case, the vaccine debate, we are limited in the understanding and analysis of other sources of scientific information that users share online when discussing health issues. In future work, we would like to explore the strength of the attitudes held by each side of the debate and study if people with strong opinions differ in the usage of information sources from users with moderate or no opinions towards the debate. Eventually, we plan to expand our understanding of the use of scientific sources via social media by studying other health topics.

4.14 Categorization and Comparison of Influential Twitter Users and Sources Referenced in Tweets for the topic of MMR vaccine

Social media user networks create important implications for the distribution and nature of health-related information online. Communication modalities such as retweets, comments, likes and/or mentions allow users to cite health information from a variety of online sources (e.g., press releases, news articles), choose relevant audiences (e.g., public vs. private), and contextualize health messages by adding personal opinions in the form of captions or comments. These behaviors potentially define the impact of health messages on public discourse, information-seeking behaviors, and health choices.

Research suggests that social media users tend not to verify the accuracy of information encountered (M. Del Vicario et al., 2016) and/or rate health-related information as credible, irrespective of the source (Eastin, 2001). Hyperlinks associated with health-related discussions on social media, in particular, do not cite scientific information adequately (Sudau et al., 2014). This behavior is particularly notable for controversial health-related topics (Wikgren, 2001). Additionally, past studies highlight that both the type of social media users (e.g., bots, individuals, organizations) and the source of these messages (e.g., fake news, rumors) contribute to health misinformation, controversies, and polarization of health topics. Recent evidence also suggests that users are unable to distinguish between messages posted by social media bots or humans (Edwards, Edwards, Spence, & Shelton, 2014). Social media bots are particularly associated with strategically spreading scientific disinformation or unverified claims (Allem & Ferrara, 2018; David A. Broniatowski et al., 2018). Among individual social media users, influential users such as celebrities or those with high Klout scores (a measure of online social influence) have often stirred debates about health-related topics (Wagner, 2016) or participated in polarized health discussions (Cavazos-Rehg et al., 2015). Government regulatory or health organizations such as the Food and Drug Administration (FDA) or the Centers for Disease Control and Prevention (CDC), despite their efforts in disseminating mass social media health messages, are increasingly perceived to be less credible sources of information (Kowitt, Schmidt, Hannan, & Goldstein, 2017). This nuanced understanding of the differential impact of social media users makes it possible to assess the impact of these users on health-related topics. A critical next step to counter health misinformation more effectively is to identify key information sources cited by these social media users.

Examining hyperlinks or Uniform Resource Locations (URLs) embedded in social media messages offers tremendous potential in identifying health information sources. Social media posts that include URL "citations" are known to increase perception of information credibility (Borah, 2014), accelerate information mobility, and positively influence information seeking and sharing practices (Bao, Shen, Chen, & Cheng, 2013; Borah, 2014; Dumbrell & Steele, 2017; H. Park, Reber, & Chon, 2016; Son, Lee, & Kim, 2013; Suh, Hong, Pirolli, & Chi, 2010).

Social media users evaluate the quality of posts by clicking URLs (Rieh, 2002), and then act on those posts for their own followers even though the source information is often inaccurate or false (Bian, Topaloglu, & Yu, 2012). However, evidence also suggests that URLs cited in social media posts are linked to misinformation/inaccurate information (Shema, Bar‐Ilan, & Thelwall, 2015; Tanaka, Sakamoto, & Honda, 2014).

This research develops a classification schema for influential users who tweet about the Measles-Mumps-Rubella (MMR) vaccine, a controversial health-related topic, and examines key information sources cited by these users in the form of URLs. In general, research suggests that the online discussions about vaccination are associated with unverified claims and incorrect information on the health effects of vaccines (Kata, 2012; Witteman & Zikmund-Fisher, 2012). Anti-vaccine groups link vaccination to harmful health effects such as autism and brain injury to increase skepticism about the effectiveness of the vaccines, use rhetoric of individual freedom to encourage refusal to vaccinate, and co-promote alternative options such as homeopathy (Moran, Lucas, Everhart, Morgan, & Prickett, 2016). The controversy surrounding vaccination has complex public health implications, including widespread vaccine-denial or hesitancy (Dube, Vivion, & MacDonald, 2015), low confidence in medical practitioners (Dube et al., 2015), and low public trust in vaccine effectiveness (Kata, 2010, 2012; Larson, Cooper, Eskola, Katz, & Ratzan, 2011).

The MMR vaccination debate, in particular, has been characterized by similar controversies (Hilton, Petticrew, & Hunt, 2007; Mann, 2018) and gaps in knowledge about its health effects (Ramanathan, Voigt, Kennedy, & Poland, 2018) since the publication of the now retracted scientific article linking MMR vaccine to autism (Beck, 2006; Begg, Ramsay, White, & Bozoky, 1998; J. S. Gerber & Offit, 2009; Speers & Lewis, 2004). Although the United States has been declared measles-free, outbreaks continue to be common. Data on county-level estimates of MMR uptake are inadequate (Kluberg et al., 2017), and children/young adults aged 10-19 are most at risk (Livingston, Rosen, Zucker, & Zimmerman, 2014). The most recent measles outbreak occurred in 2018 with 137 cases reported across 24 states. The majority of these cases were unvaccinated. In 2014, a record measles outbreak was confirmed in an amusement park in California with 667 cases spreading across 24 states. In the same year, another outbreak occurred among 383 unvaccinated cases in Ohio.((CDC), 2018) Recent reports of increasing medical exemption requests for vaccination validated by the medical community in California raise further concerns about more widespread measles outbreaks (Hiltzik, 2018; Mohanty et al., 2018).

MMR debates on social media are generally event- or media-driven. A recent study suggests that pro-MMR posts tend to follow patterns of MMR outbreaks and typically attribute blame for the outbreaks to those who do not vaccinate (Deiner et al., 2017). In contrast, anti-MMR posts appear on social media platforms more regularly and continue to question the health benefits of the vaccines (Deiner et al., 2017). Recent analysis also shows that during measles outbreaks, MMR-related articles containing statistical information are more likely to be shared on social media in the form of embedded URLs (D. A. Broniatowski, Hilyard, & Dredze, 2016).

Previous research provides valuable insights about the nature and flow of MMR-related health messages. A gap in the literature pertains to patterns related to preferred MMR-information sources and types of social media users. In this research, we classified URLs shared in MMR-related tweets from January 1, 2016 to October 1, 2018, based on a lexicon consisting of 10 different categories: *Social Media; Videos; Government; Scientific; National Medical Professional Societies; News; Health Magazines; Health Insurance; Fake News; Commercials; Blogs*. The lexicon was constructed by assimilating publicly available categorizations of web domains into the aforementioned categories. Second, we identified and categorized influential Twitter users into several categories, including *Broadcast News; Blog posts; Scholarly and/or Scientific Sources; Federal, State or Local Government Agencies; National or State Professional Medical Societies and Associations; Educational Instituions*. Finally, we analyze the correlation between different user categories and their URL sharing patterns.

4.15 Data Collection

Data was obtained from Crimson Hexagon[13], a public social media analytics platform. Tweets posted between January 1, 2016 and October 1, 2018 were collected from accounts that set English as their language. The search criteria consisted of the following root terms:

("vaccinations" OR "vaccination" OR "vaccines" OR "vaccine" OR "measles-mumps-rubella" OR "MMR" OR "mmr") AND ("autism" OR "autistic disorder")

The root terms could have appeared in the post or in an accompanying hashtag, for example, vaccine or #vaccine. Retweets were excluded from the sample. The root terms used to collect tweets during the study period resulted in an initial corpus of (N= 222,073) public tweets.

4.16 Types of Social Media Users

4.16.1 Bot Detection

Bots are defined as Twitter accounts that behave and operate in a similar fashion as Twitter accounts operated by humans (Allem et al., 2017), using an openly accessible solution called Botometer, a.k.a. BotOrNot (Davis et al., 2016). Typically, Botometer returns bot likelihood scores greater than 50% for accounts that demonstrate bot-like characteristics, such as high numbers of friends and followers, high numbers of retweets, being mentioned by others, and user activity.

Botometer is based on a supervised machine learning approach (Davis et al., 2016; Varol, Ferrara, Davis, Menczer, & Flammini, 2017). For Twitter accounts, Botometer extracts over 1,000 features relative to the account from data easily provided by the Twitter API and produces a classification score called the bot score: the higher the score, the greater the likelihood that the account is controlled completely or in part by software, according to the algorithm (Yang et al., 2019).

4.17 Influential User Categorization

To identify influential users, we used pre-defined Klout scores obtained from Crimson Hexagon, which are acknowledged as a valid measure of influence (Rao, Spasojevic, Li, &

[13] https://www.crimsonhexagon.com/

DSouza, 2015). Based on the Klout score, which ranges from 1 to 100, the top 25th percentile of users were considered to be influential (Krauss et al., 2015). Within the top 25th percentile of users, 300 users were randomly chosen. Next, informed by past work (Liu, 2016), definitions of each category of social media influencers were specified in a codebook (shown in Appendix A, table 32). Two independent, trained coders assigned codes for (1) individual or organization, and (2) sub-category of individual or organizational tweet for each Twitter user profile. The coders agreed on 252 out of 300 profiles (Cohen's Kappa coefficient 81.5%).

4.18 Information Source Categorization

Past work has manually annotated information sources such as URLs (Addawood & Bashir, 2016). This approach has limitations because of resource intensiveness and limited scalability. One previous work has developed an information source lexicon to classify types of health information sources referenced in social media texts (Addawood, Rezapour, Mishra, Schneider, & Diesner, 2017).

In the same previous work (Addawood, Rezapour, et al., 2017), the authors used a corpus of tweets about the Measles, Mumps and Rubella (MMR) vaccine debate. They demonstrated the application of their lexicon by contrasting the distribution of various information sources. They developed a large-scale information source lexicon by combining data from various open resources. The focus of the lexicon is on identifying different content types present by URL domain; e.g., video, social media, blog, news, fake news, and scientific communication. This lexicon allows for a simple and high-recall identification of information source types present in social media content.

4.19 Lexicon Construction

Most URLs used in Twitter start with https://t.co, which does not reference an original URL but instead is a URL shortener that helps to stay within Twitter's character limit. These URLs need to be expanded to extract the original domain names.

Since there are many URLs that our manually built lexicon was not able to categorize, we decided to manually annotate each of these URLs to have a better picture of the usage of different information sources. To construct the lexicon, we first identified dominant categories of online health information sources as cited in the sample by manually going over a subsample of URLs (n=100), referring to previous work in this area (Addawood & Bashir, 2016) and consulting with a health librarian. The emerging 10 categories were *Broadcast News; Blog Posts; Scholarly and/or Scientific Sources; Federal, State, or Local Government Agencies; Health Corporations; Commercial Content; Health Magazine Websites; Videos; Educational Institutions; Other.* A subsample of tweets (n=50) were coded using these initial categories. A new category of Other was generated when many URLs seemed to belong to an identifiable new type. The final lexicon was tested on a subsample (n=100) to ensure that this lexicon contains comprehensive categories of the different types of health information sources. Moreover, since Twitter users may share tweets with each other, we decided to consider Twitter as a separate category. The final version of the lexicon included these categories (n=12): *Broadcast News; Blog Posts; Scholarly and/or Scientific Sources; Federal, State, or Local Government Agencies; Commercial Content; National or State Professional Medical Societies and Associations; Commercially Subsidized Health Websites; Health Magazine Websites; Health Insurance; Educational Institutions; Social Media; Twitter; Videos; Biased/Opinion-Driven News; Other* (Table 33 in Appendix A contains a description of each information source type, the count of each type, and examples of each source type). For each source type, we list the sources where they were constructed; please see Appendix A for more information.

Blogs: Most popular domains for blogging, such as Word Press, Tumblr and BlogSpot. We leveraged a curated list from Wikidata.[14]

Commercial Content: Well-known commercial websites, such as Amazon and eBay.

Fake News Sources: 1) A list developed by Melissa Zimdars and her research team at Merrimack College,[15] which contains a curated resource for assessing online information sources, available for public use; 2) a list of fake news websites from Wikipedia;[16] and 3) FakeNewsChecker.com, which relies on several reputable sites and third-party sources to identify fake news websites.

News Outlets: As a starting point for collecting news sources, we used a list of news media websites created by Facebook.[17] This list, which includes trusted domains for news, is curated by

[14] https://www.wikidata.org/

[15] http://www.opensources.co

[16] https://en.wikipedia.org/wiki/List_of_fake_news_websites

[17]http://newsroom.fb.com/news/2016/05/information-about-trending-topics/

Facebook's trending topic team. We also used usa.gov[18] to collect direct links to every federal agency and state, local, and tribal government in the USA.

Scholarly and/or Scientific Sources: We manually collected the most well-known scientific publishers, such as NIH, BMJ, and Springer.

Federal, State, or Local Government Agencies: Any website with a *.gov* extension or any state-supported website with a two letter state extension.

National or State Professional Medical Societies and Associations: Established national or state organizations, societies, or associations with a *.org* extension.

Commercially Subsidized Health Websites: Websites that provide health information from for-profit entities such as hospitals, health systems, clinics, and independent groups; for example, WebMD.

Health Magazine Websites: Magazines that cover a variety of topics, including physical fitness and well being, nutrition, beauty, strength, bodybuilding, weight training, etc.

Health Insurance: Insurance that covers the whole or a part of the risk of a person incurring medical expenses.

Educational Institutions: Information from colleges and universities with a .edu extension.

Social Media: We added the most well-known social media domains, including YouTube and Facebook, to the lexicon.

Twitter: Since Twitter users may share tweets with each other, we decided to consider Twitter as its own category.

Videos: To compile a list of video sharing websites such as youtube.com and vimeo.com, we used a curated list from Wikipedia.[19]

Biased/Opinion-Driven News: A semi-official broadcast channel that has some kind of an agenda with selected information. Websites such as Trump.tv can fall under this category.

Due to the uncertain credibility of biased/opinion-driven news, we decided to include more guidelines to understand each of these websites' credibility. A domain that fell into the blog or biased/opinion-driven news category required further assessment of whether sources were fact- or opinion-based. Fact-based sources consisted of articles from peer-reviewed journals, information from government agencies, or quotes from primary sources. Opinion-based sources were those that linked to articles from unaccredited domains or those that did not link to any

[18] https://www.usa.gov/federal-agencies/a

[19] https://en.wikipedia.org/wiki/List_of_video_hosting_services

sources at all. The URL categorization was not mutually exclusive in that some URLs belonged to more than one category (e.g., youtube.com was placed in both the social media and video categories).

The lexicon constructed was first applied to our data computationally using Python. We matched all URLs in the tweets with the domain names in the lexicon. After applying this lexicon to our dataset, we found that some domain names that did not fall into any of the identified categories. We identified this category as *unknown*. For this category, a researcher with a library and information sciences background went through the list of domains and labeled each URL using a set of guidelines and a codebook (as seen in Table 2, Appendices A and B). A second annotator annotated 35% (n= 525) of the domains to validate the annotation.

4.20 Bot Analysis

As seen in table 21, about 29% of user accounts (n= 65,375) in the sample did not have a bot score because these were either not found or had restricted access. Approximately 10,207 accounts had bot scores greater than 0.5, resulting in a sample size of 27,087 tweets. Exclusion of bot accounts yielded 64,539 accounts and 129,611 tweets.

Table 21: Distribution of accounts in our dataset

Account type	User count	Tweet count	Tweet %
Bot	10,207	27,087	12.19
Not bot	64,539	129,611	58.36

4.21 Information Source Analysis

4.21.1 Overall URL analysis

First, we looked at the top domains in our dataset as a whole. Our results showed that overall, we have 104,093 tweets (44.7%) containing a total of 118,498 URLs. This resulted in a dataset of 49,137 unique URLs from 4,131 unique domains.

4.21.2 Information Source Type Distribution

The distribution of the different types of information sources in the dataset based on the created lexicon is shown in figure 11 below. *Twitter* as a type of information source was excluded from the figure since there were around 31,214 URLs with this domain type. We believe that there is this number of URLs from Twitter because people on Twitter tend to reference each other or reply to each tweet. This behavior increases the number of Twitter references in the text of the tweet. Moreover, we found 1,171 URLs that are invalid because either the URL was removed or we could not access it due to firewall restrictions.

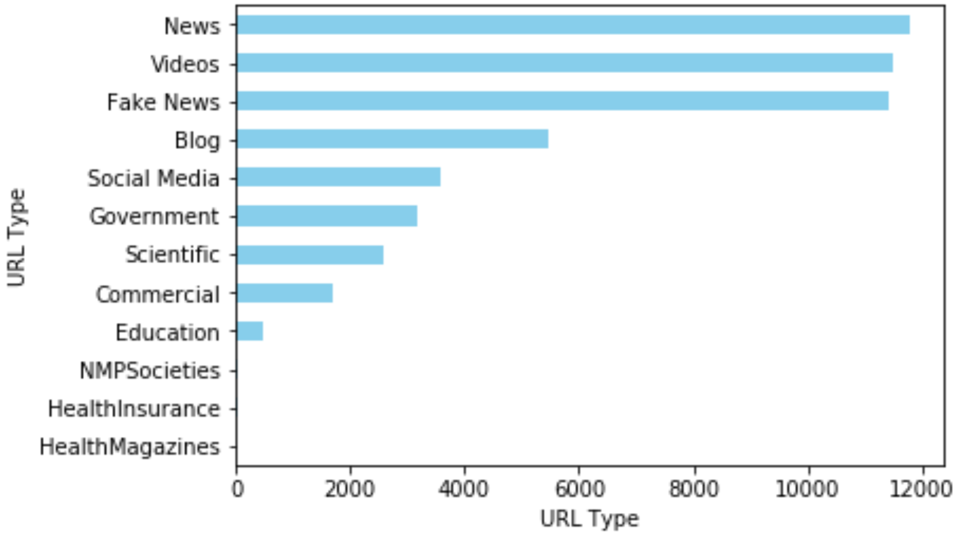


Figure 11: Distribution of most Information Source Types

4.21.3 Unknown URL analysis

The *Unknown* URL category contained 33,867 URLs in total. 3,605 of these URLs were unique. Table 22 below shows the five most frequent unknown URLs. *Inshapetoday.com* is the most website that was most frequently not categorized into any information source type. When investigating the website, it has been found that it contains different health-related topics and there was not an “about us” page. This website and similar ones can be classified as *biased /opinion-driven news* websites.

Table 22: The frequency of the top five unknown URLs

URL	Frequency
http://inshapetoday.com	2269
https://newspunch.com	731
https://www.hopkinsmedicine.org	684
http://www.alternativenewsnetwork.net	670
https://vaccineimpact.com	622

In total, 1,000 URLs were annotated by a health librarian. 258 URLs were categorized as N/A, primarily because these websites no longer exist on the internet. Figure 12 below shows the distribution of URLs categorized by human annotator. As shown below, *Other* and *Blogs* have similar frequencies (~18%). *Biased/Opinion-Driven News* comprises 16.8% of total unknown URLs. For *Biased/Opinion Driven News* URLs, each URL was annotated based on six questions (shown in Appendix B). These questions can help with better identification of the credibility of each website.

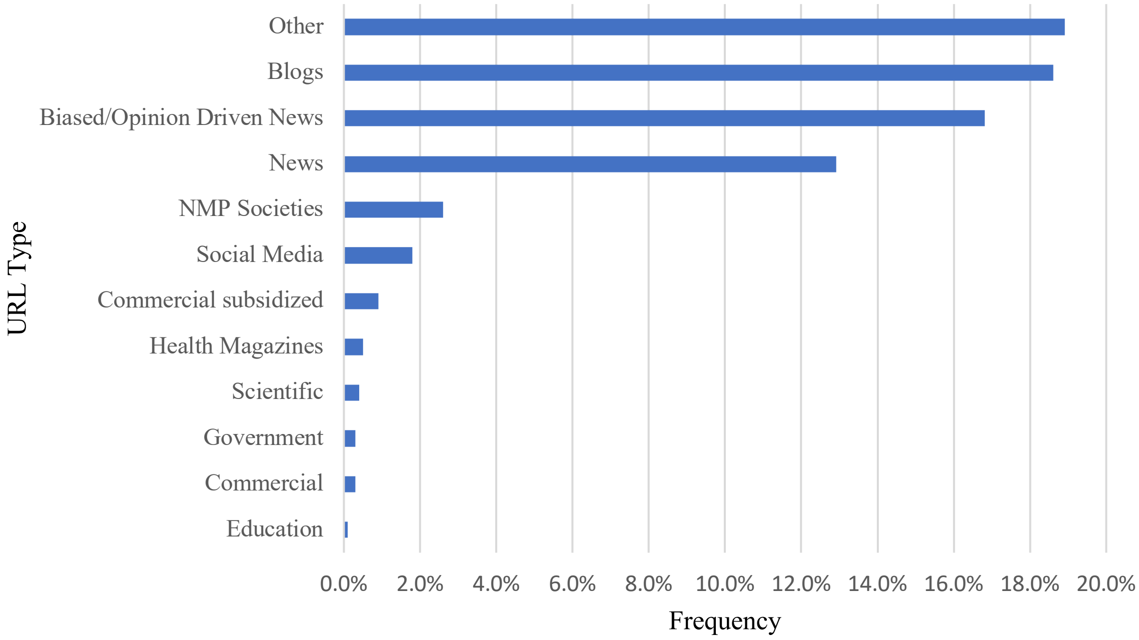


Figure 12: The frequency of annotated unknown URLs

The first question investigates who runs or created the website. The annotator attempted to find an "about us" page on the website. The website would either be a *for-profit organization*, where the website is either selling something or promoting its own products, or a *not-for-profit organization* if it is self-identified as non-profit. Finally, the website will be labeled as *not clear who created the website* if it was not clear whether it is for profit or not. Our results showed that out of 168 *biased/opinion-driven news websites*, 55% were not clear about who created the website, while only 10% were created by not-for-profit organizations and 35% created by for-profit organizations.

The second question addressed when the information was written or reviewed, as old information is generally deemed to be less credible. The website is labeled *up-to-date* if it uses scholarly articles published within five years of 2018 and *not up-to-date* if it uses scholarly articles published more than five years prior to 2018. The website is labeled *not sure when the source(s) was posted* in the case of non-scholarly source(s). Our results showed that 90% of the websites were labeled as *not sure when the source(s) was posted*, where only 3% had up-to-date sources.

The credibility of the source of the information is an important factor. For the third question, the website is labeled *fact-based* (i.e., based on scientific research) if it contains links to scholarly articles and known reputable authorities. The website is labeled *opinion-based* if it contains links to unreliable or non-authoritative sources. Finally, the website is labeled *not clear where the information came from* when the website contains links to both authoritative and non-authoritative sources, or if it contains no links to other sources. Our results show that 54% of the websites were unclear as to where the information came from, while only 8% used fact-based sources.

The reason behind creating a website is important, as whether or not it was created based on advertisements can aid in identifying the website's sponsor(s). For the fourth questions, websites can be labeled as either containing *advertisements with disclosure,* if there is a statement about the advertisements and whether the site is making money from them, or *advertisement without disclosure*, if the website is lacking a statement about the advertisements. Finally, the website will be labeled as *No Advertisements* if there are no advertisements on the website. Our results show that 38% of the websites contained no advertisements, while 32% contained advertisements without disclosure.

Having a "Contact Us" page is another indication of the credibility of the website. Our results show that 67% contained a working "Contact Us" link, compared to 32% that did not.

The final question addressed whether or not the website is health-related. A website is labeled *health-related* if it is based solely on a health-related subject(s), and labeled *not health-related* if it does not discuss health-related topics or has a mix of health- and non-health-related topics. Our results showed that 77% of the websites are actually not health-related, while only 21% are solely based on a health-related subject.

4.21.4 Bot vs. Non-Bot Accounts

Bot accounts could have an effect on correctly identifying the sharing behaviors of URLs in Twitter regarding MMR vaccines. Figure 13 below shows the distribution of URL types between bot and non-bot accounts. The most frequently shared domain was Twitter (not shown in the figure), with 13.84% shared by non-bot accounts compared to only 3.86% shared by bot accounts. As shown in the figure below, bot and non-bot accounts mostly share *Videos*, *News*, and *Fake News* sources.

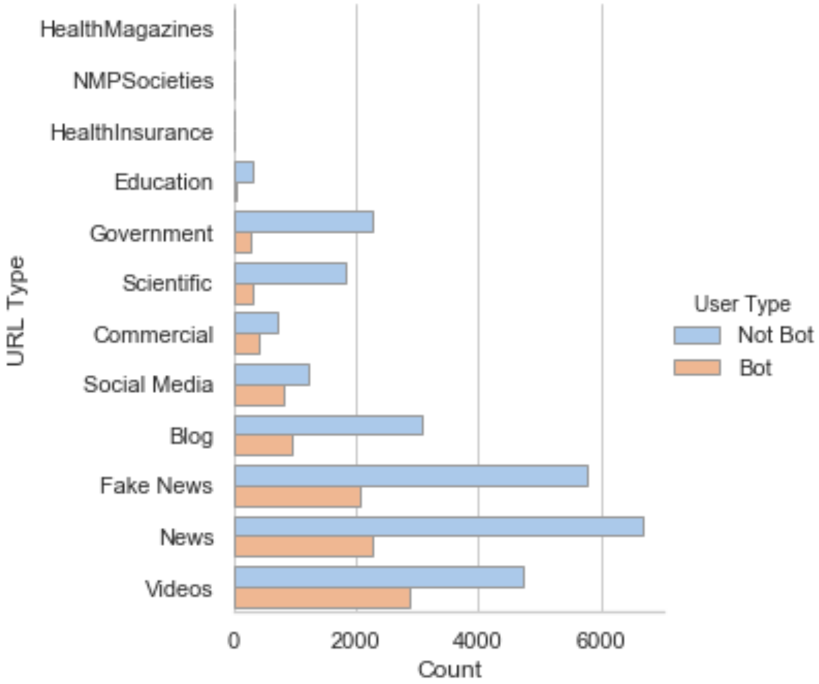


Figure 13: Distribution of URLs between bot and non-bot accounts

4.22 Influential User Analysis

One important part of our analysis is understanding who shares vaccine-related information on social media. To do that, three annotators annotated 287 user accounts. Each account was classified by user category (individual or organization) and a specific dimension within these categories; 100 users were cross-annotated and resulted in an inter-annotator agreement score of 0.83. Out of these annotations, approximately 2.78% of user accounts were not categorized due to one of the following reasons: Account Suspension; Protected Account; No Bio Available; No longer existing account. Within the annotated Twitter users, 78.39% of users were categorized as individuals and the remaining 18.46% as organizational accounts. Figure 14 below shows the frequency of individuals' account types. Layperson is the account type that most frequently participates in vaccine-related discussions on Twitter at 30.22%. Health writers come after that, with 16.4% of accounts identified as writers.

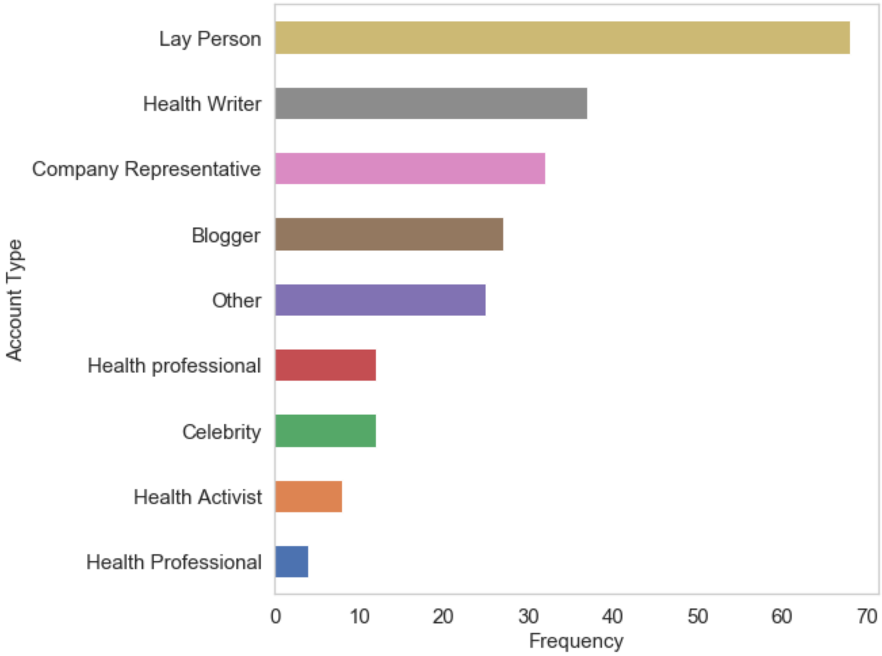


Figure 14: The frequency of individual account types

Figure 15 below shows the frequency of accounts that are labeled as organizational accounts. Media accounts (official, non-official, and blogs) have the highest frequency, with 35.8% official, 24.52% non-official, and 5.66% blogs.

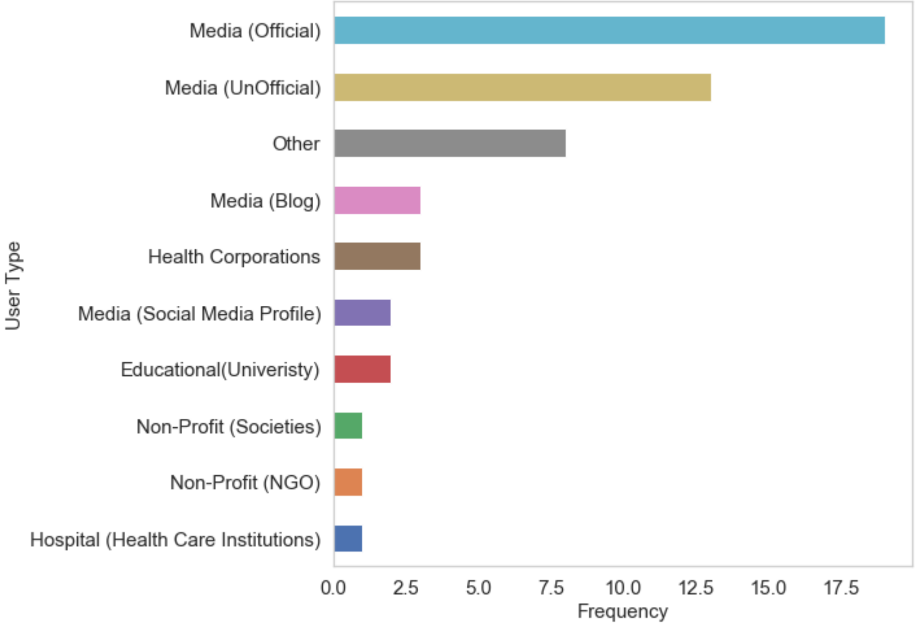


Figure 15: The frequency of organizational account types

4.23 URL and User Relationship

One important insight that we were interested in understanding is the type of sources certain online users share. For example, do doctors share more valid sources compared to other users? To accomplish that, we analyzed individual and organizational accounts' URL sharing behaviors. Figure 16 below shows the frequency of different URL types for each individual user account. We can see that laypeople usually share Twitter URLs, which might be responses to others or retweeting other tweets. Also, we can see that they share fake news at the highest rates. Laypeople and bloggers are the highest in sharing news compared to health professionals, who do not share any news.

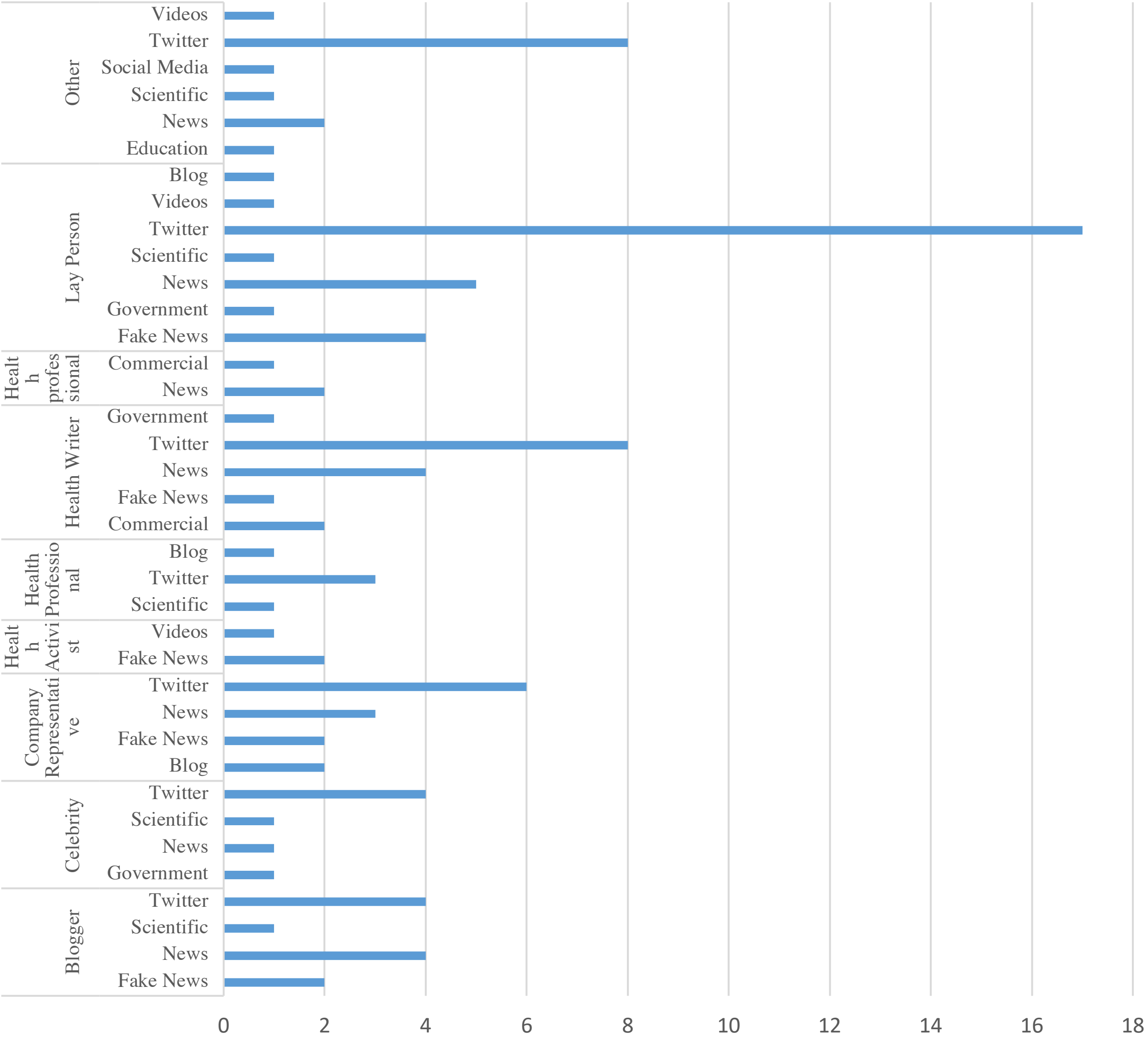


Figure 16: URL and individual user account relationship

Figure 17 below shows the frequency of different URL types for each organizational Twitter account. As shown, official media accounts frequently cite Twitter URLs compared to other organizational accounts. Official and non-official media accounts frequently cite news URLs. Non-profit societies are the only ones citing scientific URLs.

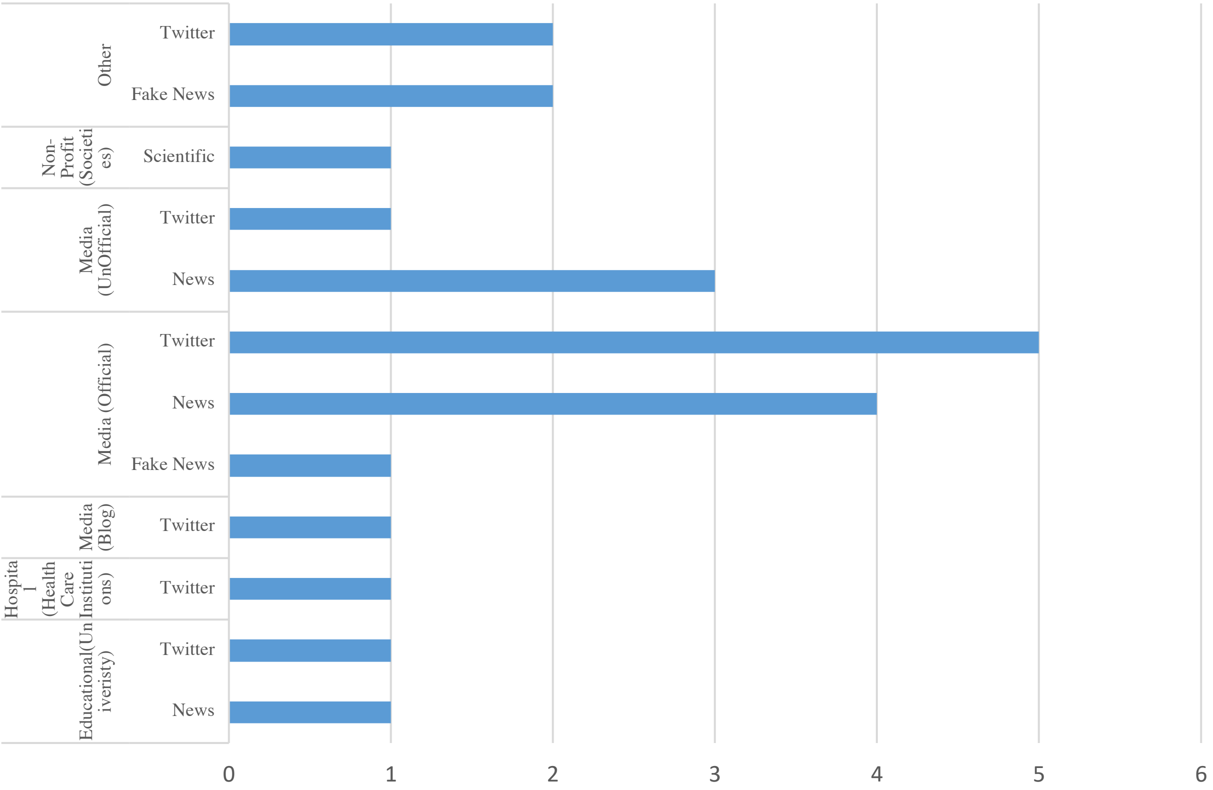


Figure 17: URL and organizational account relationship

4.24 Discussion and Conclusion

This research offers evidence supporting the Fuzzy Trace Theory (FTT) – a leading theory of medical decision-making that explains the process by which individuals derive meaning from information they are given (Cornelia Betsch, Renkewitz, & Haase, 2013). It appears that URLs cited in social media posts are linked to misinformation/inaccurate information. In our study of a controversial health-related topic – MMR vaccines – we find that the online discussions about vaccination are associated with unverified claims and incorrect information about the health effects of vaccines. Our research aligns with other studies on effective vaccine communication during the Disneyland Measles Outbreak, as they found that both social media and provider recommendations can influence parental decision-making about vaccination, and the findings suggest practical implications and future research questions for public health communicators and clinicians (D. A. Broniatowski et al., 2016).

Other studies tried to see whether people articulate their information needs and provide information to others differently in online sites of various types, specifically blogs and internet discussion forums (Savolainen, 2011). The findings reported that the bloggers, blog readers, and discussion group participants mainly articulated needs related to getting an opinion or evaluation of an issue, while needs for factual information and procedural information about possible ways of action were presented less frequently. Information provision drew strongly on the use of personal knowledge. There were no remarkable differences between the types of online sites with regard to the articulation of information needs and using sources for providing information to others. On the other hand, Bryan et al. (2018) studied the content and accuracy of vaccine information on pediatrician blogs. A national sample of pediatrician blogs was identified using a search rubric of terms applied to multiple search engines. The objective of that study was to assess content, citations, audience engagement, and accuracy of vaccine information on pediatrician blogs. The conclusions of that study showed that pediatrician bloggers frequently address vaccinations; most provide accurate information. Pediatrician blogs may be a new source to provide vaccine education to parents via social media. Our study had a different view, as blogging differs from Twitter. In our research, the annotated Twitter users (78.39%) were categorized as individuals, with the remainder (18.46%) categorized as organizational accounts.

In their study entitled "Measles, the media, and MMR: Impact of the 2014–15 measles outbreak," Catali et al. (2016) investigated how knowledge and attitudes varied with the type of media sources mothers trusted most. The results showed that new mothers had high levels of knowledge and favorable attitudes about vaccination after the 2014–15 measles outbreak. The most frequently used media sources are not the most trusted ones. Communication about outbreaks of vaccine-preventable diseases should include spreading accurate information to new media sources and strengthening existing trust in traditional media. Other researchers found that the online discussions about vaccination are associated with unverified claims and incorrect information about the health effects of vaccines (Kata, 2012; Witteman & Zikmund-Fisher, 2012). Our study found that the layperson is the account type participates most frequently in vaccine-related discussions on Twitter (30.22%).

(Aquino et al., 2017), MMR vaccination rates in Italy have been decreasing since 2012; at present, none of the Italian regions has achieved the target vaccination rate of 95%. Their study aimed to explore the relationship of MMR vaccination rates to online search trends and social network activity on the topic "autism and MMR vaccine" during the period 2010-2015. A significant inverse correlation was found between MMR vaccination rates and Internet search activity, tweets, and Facebook posts. New media might have played a role in spreading misinformation. Media monitoring could be useful to assess the level of vaccine hesitancy and to plan and target effective information campaigns. In our research, our results showed that 77% of the websites are actually not health-related, while only 21% are solely based on a health-related subject.

Semantic network analysis of vaccine sentiment on social media aims to examine current vaccine sentiment on social media by constructing and analyzing semantic networks of vaccine information from websites frequently shared by U.S. Twitter users, as well as to assist public health communication about vaccines (Kang et al., 2017). This study concluded that semantic network analysis of vaccine sentiment in online social media can enhance understanding of the scope and variability of current attitudes and beliefs toward vaccines. The study synthesizes quantitative and qualitative evidence, using an interdisciplinary approach to better understand complex drivers of vaccine hesitancy for public health communication and to improve vaccine confidence and vaccination rates in the United States. Our study aligns with other studies of social media posts about measles vaccination. These posts were classified as pro-vaccination, expressing vaccine hesitancy, uncertain, or irrelevant. The findings may result from more consistent social media engagement by individuals expressing vaccine hesitancy, contrasted with media- or event-driven episodic interest on the part of individuals favoring current policy (Deiner et al., 2017).

CHAPTER 5: SOCIAL DEBATES IN TWITTER

5.1 Introduction

Twitter provides a window into peoples' opinions about issues of public interest. In this paper, we analyzed the stance, gender and location of tweets and tweeters related to the controversial issue of women driving in Saudi Arabia. We used a sample of tweets between 2012, when the first campaign for women driving began, until 2017, when the government issued a policy that allowed women to drive. We manually labeled 4089 tweets for stance (i.e., being in support, against, or neutral on this topic.

Most of the contents in this chapter are published in (Addawood, Alshamrani, Alqahtani, Diesner, & Broniatowski, 2018). It is joint work with Amal Alqahtani, Amirah Alshamrani, Jana Diesner and David Broniatowski.

5.2 Women's Driving in Saudi Arabia

Social media platforms such as Twitter are popular around the world. With 13.8 million active users out of 24 million internet users (91% of total population), Saudi Arabia is among the countries with the highest number of Twitter users among its online population(Statistic, 2018). Moreover, Saudi Arabia is producing 40% of all tweets in the Arab world (Mourtada & Salem, 2014). One of the issues discussed on Twitter is the permission for women to drive in Saudi Arabia (Addawood et al., 2018). This longstanding issue has been more than a regulatory in Saudi Arabia; a country that has had undergone social and economic changes with their new governmental regime. The issue of women driving highlights the historical and ideological conflict in Saudi Arabia between conservative and more liberal voices. On October 26th, 2013, a social movement began when 60 women drove their cars in the streets of Riyadh, the capital of Saudi Arabia (agencies, 2013). This movement sparked heated debates between people for and against women driving in Saudi Arabia. On 26 September 2017, the policy changed to grant women permission to get Saudi drivers' licenses was officially announced. Measuring public opinion is challenging in Saudi Arabia, e.g., due to a lack of polling data. In addition, previous studies based on other societies and languages might not generalize to the culture of Saudi Arabia due to differences in traditions and norms. In this study, we capture a piece of this culture by collecting and analyzing data from a public channel, namely Twitter, which can offer a window into public opinion and the role of social media in Saudi Arabia. More specifically, we aim to explore peoples' attitudes towards the topic of allowing women to drive in Saudi Arabia, and if these attitudes shifted when relevant events happened. We also examine the gender and location of tweet authors and if these features correlate with peoples' stances towards the given topic. This work can help explore the relationship between a change in policy and the expression of peoples' opinions on social media. This study might also inform the development of models of social behavior that fit the Saudi Arabic culture.

5.3 Online User's Opinion Analysis

Social media can have an impact on people, e.g., by challenging existing norms and integrating different opinions. Borge-Holthoefer and colleagues (2015) explained how Twitter can provide a platform for modern protests such as the 2013 Egyptian coup. They used content analysis and network analysis to trace opinion changes in Egypt's population during the protests. Their results show little evidence of users changing sides between pro-military/anti-military and Secularist/Islamist camps. Another study by Magdy and colleagues (2016) argued that social media can be used to predict future attitudes and stances. This study investigated the attitudes of U.S. Twitter users towards Muslims reactions to the 2015 terrorist attacks in Paris. The authors used tweet contact and network interactions to make a distinction between online speech that attacks/blames, defends, and is neutral towards Muslims after a terrorist attack. Their results showed that it is possible to predict users' stances toward a social issue on Twitter since people tend to agree with like-minded others (homophily). Abokhodair and colleagues (2016) attempted to understand the opinion of online users regarding online privacy and its value in Arab golf countries, mainly Qatar. They analyzed tweets that mentioned "privacy", focusing on how digital contexts can lead to different opinions about privacy. Their results showed that users from Arab golf countries value their privacy widely because of religious reasons. Moreover, when men are discussing women's' privacy, they tend to use authoritarian language because they link their privacy to their honor. Another study by Abokhodair and Vieweg (2016) also discussed the concept of privacy and social media use in two of the Arab Gulf country, Qatar and Saudi Arabia. Their results confirmed that privacy is not a choice but can be imposed by cultural norm, specifically for women.

A related study by Al-Dawood and colleagues (2017) indicated that Saudis are restricted by their cultural boundaries when it comes to online dating and matchmaking. They found that it is very hard for males and females to meet or find potential partners face-to-face because gender segregation is culturally enforced. Moreover, they showed how social media, and technology in general, can help with such cultural restrictions. Despite multiple papers that address the engagement of individuals with societal issues on social media, our body of knowledge about the discourse of controversial social and cultural issues in Saudi Arabia is still limited.

5.4 Data Collection

For this research, we collected publicly available social media data from Twitter by using Crimson Hexagon (S. Etlinger & W. Amand, 2012). First, we sampled public posts from March 1, 2012 (before the first campaign on women driving) through September 30, 2017 (governmental announcement of allowing women to drive). The sample only included tweets from accounts that set Arabic as their language. The following query was used (translated from Arabic to English below):

English	**Arabic**	**Transcript**
Driving OR Drive	قيادة - قياده- سواقة - سواقه	sewaqah – sewaqat – qeyadah - qeyadat
Women OR Females	حريم - النساء - المراة - المرأه	almara'h – almara't – alnisa' - hareem
Car	سيارة - سياره	sayarah -sayarat
With, Against, Ban, Refusal, OR Cancellation	- رفض - منع - ضد - مع لن تقودي	Mae – dhd –manae – rafdh – lan taqudi

Using this query, we collected 106k tweets, which we then divided into four time periods as follows: Time period 1: March 1st, 2012 - September 30th, 2013; represents the time before the first driving campaign movement (9,628 tweets). Time period 2: 1st – 31th, October 2013; the month during which the driving campaign began (10,826 tweets). Time period 3: 1st, November 2013- 31th, August 2017; before the governmental announcement of allowing women to drive in Saudi Arabia (65,437 tweets). Time period 4: 1st - 30th, September 2017; the month during which the government announced the permission for women to drive (10,247 tweets). During annotation, we noticed a considerable number of repeated tweets written by different users (10% of the total dataset), which were removed to avoid bias. This left us with 96,138 tweets.

5.5 Stance Identification

Stance identification typically seeks to identify whether a person is for or against some given issue (Addawood, Schneider, et al., 2017). To have an equal distribution of data per time period we randomly selected 10% of the tweets from each time period. That resulted in 4,089 tweets to be our sample for manual tagging of the stance per tweet. We annotated each tweet as either being "positive" (i.e., in favor of women driving), "negative" (i.e., against women driving), or "other" (i.e., tweets with no clear or strong stance for or against women driving, and tweets with contradicting or unclear stances). The first two authors, whose native language is Arabic, annotated the tweets. The third author, who is also familiar with the Arabic language and Saudi culture, intervened when there were any disagreements. By the end of the annotation process, approximately 5% of the total annotated tweets were labeled by both coders. We used Cohen's Kappa (Cohen, 1960) to measure inter-annotator agreement. which was $\kappa = 0.756$ (95% CI, 0.6742 to 0.8382), $p < .0005$.

5.6 Identification of Author Gender

For Arabic names, automatic annotation of gender might not give accurate results. Thus, we decided to perform another type of manual labeling for the same 4,089 tweets used for stance identification. Annotators looked at each user's name, screen name, and link to their Twitter profile. Annotators relied on their cultural knowledge to infer the correct gender as either being "female", "male", or "not sure".

5.7 Identification of Author Location

Although allowing women to drive is a Saudi Arabic concern, this topic received international attention. To find out if Twitter users from other countries are also involved in this discussion, we retrieved locations as provided by Crimson Hexagon (S. Etlinger & W. Amand, 2012). To infer locations, Crimson Hexagon uses two types of information: 1) geotagged locations, which are only available for about 1% of Twitter data (Jurgens, Finethy, McCorriston, Xu, & Ruths, 2015; Morstatter, Pfeffer, Liu, & Carley, 2013); and 2) for tweets that are not geotagged, an estimation of the users' countries, regions, and cities based on "various pieces of contextual information, for example, their profile information", as well as users' time zones and languages[20].

5.8 Stance Analysis

Overall Stance Distribution. Out of 4,089 tweets, 1,689 (41.3%) had a positive stance on the issue, while only 931 tweets (22.7%) had a negative stance, and the rest of the tweets were classified as "others". In other words, about 2/3 of the tweets had a clear stance, and out of those tweets, the majority was in support of women driving.

Stance Distribution over Time. Figure 18 shows the distribution of stances across the four time periods. There are only minor differences in stance between the time periods. Our results show that before the social movement and after the announcement of the new regulations (time periods 1 and 4), there was a higher agreement for women to drive. Most importantly, when the new regulation was announced, the percentage of agreeing to the new law was the highest.

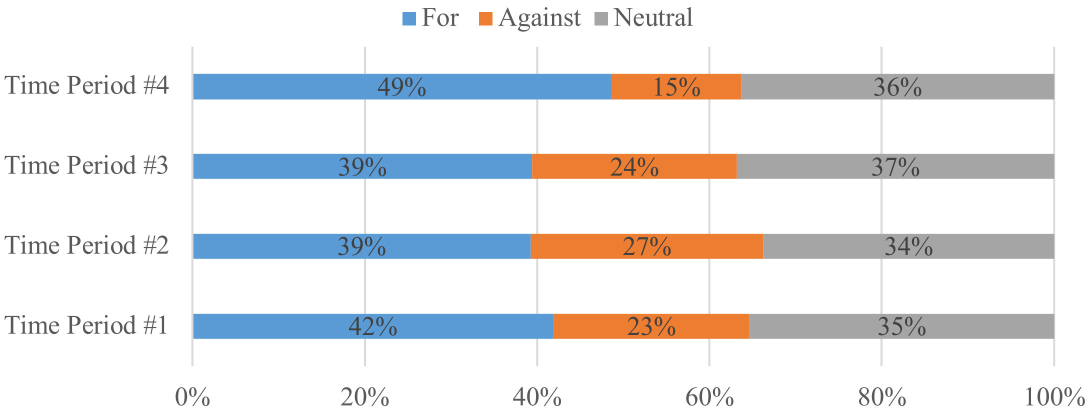


Figure 18: Stance distribution over the four time periods.

5.9 Gender Analysis

- **Overall Gender Distribution.** Since Saudi culture is a gender-unequal society (Elamin & Omair, 2010), we were interested in measuring if author gender correlates with stance. In our annotated sample, 2,443 (59.7%) of 4,089 tweets were posted by male authors, and 1,428 tweets (34.9%) by female authors. This matches previous studies that report low percentages

[20] https://crimsonhexagon.com/

of Saudi women participating in social media (Mourtada & Salem, 2014), which might be due to ongoing societal and cultural constraints imposed on some females when it comes to using social media (Mourtada, Salem, Al-Dabbagh, & Gargani, 2011).

- **Gender Distribution over Time Periods.** Figure 19 shows the gender distribution across over time. The participation of women slightly decreases as the events unfold.

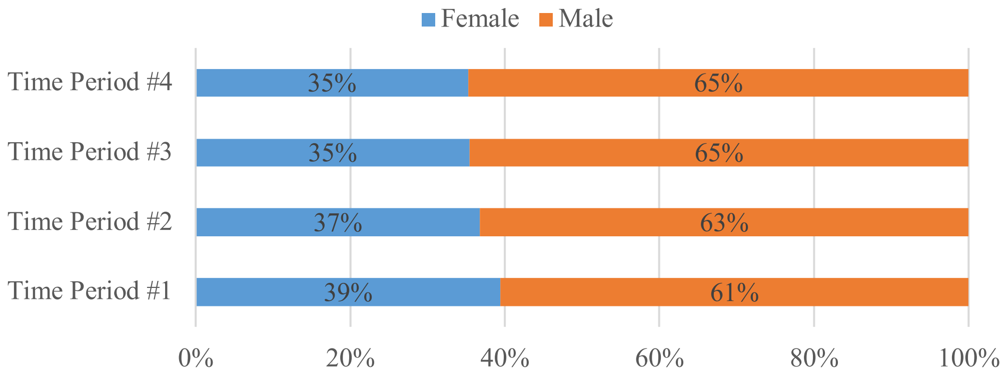


Figure 19: Gender distribution across the four time periods

- **Relationship between Gender and Stance.** To gain a better understanding of the relationship between author gender and stance, we analyzed the gender distribution across stances. A previous study has shown that some Saudi males report traditional attitudes towards working females (Elamin & Omair, 2010). Our results, shown in figure 20, show that in our sample, women are more opinionated (as opposed to neutral, i.e., the "other" category) on this topic: more women (47%) than men (40%) were in support of women driving in Saudi Arabia, and women (26%) than men (22%) were against women driving in Saudi Arabia. One explanation for this observation might be that the Saudi people are primed by the long tradition of driving being only permitted for men or stereotypes about male leadership roles (Al-Ahmadi, 2011). Moreover, this effect might be also attributed to the language used to describe driving in Arabic, as the word "قيادة" means "driving" as well as "leading" in English. Overall, males (38%) had more "other" opinions then females (27%). A recently conducted study in a start-up company in the United States had shown that men tend to refrain from participating in gender-parity initiatives because they feel that it is not their place or psychological standing to be involved in such discussion (Sherf, Tangirala, & Weber, 2017).

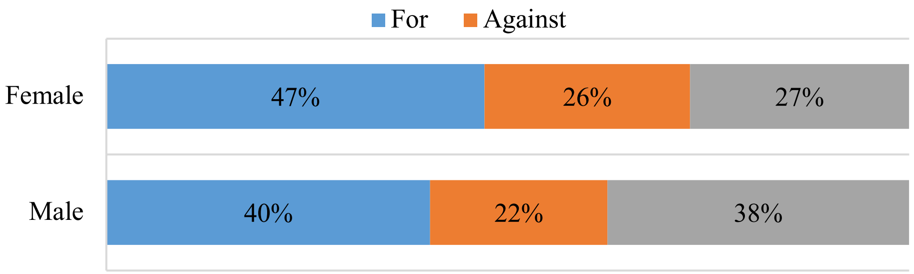


Figure 20: Gender distribution across stance.

5.10 Country Analysis

- **Overall Country Distribution.** We first analyzed the distribution of the tweet authors' country to see where most posts originated from. Out of the original dataset that contained 96,138 non-repeated tweets, 42.25% had a location. Of these tweets, 29,006 (30.17%) originated from Saudi Arabia. The second highest number (4,097 tweets, 4.26%) came from Turkey. We were curious about this relatively high number of tweets from Turkey (the third highest contribution came from the U.S, with 1,818 tweets/ 1.89%). Moreover, we noticed that several tweets with similar, spam-like content originated from Turkey. To investigate this effect in more detail, we attempted to identify whether these accounts were bots or not. We used Botometer[21] to evaluate 3,618 tweets (out of the 4,097 from Turkey) and found that approximately 25% of the accounts were no longer accessible for analysis. Among the accessible accounts, Botometer identified 3% of the accounts as being bots, indicating that the majority of the tweets from Turkey were not from bots.

- **Relationship between Location and Stance.** To gain a better understanding of the relationship between author location and attitude towards women driving, we analyzed the location distribution across stances. We selected the three top countries with the highest ratio of tweets in our labeled corpus of 4,089 tweets. These were Saudi Arabia (51%), Turkey (8%), and the USA (3.6%). Figure 21 shows the distribution of stance for the tweets originating from these countries. The highest ratio of tweets in favor of women driving in

[21] https://botometer.iuni.iu.edu/

Saudi Arabia came from the USA, and the highest number of tweets voicing opposition to this concept came from Turkey. One limitation to these findings is that location might not equate nationality.

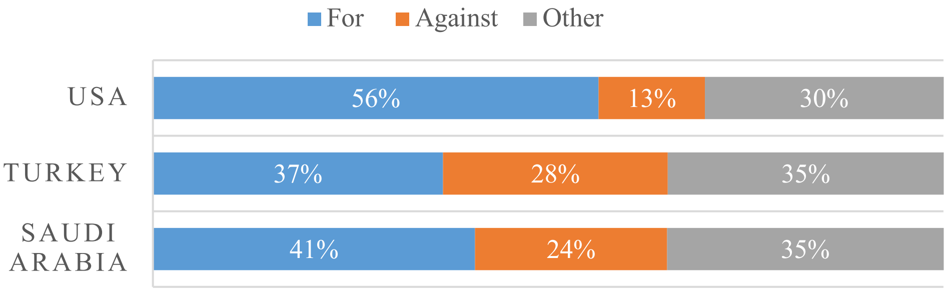


Figure 21: Location distribution across stance.

5.11 Conclusion

In this chapter, we explored the stance towards a controversial topic of public interest (allowing Saudi Arabic women to drive) based on a sample of tweets. To achieve that, we partitioned a set of tweets into four time periods that represents major events that relate to that topic. Our results show that there was less opposition than support for this topic, and that the ratio of opposing tweets was lowest after the change was officially and publicly announced. More men than women are represented in our sample. The women in our sample were more opinionated (both, pro and against women driving) than the men.

This analysis has several limitations. First, the manual mark-up required annotators to understand Arab and Saudi culture. Therefore, our dataset is comparatively small. Second, using Twitter and Crimson Hexagon as a data source and collection tool involves multiple types of potential sampling biases (González-Bailón, Wang, Rivero, Borge-Holthoefer, & Moreno, 2014; Hargittai, 2015; Ruths & Pfeffer, 2014). For future work, we are expanding the dataset by collecting current and upcoming tweets, with the new policy taking effect in June 2018. Our work will focus on building a classifier and a model to help predict the publics' acceptance of this policy change.

CHAPTER 6: SOCIAL MEDIA AND POLITICAL CAMPAIGN

6.1 Introduction

The ease with which information can be shared on social media has opened it up to abuse and manipulation. One example of a manipulation campaign that has garnered much attention recently was the alleged Russian interference in the 2016 U.S. elections, with Russia accused of, among other things, using trolls and malicious accounts to spread misinformation and politically biased information. To take an in-depth look at this manipulation campaign, we collected a dataset of 13 million election-related posts shared on Twitter in 2016 by over a million distinct users. This dataset includes accounts associated with the identified Russian trolls as well as users sharing posts in the same time period on a variety of topics around the 2016 elections. To study how these trolls attempted to manipulate public opinion, we identified 49 theoretically grounded linguistic markers of deception and measured their use by troll and non-troll accounts. Finally, we show that deceptive language cues can help to accurately identify trolls, with average F1 score of 82% and recall 88%. Most of the content in this chapter is going to be published in ICWSM conference 2019 under the title "*Linguistic Cues to Deception: Identifying Political Trolls on Social Media*". It is a joint work with Emilio Ferrara, Kristina Lerman, and Adam Badawy.

6.2 Identifying Political Trolls on Social Media

According to Pew Research Center (Gottfried & Shearer, 2016), two-thirds of Americans get their news from social media. However, even as social media has become a vital source of information for many, it has also become a source of misinformation, hoaxes, and fake news. This is because, unlike traditional news outlets, social media platforms provide little in the way of individual accountability or fact-checking. Misinformation, including conspiracy theories, hoaxes, and rumors, propagate on social media just as readily as factual information. For example, a study showed that when the Ebola crisis broke out in 2014, lies, half-truths, and rumors spread as quickly as accurate information on the Twitter social media platform (Jin et al., 2014).

An oft-cited example that shows how misinformation can affect real-world events is the 2013 hacking of the Associated Press Twitter account. Using the compromised account, hackers tweeted that Barack Obama had been injured in an explosion at the White House. The tweet triggered a drop that wiped out 130 billion dollars in stock value in a matter of seconds (Matthews, 2013). This issue becomes more prominent when the topic of discussion is related to a highly controversial issue, such as politics, since online users are being exposed to more political content written by ordinary people than ever before. Bakshy et al. (2015) report that 13% of posts by Facebook users who report their political ideology are political news. Moreover, these posts may not be even generated by humans. Troll accounts and social bots for example, have attempted to manipulate the 2016 U.S. presidential elections by injecting false tweets, or "fake news", in support of or against certain candidates (Pennycook & Rand, 2018). (Allcott & Gentzkow, 2017; Guess, Nyhan, & Reifler, 2018; Lazer et al., 2017; Pennycook & Rand, 2018; Shao et al., 2018; Zannettou et al., 2018).

This deceptive, made-up content was shared with millions of Americans, both on Twitter and Facebook, before the 2016 election. In this study, we study the language used by Russian trolls during Russia's campaign to interfere in the 2016 US presidential election. Trolls are user accounts whose sole purpose is to sow conflict and deception. In the context of the 2016 elections, their intent is to harm the political process and create distrust in the political system. These trolls were allegedly funded by the Russian government to influence conversations about political issues, with the goal of creating discord and hate among different groups (T. P. Gerber & Zavisca, 2016). Stanley Renshon notes that deception in U.S. presidential politics has become more pervasive over the past several decades (Borenstein, 2016).

However, the topic of automatic detection of deceptive information has not been widely studied until recently. Our study addresses this gap with an empirical study of deceptive language used by Russian trolls in their attempts to influence U.S. elections. This may lead to better tools to detect misinformation in the Twitter sphere produced by fake accounts.

6.2.1 Contributions of this work

The focus of our ongoing research is to understand the effects of trolls' interference in the U.S. election. To do so, we plan to answer the following questions:

- How do trolls insert themselves into political discussions on Twitter? What topics do they discuss?
- What deceptive linguistic cues do trolls rely upon to generate tweets?
- Can we automatically detect troll accounts using these deceptive linguistic cues?

The goal of these questions is to understand how these agents camouflage themselves among U.S. Twitter users in order to be more appealing to them. We use the markers of deceptive language to measure how deceptive trolls' tweets are compared to legitimate users. Since deception generally entails messages and information knowingly transmitted to create a false conclusion (Buller, Burgoon, Daly, & Wiemann, 1994), it stands to reason that trolls use deceptive language to mislead others into believing the information they share. In social media, people tend to be truth-biased on assessing messages they receive (Levine, Park, & McCornack, 1999). Because of that, the accuracy of human detection of deception remains little better than chance (Frank & Feeley, 2003). There is compelling evidence from prior deception research that a variety of language features, either spoken or written, can be valid indicators of deceit (Buller & Burgoon, 1996; Burgoon, Buller, Guerrero, Afifi, & Feldman, 1996). One example of the psychological side effects of deception is the observation that people manage the discomfort caused by lying by distancing themselves from the deceptive message they created (DePaulo et al., 2003). Psychological distancing was found to manifest itself through a decrease in self-reference (e.g., “I,” “me,” “myself”) and an increase in group reference (e.g., “they”, “he”), which are strategies that indicate a lack of commitment toward the deceptive statement (DePaulo et al., 2003; Hancock, Curry, Goorha, & Woodworth, 2007). These pronouns become effective linguistic markers of deceptive language.

6.3 Features of Deception

We identify deception as misleading the audience via a piece of information. Deceptive information includes but is not limited to lies, fake news, and rumors disseminated to change peoples’ cognition or beliefs (Rubin, 2017). Social media that focus primarily on content are highly susceptible to deception, since most communication is text-based and done asynchronously (Tsikerdekis & Zeadally, 2014).

A growing body of research suggests that we can learn a great deal about people’s underlying thoughts, emotions, and motives by counting and categorizing the words they use to communicate, where the communication can be verbal or written. Several studies on deception detection have demonstrated the effectiveness of linguistic cue identification, as the language of truth-tellers is known to differ from that of deceivers---see (Larcker & Zakolyukina, 2012).

Prior work has examined deceptive language in several domains, including fake reviews (S. Feng, Banerjee, & Choi, 2012; Ott, Choi, Cardie, & Hancock, 2011), online games (Zhou, Burgoon, Nunamaker, & Twitchell, 2004), online dating profiles (Toma & Hancock, 2012), interview dialogues (Levitan, Maredia, & Hirschberg, 2018), and opinions on controversial topics (Mihalcea & Strapparava, 2009). However, deception detection in social media has not been studied yet since the type of communication is different from interviews and emails.

Even though there is no clear consensus on reliable predictors of deceptive language, prior work has identified several deceptive cues that can be identified in text, extracted and constructed conceptually, to represent several categories, such as complexity, specificity, and non-immediacy. Ott et al. (2011) compared approaches to automatically detecting deceptive opinion spam using a crowd-sourced dataset of fake hotel reviews. Other research has collected deceptive data by asking subjects to write or record deceptive and truthful opinions about controversial topics such as the death penalty or abortion, or about a person that they like or dislike (Mihalcea & Strapparava, 2009). (Zhou et al., 2004) consider computer-mediated deception in role-playing games designed to be played over instant messaging and e-mail.

Literature on linguistic analysis of deception suggests that changes in word quantity, pronouns, emotional terms, and distinction markers may reflect deception (Burgoon, Blair, Qin, & Nunamaker, 2003; DePaulo et al., 2003). Linguistic features such as n-grams and language complexity have been analyzed as cues to deception (Pérez-Rosas, Kleinberg, Lefevre, & Mihalcea, 2017; Yancheva & Rudzicz, 2013). Moreover, expressing emotions, specifically negative ones, has been shown to be linked to deception (Burgoon et al., 2003; Zhou et al., 2004). Syntactic features such as part of speech tags have also been found to be useful for structured data (S. Feng et al., 2012; Ott et al., 2011).

Building on previous research on deception detection using language, new ways to analyze such data have emerged, such as developing software that can automate the detection of linguistic cues. One of the best-known software platforms used for text-based deception detection is Linguistic Inquiry and Word Count (LIWC) (Pennebaker & King, 1999), which groups words into psychologically motivated categories. The main idea of LIWC coding is text classification according to truth conditions. LIWC has been extensively employed to study deception detection (Hancock et al., 2007; Mihalcea & Strapparava, 2009; Vrij, 2000).

6.4 Deception Detection Methods

When deception detection is implemented with standard classification algorithms such as decision trees and logistic regression, it achieves an accuracy of 74% (Fuller, Biros, & Wilson, 2009). When using existing psycholinguistic lexicons as LIWC for detecting deceptive opinions, the accuracy of the classifier achieves an average accuracy rate of 70% (Mihalcea & Strapparava, 2009). By comparison, human judges only achieve a 50-63% success rate in identifying deception (Rubin & Conroy, 2011).

Researchers have proposed various methods to automatically detect deception. (Rubin, Chen, & Conroy, 2015) divided fake news identification into three types: “fabrication, hoaxing and satire detection.” Their research not only promotes a more nuanced view of fake news, but also suggests different methods to detect each type; e.g., network analysis for hoaxes and binary text classification for satire detection. (Chen, Conroy, & Rubin, 2015) proposed a hybrid approach to automatically detect clickbait. They believe that using contextual cues such as lexical, semantic, and pragmatic analysis in addition to various classification algorithms - such as Naïve Bayes - as well as non-contextual cues, such as image and user behavior analysis, are important factors that may assist researchers in distinguishing clickbait headlines from legitimate ones.

6.5 Data Collection

6.5.1 Trolls

To collect Twitter data on Russian trolls, we used a list of 2,752 Russian troll accounts compiled and released by the U.S. Congress[22]. After that, we collected all of the trolls' discussions. To collect the tweets, we used Crimson Hexagon[23], a social media analytic platform that provides paid data stream access. This tool allowed us to obtain tweets and retweets produced by trolls and subsequently deleted in 2016. We were interested in understanding troll activity during the election year. We collected data starting from 2015.

- ***The Year 2015***

[22] https://www.recode.net/2017/11/2/16598312/russia-twitter-trump-twitter-deactivated-handle-list

[23] https://www.crimsonhexagon.com/

During 2015, trolls posted 1,128,615 tweets, 43.7% of them written in Russian. Seventy four percent of the posts have an identifiable location, 58.30% of which were from the U.S., while 31.32% were from Russia. As we can see in Figure 22 below, these accounts started a campaign against Ukrainian nuclear power plants at the beginning of January. Hashtags used in this campaign include #FukushimaAgain and #Chernobyl2015. We can observe another attack on Ukrainian President Petro Poroshenko on January 20th, when the hashtag "#SomeoneWhoKillsChildren" was trending. This hashtag is related to the war between Russia and Ukraine in Donbass. There is a large spike in activity on March 18th. A closer look at these posts reveals them to be mostly random tweets talking about love and life with multiple borrowed quotes. Examples include:

- "Only put off until tomorrow what you are willing to die having left undone." – Pablo Picasso
- "I wasnt poor, i was po', i couldnt afford the or" – Big L

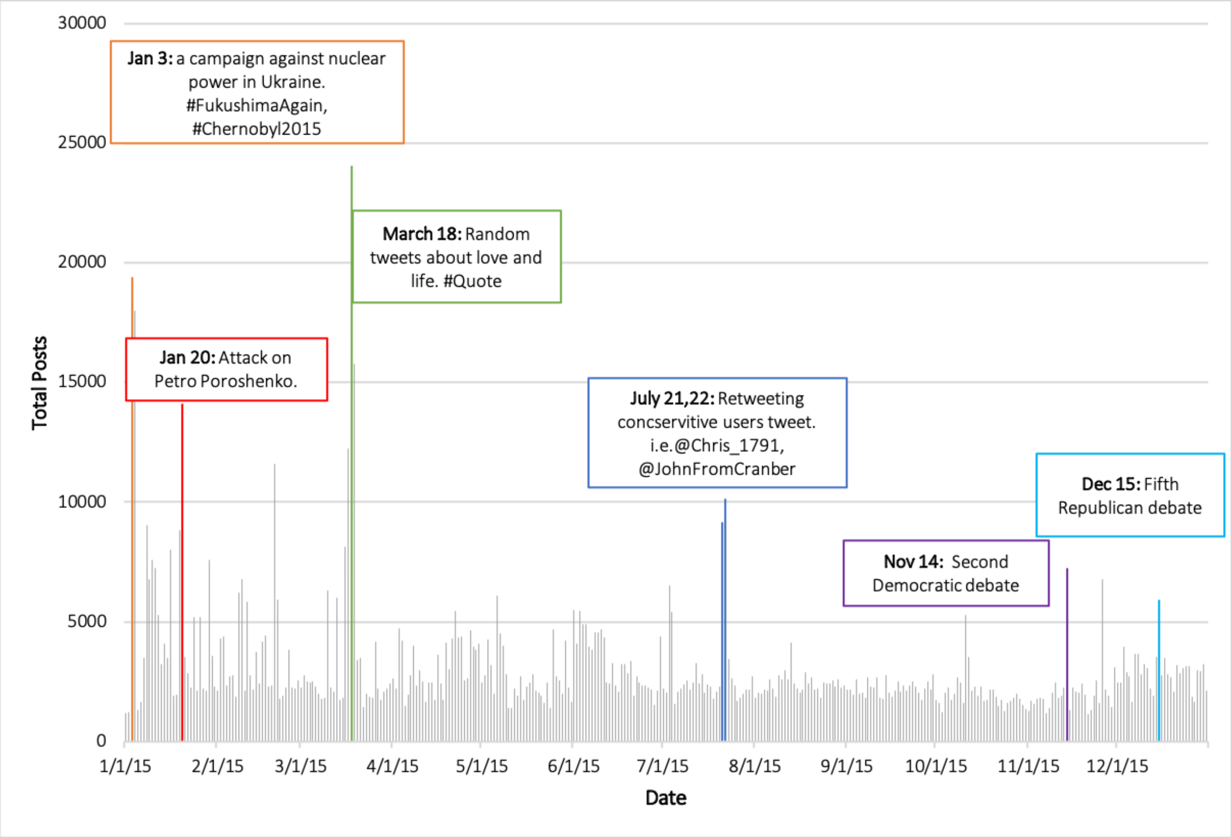


Figure 22: Daily volume of troll tweets in 2015. Upticks in activity are highlighted in the plot.

Trolls were already active in 2015, posting over a million tweets, 44% of them in Russian, with 31% of the posts with an identifiable location coming from Russia. These accounts were actively demonizing Ukrainian President Petro Poroshenko and campaigning against Ukrainian nuclear power plants. Late in the year, the accounts started tweeting about U.S. elections, talking about debates between Republican and Democratic presidential candidates.

- ***The Year 2016***

In 2016, the 1,148 trolls posted 1,226,185 tweets, of which 27% were written in Russian. Over 90% of the tweets had identifiable locations, with 65% from the U.S., 27% from Russia, and 2% from Belarus. As we can see below in figure 23, troll activity increased in the months leading to the elections, with spikes in activity related to external events. Interestingly, the biggest spike of activity was on October 6th. The tweets were mainly pro-Trump, although no specific topics are discernible. The next day, the Access Hollywood tape was released, which showed Trump using derogatory and sexist language. The timing of the spike is curious. Was it meant to serve as a distraction from the tape?

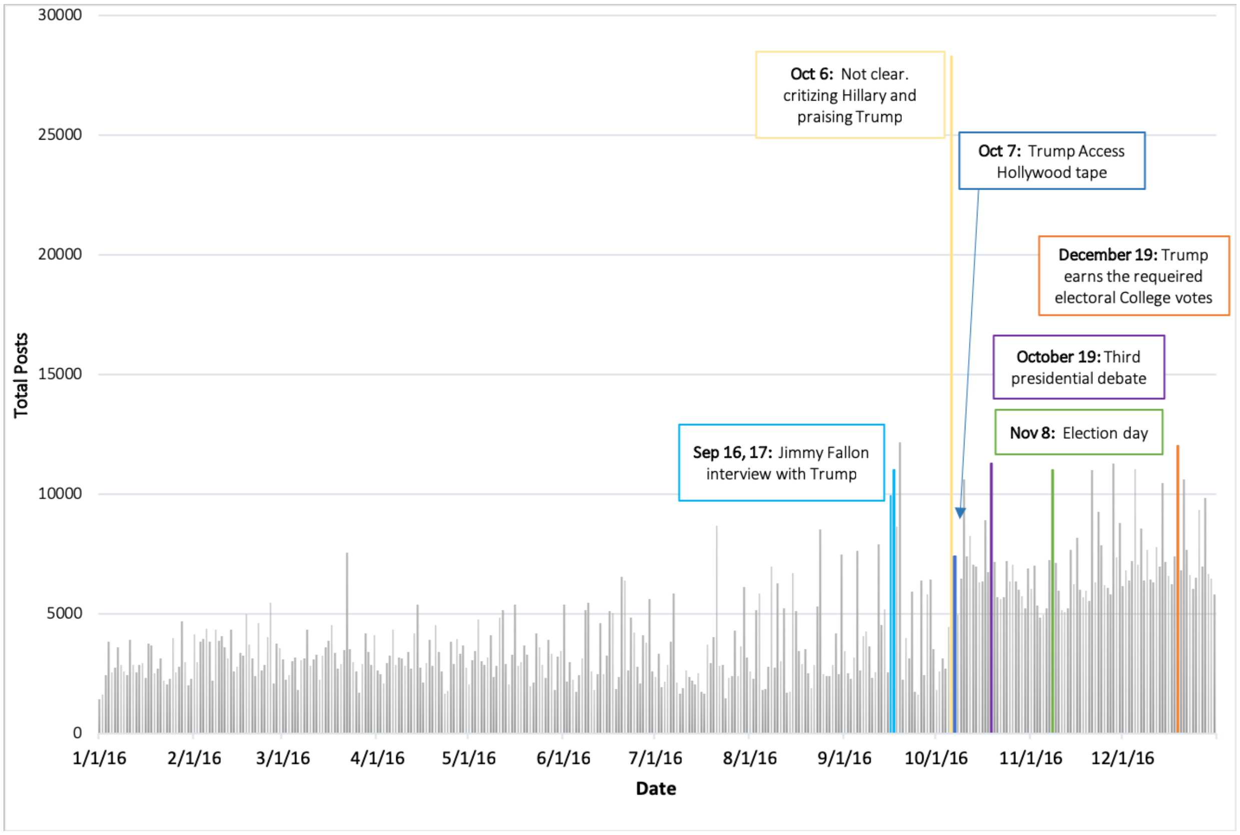


Figure 23: Daily volume of troll tweets in 2016. Upticks in activity are labeled with dominant topics.

Table 23 presents descriptive statistics of the troll accounts. They tweeted over one million times, with 688,019 retweets. 1,148 of these trolls (42%) exist in our dataset, with 1,032 (~90%) of them producing original tweets.

Table 23: Descriptive statistics of the troll dataset

Trolls	Number
# Unique Russian trolls in the data	1,148
# Tweets	1,226,155
# Retweets by trolls	688,019
# Original tweets by trolls	538,136
# Trolls who posted original tweets	1,032

6.5.2 Non-Trolls

To collect non-troll tweets, we use two strategies. First, we collect such tweets using a list of hashtags and keywords that relate to the 2016 U.S. Presidential election. This list is crafted to contain a roughly equal number of hashtags and keywords associated with each major Presidential candidate: we select 23 terms, including five terms referring to the Republican Party nominee Donald J. Trump (#donaldtrump, #trump2016, #neverhillary, #trumppence16, #trump), four terms for Democratic Party nominee Hillary Clinton (#hillaryclinton, #imwithher, #nevertrump, #hillary), and several terms related to debates. To make sure our query list was comprehensive, we add a few keywords for the two third-party candidates, including the Libertarian Party nominee Gary Johnson (one term), and Green Party nominee Jill Stein (two terms). Our second strategy is to collect tweets from the same users that do not include the same key terms mentioned above and making sure that we exclude any users who have re-tweeted a troll.

Users who did not retweet a troll may help with shaping a better understanding of troll behaviours online. Our collection yielded a total of 12,361,285 tweets produced by 1,166,760 unique users. Table 24 shows descriptive statistics of non-troll accounts.

Table 24: Descriptive statistics of the non-troll dataset.

Non-Trolls	Number
# unique non-trolls	1,166,760
# tweets by non-trolls	12,361,285
# retweets by non-trolls	9,868,403
# original tweets by non-trolls	2,492,882
# of non-trolls who posted original tweets	140,062

6.6 Deceptive Language

We conjecture that political trolls use deception to deliberately mislead others about their true intention. The hypothesis we carry in this study is that trolls deceive other online users to believe in a certain agenda. Where deception can be defined as a deliberate act with the intent to mislead others while the recipients are not made aware or expect that such an act is taking place. Moreover, the goal of the deceiver is to transfer that false belief to the deceived ones.

Deception has an emotional and cognitive cost to the deceiver, which can often emerge through the language used to deceive. Studies examined physiological responses of the deceiver utilizing behavioral coding with well-trained experts, or applying content-based criteria to written transcripts for deception detection (Zhou et al., 2004). After that, automated linguistic techniques were developed to analyze the linguistic properties of texts to examine the linguistic profiles of deceptive language (Newman, Pennebaker, Berry, & Richards, 2003) (Zhou et al., 2004).

Deceptive (and truthful) language has been studied through different approaches (Shuy, 1998) based on theoretical assumptions of how deception should be reflected in language. Interpersonal Deception Theory (IDT) explains deception in interpersonal contexts (Buller & Burgoon, 1996). While not developed for online text, it provides a theoretical and evidentiary foundation for the cues in our study. Verbal immediacy theory (VI) was proposed to infer people's attitude or affect. The general construct of immediacy refers to verbal and nonverbal cues that create a psychological sense of closeness or distance (Zhou et al., 2004).

Criteria-based content analysis (CBCA) was developed to determine the credibility of child witness' testimonies in trials for sexual offenses and recently applied to assess testimonies by adults (Raskin & Esplin, 1991). It holds that a statement derived from memory of an actual experience is different in content and quality from a statement based on fantasy (Steller & Koehnken, 1989; Undeutsch, 1989). A similar theory, reality monitoring (RM), was designed to study memory. It holds that a truthful memory differs in quality from remembering a made-up event (Johnson & Raye, 1998). Previous research has used this framework extensively to distinguish truth from lies (Bond & Lee, 2005).

Twitter messages lack facial expressions, gestures, and conventions of body posture and distance, so text itself is the only source for us to infer personal opinions and attitudes and verify message credibility. Moreover, previous work has identified deception as a characteristic that can be measured through verbal cues (Tsikerdekis & Zeadally, 2014).

Lately, automated linguistic techniques in which computer programs are used to analyze the linguistic properties of text have been used to examine the linguistic profiles of deceptive language---see (Bond & Lee, 2005; Newman et al., 2003; Zhou et al., 2004).

Linguistic cue dictionaries are borrowed from different sources. The first is the Multiple Perspective Question Answering (MPQA) opinion corpus developed by University of Pittsburgh. This lexicon includes patterns to account for the various ways in which speakers argue. Lexicon entries are in the form of regular expression patterns. The second is Linguistic Inquiry and Word Count (LIWC) (Tausczik & Pennebaker, 2010). LIWC is a text analysis program that computes features consisting of normalized word counts for 93 semantic classes. LIWC dimensions have been used in many studies to predict outcomes including personality (Pennebaker & King, 1999), deception (Newman et al., 2003), and health (Pennebaker, Mayne, & Francis, 1997). LIWC produces the percentage of each variable type by dividing the frequency of the observed variable by the total number of words in the sample, with the exception of word count, words per sentence, and question marks, which are reported frequencies. All features computed for the users are normalized by the number of tweets each posted except for LIWC features since they are computed as percentages. Building on this research, we identified 49 linguistic cues as potential markers of deceptive language. We used specialized lexicons designed to operationalize language-based measures. Below we justify our choice of each measure as a potential deception marker.

- **Uncertainty.** Based on IDT theory, deceivers tend to use less structured and more evasive language. %Being certain means not having doubts, which in contrast, truth-tellers tend to be more certain about their statements.

  Linguistic markers of certainty, such as "always" or "never," are strong indicators of truthfulness (Levitan et al., 2018; Rubin, Liddy, & Kando, 2006). Prior research has shown that subjective language can help recognize certainty in textual information (Rubin et al., 2006). Deceivers express greater uncertainty by using more modifiers and model verbs in their text than truth tellers (Buller & Burgoon, 1996; Zhou et al., 2004). The increased use of hedges has been linked to more uncertainty (Levitan et al., 2018; Rubin et al., 2006).

- **Modifier.** Is a word, phrase, or sentence element that limits or qualifies the sense of another word, phrase, or element in the same construction[24]. Inspired by previous research, we use a list of modifier words borrowed from MPQA. We count occurrences of each modal word in each user's list of tweets and follow the same technique for other lexicon-based measures.
- **Modality** is an expression of an individual's "subjective attitude" (Bybee, Perkins, & Pagliuca, 1994) and "psychological stance" (Mitra, Wright, & Gilbert, 2017) towards a proposition or claim. Words as "should" and "sure" denote assertion of a claim, while "possibly" and "may" express speculation. Modality can be identified as an auxiliary verb that is characteristically used with a verb of predication and expresses necessity or possibility[25]. We measure modality expressed in the text by using a list of necessity and possibility words borrowed from MPQA.
- **Subjectivity.** Is an aspect of language used to express opinions and evaluations (Banfield, 1982; Wiebe, 2000). Since being certain can be identified as being objective, we hypothesized that subjectivity can provide meaningful signals for deception detection and used OpinionFinder's subjectivity lexicon comprising 8,222 words (Wilson et al., 2005).
- **Quotations.** Serve as a reliable indicator for accuracy, where quoted content is correlated with being uncertain about its content (De Marneffe, Manning, & Potts, 2012). We hypothesize that trolls use more quoted content in their tweets. We compute this measure by counting the number of quotations present in a user's tweets.
- **Questions**. Based on IDT's interactivity principle, deceivers attempt to increase the interactivity of the communication in an effort to increase believability. Thus, in such interactions, deceivers are expected to ask more questions. Previous work has showed that deceivers use more questions during their discussions (Hancock et al., 2007). Hence, we include questions, measured as question marks in each user's tweets, as a potential indicator of deception.
- **Hedges.** Are words that express lack of commitment to the truth value of a claim, reveal skepticism, caution, or display an open mind about a proposition. Previous research has shown

[24] https://www.dictionary.com/browse/modifier
[25] www.webster.com

that deceptive speech contains more hedges (Tausczik & Pennebaker, 2010). We included hedges as potential deception markers in tweets. To measure hedges, we used a curated set of hedging cues from (Hyland, 1998, 2018).

- **Non-immediacy.** Following IV theory, being non-immediate is related to being deceptive. Deceivers tend to acquire more avoidance strategies. For example, "you and I worked" is equivalent to "we worked" in meaning; however, the former is more non-immediate than the latter. Moreover, IDT theory describes non-immediacy as a method of dissociation where deceivers may use language to distance themselves from the content of their messages. Non-immediacy can be measured through lack of self-reference, group reference, and generalization.
- **Self-reference.** Measured through first person singular pronouns (i.e., "I", "me", or "my"), is one of the ways deceivers can express non-immediacy. Theoretical and empirical observations suggest that deceivers attempt to distance themselves from their deception and not take ownership of a statement by using fewer first-person singular pronouns (Hancock et al., 2007; Newman et al., 2003; Toma & Hancock, 2012; Zhou et al., 2004).
- **Group reference**. Is measured by using third-person pronouns (i.e., "they", "she"). Research suggests that liars are less likely to use third-person pronouns in their deceptive interactions than in truthful ones (Newman et al., 2003). In contrast, (Zhou et al., 2004) showed that deceptive senders used more group reference compared to truthful senders. This is a strategy to distance themselves from the deceptive message they created (Ickes, Reidhead, & Patterson, 1986). This feature is obtained from LIWC.
- **Generalization**. Refers to a person (or object) as a class that includes the person (or object). Hypothesizing that a non-immediate and more general narrative can be associated with higher deception, we employed MPQA's list of generalization words to incorporate features corresponding to these language markers.
- **Indefinite articles.** Another way to be general is the usage of indefinite articles like "a", "the", and "an", which signal an upcoming noun (Tausczik & Pennebaker, 2010). Indefinite articles are more likely to refer to general concepts than definite articles since they suggest concreteness (Danescu-Niculescu-Mizil, Cheng, Kleinberg, & Lee, 2012). To measure indefinite articles, we used LIWC's list of articles.

- **Specificity.** Based on IDT, RM and CBCA theories, being specific in describing an event or a situation has been proven to relate to truthfulness. Previous research has shown that deceivers are less specific in their text (Burgoon et al., 2003). Being specific includes the usage of discourse markers, causation cues, emotional words, and sense terms.

- **Discourse Markers.** Liars may be particularly wary of using discourse markers that delimit what is in their story and what is not (Newman et al., 2003). Exclusion words, conjunctions, and negations are discourse markers that require a deceiver to be more specific and precise when communicating their messages. We hypothesize that trolls use fewer discourse markers compared to non-trolls. We employed LIWC's list of exclusion, negation, and conjunction words to incorporate features corresponding to these language markers.

- **Conjunction.** A conjunction is a word like "and", "but", etc., which is used to join two ideas together into a complex sentence. They are useful for creating a coherent narrative and require a deceiver to be specific and precise in his language. Hypothesizing that a coherent narrative cannot be associated with deceivers, we employed LIWC's list of "conjunction" words to incorporate features corresponding to these language markers.

- **Negation.** Previous research has showed that liars will produce fewer negation words during deceptive discussions compared to truthful ones (Hancock et al., 2007) (Toma & Hancock, 2012); this includes words like "no", "not", and "never". Also, research has identified that legitimate content includes more negation (Pérez-Rosas et al., 2017). For that, we hypothesize that trolls use fewer negation words, using LIWC's list of negation words.

- **Exclusion words.** such as "rather", "but", and "however" are useful in determining if something belongs to a category (Tausczik & Pennebaker, 2010), while distinctions between alternative concepts are indicative of greater cognitive complexity (Newman et al., 2003). Previous research has shown that increased usage of exclusion words means more truthful information (Hancock et al., 2007; Newman et al., 2003).Using exclusion words in text can communicate ambiguity or equivocation, which is a characteristic of deceivers (Bavelas, Black, Chovil, & Mullett, 1990).

- **Causation.** is another linguistic marker similar to distinction markers, since it adds specificity and detail to a story and increases the possibility of self-contradiction. Causation words include "because", "effect", and "hence". Previous research has showed that deceivers use fewer causation terms when lying (Hancock et al., 2007). We hypothesize that trolls use fewer causation words. We used LIWC and MPQA's list of causation words.

- **Emotions**. One strategy to avoid being specific is to express more emotions. Previous works have %investigated how deceivers express emotions in their communications. In general, they found that deceivers tend to use more emotional language compared to truth tellers (Burgoon et al., 2003; Zhou et al., 2004). Fake content uses more positive words (Pérez-Rosas et al., 2017) and deceivers use negative emotion words (Newman et al., 2003). To measure the extent of emotions expressed in tweets, we used LIWC's comprehensive list of positive and negative emotion words.
- **Sense Terms.** like "see", "touch", and "listen" are used to add more details and specifics to narrative. Previous research has suggested that providing such sensory details may be more difficult for a person who is fabricating an opinion or a memory (Johnson & Raye, 1998; Vrij, 2000). Other studies have confirmed that deceivers are more likely to use words that pertain to the senses when lying (Hancock et al., 2007). We employ LIWC's list of sense terms.
- **Use of numbers.** Mentions of numbers is commonly used as a marker of specificity (J. J. Li & Nenkova, 2015). Since deceivers tend to be less specific, we hypothesize that trolls use fewer numbers in their text.
- **Relativity.** is a linguistic marker available in LIWC, which includes words related to motion, space, and time (i.e., "before") Previous work identified that legitimate content expresses more relativity (Pérez-Rosas et al., 2017).

6.6.1 Information complexity

Based on CBCA and RM theories, deceivers' language describing an imagined event may fail to reflect the rich diversity of an actual event, where higher sentence complexity results in lower perception of deception (Briscoe, Appling, & Hayes, 2014). Moreover, deceivers display less lexical and content diversity (Zhou et al., 2004). Information complexity is measured by average word length, sentence length, words that have more than six letters, and the amount of punctuation. We used the LIWC to produce the count of words per sentence, words with six letters, and the amount of punctuation. We calculated the average length of a user's set of tweets by summing all the tweets and normalizing by the total tweet count.

6.6.2 Information Quantity

Deceivers may be more hesitant and less forthcoming then truth-tellers and express their hesitancy by using fewer words and sentences. Previous research found deceivers' messages in text-based chats were briefer (Burgoon et al., 2003). We hypothesize that trolls use less information than non-trolls, where information quantity is measured by the number of words, verbs, adverbs, nouns, and prepositions. We use LIWC and NLTK to tokenize tweets and calculate these features.

6.6.3 Persuasion

Persuasion involves convincing a target to accept a message. We hypothesize that deceivers attempt to provide persuasive and credible statements to redirect the listener's attention from any false information.

- **URLs.** The sharing of URLs is a persuasive act that can contribute to a sophisticated and persuasive writing style. Previous research showed that persuasive arguments consistently use more links (Khazaei, Lu, & Mercer, 2017; Tan, Niculae, Danescu-Niculescu-Mizil, & Lee, 2016). Citing external evidence online is often accomplished using hyperlinks. We use the number of links used in each post and whether or not the links are featured.
- **Function words.** have little lexical or ambiguous meaning and express grammatical relationships among other words within a sentence, or specify the attitude or mood of the speaker[26]. The use of function words in communication reveals deep aspects of the communicators such as his/her honesty and sense of self (Pennebaker, 2011). Previous research has shown that persuasive comments include fewer function words (Khazaei et al., 2017). We hypothesize that trolls use fewer function words. To calculate this feature, we used LIWC's list of function words.
- **Examples.** We recorded the normalized number of any mentions of the phrases “for example”, “for instance”, “e.g.” and their synonyms in each tweet based on the notion that

[26] https://en.wikipedia.org/wiki/Function_word

providing illustrations and further explanations is another component of persuasive language, as has been shown in previous research (Tan et al., 2016).

- **Present Focus.** Linguistic cues that are used to talk about the present and the future such as "today", "is", and "now" are commonly used in non-persuasive comments (Xiao, 2018). We used LIWC to get a list of present tense words.
- **Reward.** Words such as "take", "prize", and "benefit" that reference rewards, incentives, and positive goals appear regularly in non-persuasive comments (Xiao, 2018). We hypothesize that troll tweets are less reward-focused then non-troll tweets. We used LIWC to identify the list of reward-focused words.
- **Number of Hashtags.** Previous research has shown that hashtags can serve as useful signals of rumors (Castillo et al., 2011). We include the hashtag count of tweets as a potential persuasive marker.

6.6.4 Morality

Moral foundation theory (Haidt & Graham, 2007) describes moral differences across cultures. This theory holds that there is a small number of basic moral values, and people differ in how they endorse these values. Moral foundations include care and harm, fairness and cheating, loyalty and betrayal, authority and subversion, and purity and degradation. We hypothesize that deceptive tweets contain fewer moral linguistic cues than non-deceptive tweets. We measure morality using the list of moral foundation words (Haidt & Graham, 2007).

6.6.5 Metadata

Metadata features obtained from Twitter API include the number of followers, the number of followees, total tweet count, user status count, and number of retweets. No previous work linked the predictability of such features with deception. However, we hypothesize that such features could be an indicator of deceptiveness. Previous research has showed that troll accounts usually have fewer followers and more followees (Badawy, Ferrara, & Lerman, 2018).

6.7 What Topics Do Trolls Discuss?

To have a better understanding of the trolls and their activity, we studied the top hashtags, words, and mentions used in both troll and non-troll posts. Trolls use generic hashtags, such as #news, #politics and #sports, which allows their content to be more widely viewed. Thus, when a user search for "#news" he is exposed to troll tweets. Another interesting insight is that trolls choose controversial topics that many Twitter users are discussing, such as the Black Lives Matter movement. This also makes them appear to be Americans who care about U.S. civil movements. While trolls mention Hillary Clinton with the "neverhillary" hashtag, non-trolls utilize the hashtag "imwithher" more frequently. Based on the top words used in both troll and non-troll tweets, we can get a sense of what topics these two user groups are discussing. In table 25, We see that trolls discuss recent issues in American society, such as school shootings. In contrast, non-trolls discuss the leaked “Access Hollywood tape”[27]. In table 26, we show the top 15 hashtags in trolls and non-trolls discussions.

Table 25: Top 10 meaningful words from the tweets of trolls and non-trolls.

Trolls	Count	Non-trolls	Count
trump	24476	Trump	304074
police	16952	Photo	243684
man	14382	Following	100697
black	11270	Salute	97841
year	10484	Hillary	85323
people	10391	Clinton	82507
Clinton	9914	Donald	76187
woman	9828	Sex	68416
State	9314	Video	64199
Hillary	7824	Johnson	63816
Killed	7746	Alert	60773
Shooting	7156	Woman	56031

[27] https://en.wikipedia.org/wiki/Donald_Trump_Access_Hollywood_tape

Table 26: Top 15 hashtags in trolls and non-trolls discussions.

Trolls	Count	Non-Trolls	Count
news	77,741	#communityscene	77,059
politics	28,699	#trump	56,927
#world	21,730	#pjnet	43,293
#sports	20,412	#imwithher	41,017
#tcot	13,873	#tcot	38,759
#blacklivesmatter	13,574	#events	35,478
#pjnet	10,759	#newszbreakin	31,624
#local	9,906	#nevertrump	30,703
#topnews	9,783	#topnews	28,940
#business	7,868	#windows10	24,603
#health	7,641	#freekaavan	20,799
#ccot	5,987	#lds	19,681

6.8 Do Trolls Use Deceptive Language?

To study whether trolls use deceptive language, we compare linguistic markers of deception in troll and non-troll tweets. For each linguistic dimension, we conduct a two-tailed t-test over the troll and non-troll datasets to verify the significance of differences for the mean between the two groups. Some linguistic dimensions are positively correlated deception; i.e., if a text contains more of that linguistic dimension, it is more likely to be deceptive. We show in Figures 24 and 25 the log means values for deception markers.

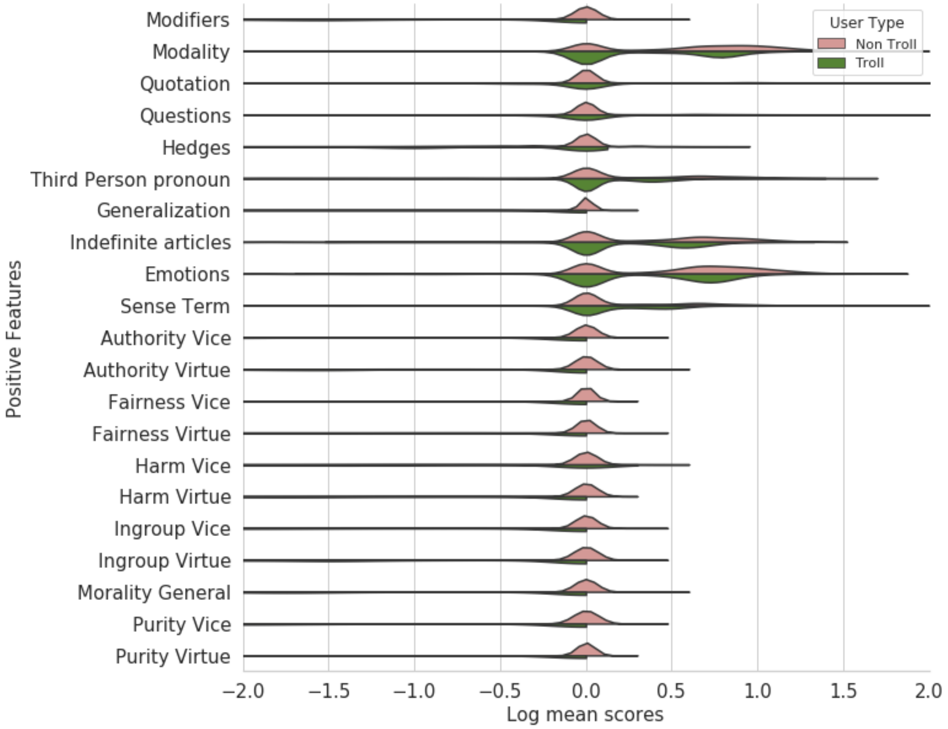


Figure 24: Log mean values for features with positive correlation with deception

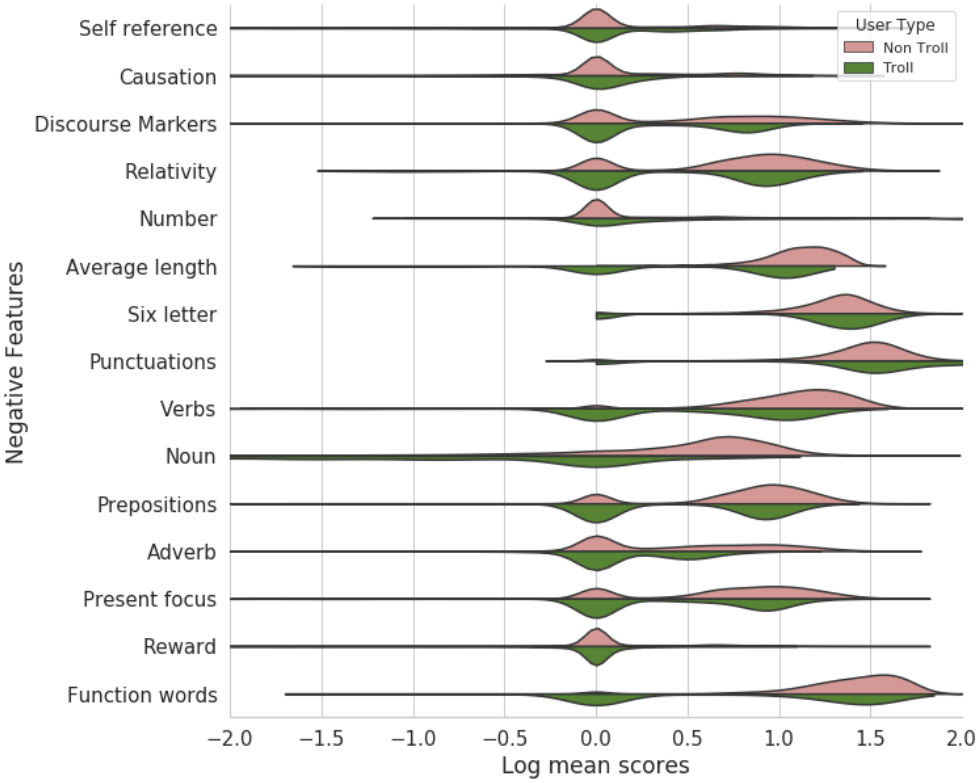


Figure 25: Log mean values for features with negative correlation with deception.

For metadata features and descriptive features such as hashtag count, URL count, etc., we show the differences between their log mean values in figure 26 below. Figure 26 shows that trolls have significantly fewer followers and more tweets and retweets than non-trolls. This finding echo finding from prior work (Badawy et al., 2018). Moreover, trolls use significantly more URLs and hashtags in their tweets, while non-trolls have more tweets and status counts. Figure 24 and figure 25 show the different linguistic measures in troll vs. non-troll tweets for features with positive and negative correlation with deception, respectively. Below we discuss the potential of linguistic measures described in the method section as markers of deception.

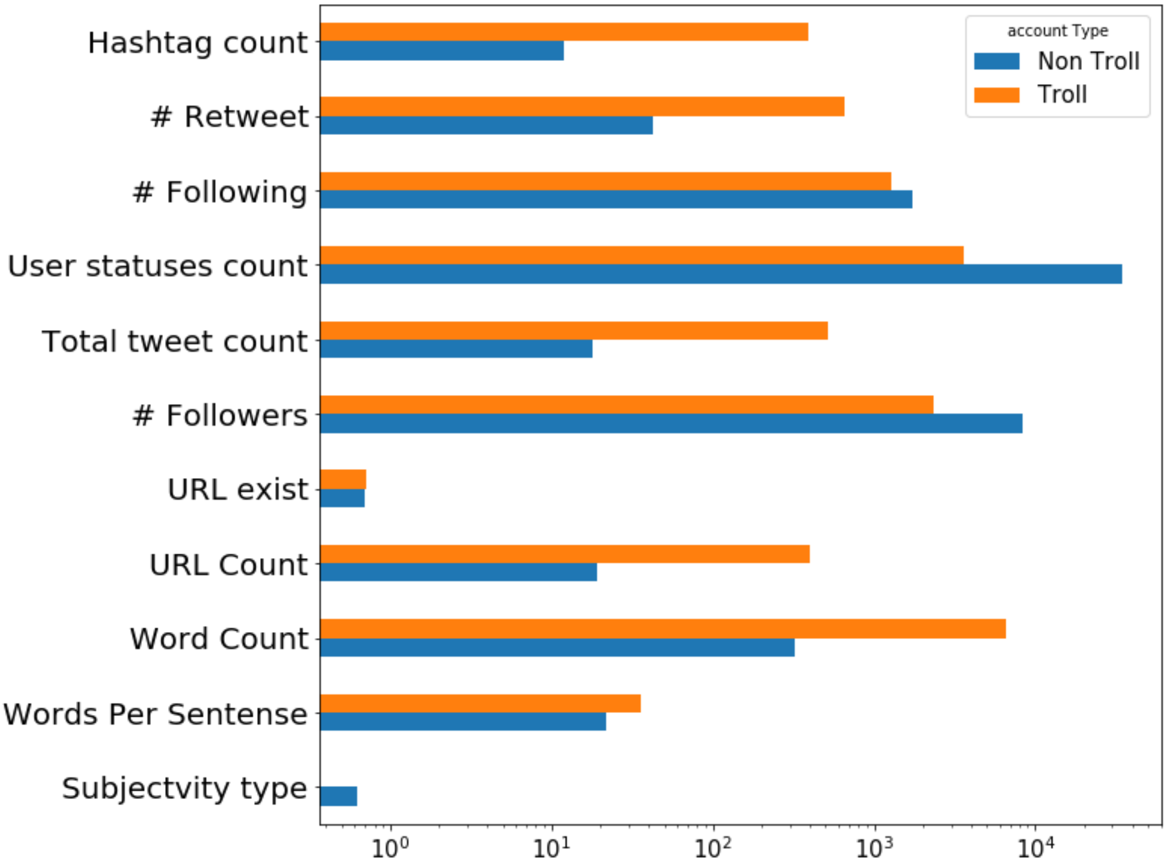


Figure 26: Log mean values for the difference between trolls and non-trolls in descriptive features

6.8.1 Uncertainty

Uncertainty was linked to deception. We hypothesized that trolls will use language that introduces uncertainty, such as modifiers, model verbs, etc. However, our results show that trolls use significantly fewer modifiers, model verbs, and hedges than non-trolls, which contradicted our hypothesis. On the other hand, other linguistic cues of uncertainty, such as the use of quotations and questions, was significantly higher in trolls compared to non-trolls. Moreover, trolls use less subjective language compared to non-trolls. Since subjectivity is used to express opinions and evaluations (Banfield, 1982; Wiebe, 2000), this implies that trolls are less certain, which leads to more deception.

6.8.2 Non-immediacy

Deceivers tend to use linguistic cues that indicate avoidance, including self-reference, group reference, and generalization. Our results show that trolls refer to themselves and others significantly less then non-trolls. This supports previous research that indicates that deceivers use less self and group reference to distance themselves from others (Newman et al., 2003; Zhou et al., 2004) (Hancock et al., 2007; Toma & Hancock, 2012). On the contrary, trolls use significantly fewer general terms and indefinite articles compared to non-trolls, which contradicts our hypothesis that they use more general narrative to distance themselves from the deception.

6.8.3 Specificity

Research suggests that liars may be wary of using discourse markers, which can delimit what is in their story and what is not (Newman et al., 2003). Our results matched previous research and show that trolls use significantly fewer discourse markers. Similarly, the usage of causation words adds specificity and details to a story and increase the possibility of self-contradiction. We found that trolls tend to use fewer causation words like "because" and fewer sense terms.

Moreover, we found that trolls tend to write with significantly less emotion compared to non-trolls. This contradicts previous work that found that deceivers tend to express more emotional language (Burgoon et al., 2003; Zhou et al., 2004). Another indicator of specificity is relativity words; we show that trolls tend to use fewer relativity words, confirming previous work (Pérez-Rosas et al., 2017).

6.8.4 Information Complexity

We find that trolls have less complex, shorter tweets, compared to non-trolls and less complex words (with fewer than six letters). However, they use significantly more words per sentence and more punctuation compared to non-trolls.

6.8.5 Information Quantity

We hypothesized that trolls use fewer words and sentences to express their hesitancy. Trolls composed tweets with significantly fewer nouns, verbs, adverbs, and prepositions, which confirms research on deception (Burgoon et al., 2003). However, trolls used significantly more words in total compared to non-trolls. Even though trolls have higher word count compared to non-trolls, these words are not important parts of speech, such as nouns and verbs.

6.8.6 Persuasion

Trolls used highly persuasive linguistic cues. For example, the use of links in text has been shown to be part of persuasive arguments (Khazaei et al., 2017; Tan et al., 2016), and we have found that trolls used significantly more URLs in their tweets compared to non-trolls. Moreover, trolls use fewer function words, which was also confirmed by previous work (Khazaei et al., 2017). Furthermore, trolls use significantly fewer present-focused words compared to non-trolls, where the use of present tense has been confirmed to be part of non-persuasive comments (Xiao, 2018). When tweets are less reward-oriented, they are considered more persuasive (Xiao, 2018). In our data, trolls use significantly fewer reward-focused words compared to non-trolls. Trolls used significantly more hashtags, which confirmed our hypothesis that persuasive tweets contain more hashtags than non-persuasive ones. The results *confirm* our hypothesis that trolls use persuasive language as a way to deceive.

6.8.7 Morality

We found that trolls show significantly fewer moral values compared to non-trolls. This confirms the hypothesis that using fewer moral cues in the text might imply that the user is trying to be deceptive.

6.9 Can Trolls be Identified?

Identifying trolls is a considerable challenge given their small number. The resulting classification task is highly unbalanced, and a trivial algorithm marking every account as non-troll will have high accuracy, but low recall. However, for the same reason, if a model were to trivially predict that no user is a troll, the model would be accurate most of the time. Provided that we want to predict trolls, this model would not be very useful in practice. In other words, our setting is a typical machine-learning example of a highly unbalanced prediction task.

To test our ability to detect trolls and to see which features are most important in distinguishing between trolls and non-trolls, we leverage two classifiers and multiple models. The first model serves as a baseline, with each model including progressively more variables. We use two off-the-shelf machine learning algorithms: Random Forest (RF) and Gradient Boosting Classifier (GBM) and train classifiers using Stratified 10-fold cross-validation with the following preprocessing steps: (i) replace all categorical missing values with the most frequent value in the column, and (ii) replace missing values with the mean of the column. We tune the GBM classifier to have a learning rate of 0.1, 500 trees, and max depth of 3 for each tree. To deal with severe imbalance between the majority labels (non-trolls) and minority labels (trolls), we use the Synthetic Minority Over-Sampling Technique + Edited Nearest Neighbor Rule (SMOTE-ENN) (Batista, Prati, & Monard, 2004) to over-sample from the minority label and under sample from the majority label, to keep a ratio of 1:5 trolls-to-non-trolls in every training fold. For GBM, the better performing classifier, we obtain an average F1-score 0.82 and average recall of 0.88 for 10-fold cross validation. For RF, we obtain 0.8 for both the average F1-score and the average recall for the 10-fold. GBM F1-scores across the 10-folds have a smaller variance than the RF scores. Thus, GBM does not only offer a better average F1-score, but the lower variance between folds shows that it is a more stable model to use.

6.9.1 Feature Importance

To better what features contribute to the accurate identification of trolls, we look at the feature importance plot of Gradient Boosting for the full model. The Variable Importance by Category plot (cf., Figure 20) provides a list of the categories of variables in descending order by a mean decrease in the Gini criterion.

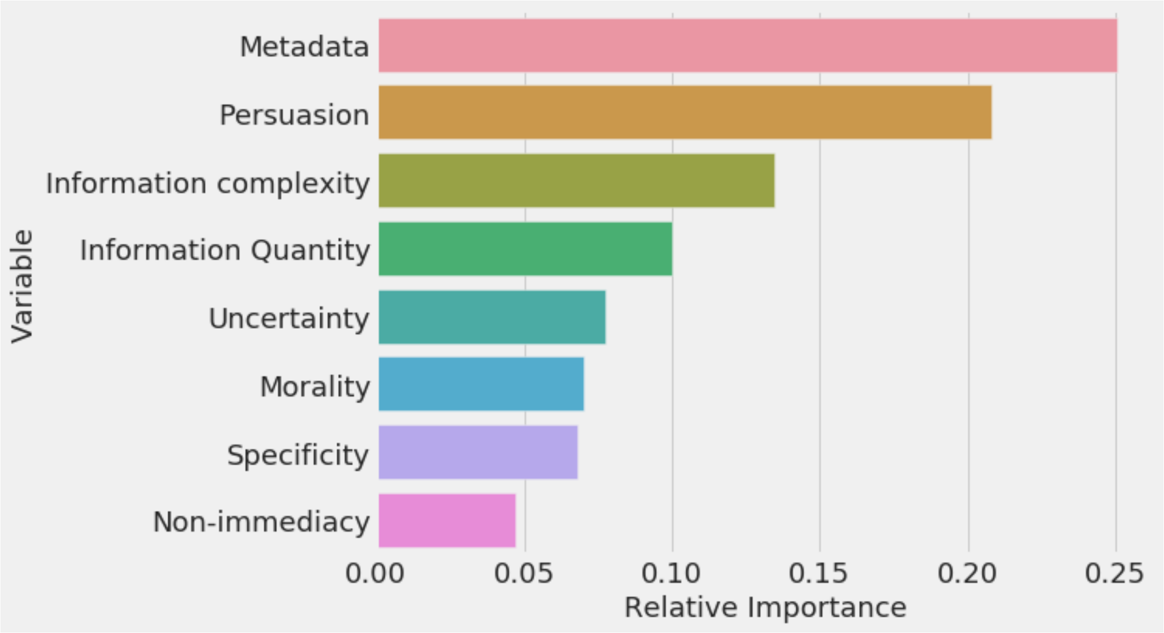


Figure 27: Relative importance of the feature categories using Gradient Boosting for the full model

As shown in Figure 27 the top category variables contribute more to the model than the bottom ones and can discriminate better between trolls and non-trolls. In other words, features are ranked based on their predictive power according to the model. Figure 28 shows the relative importance of the features using Gradient Boosting for the full model (best performing fold) in predicting users who are trolls.

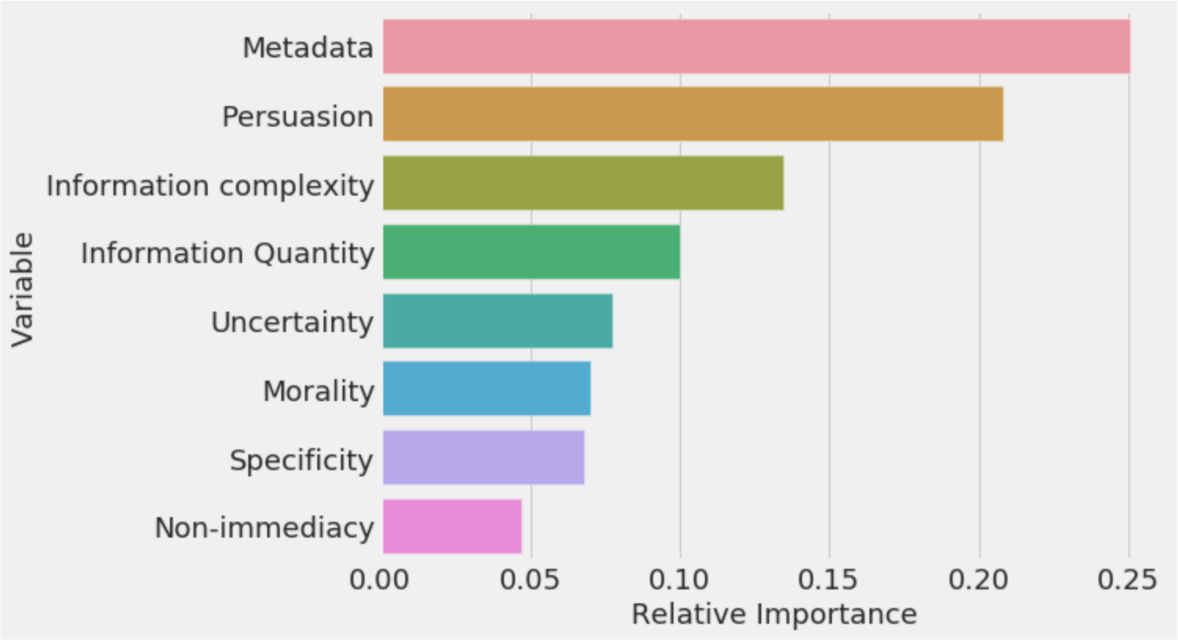
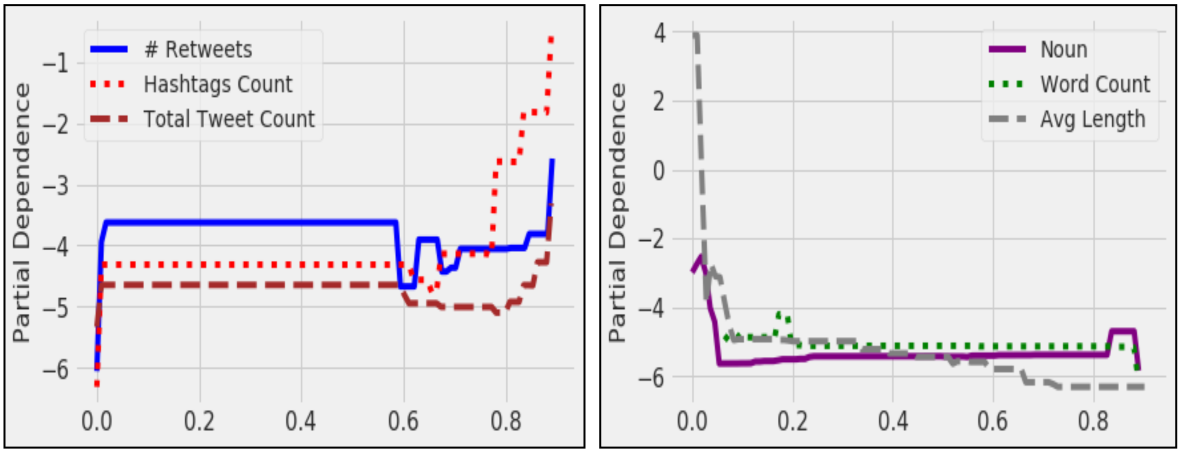


Figure 28: Relative importance of the features using Gradient Boosting for the full model (best performing fold) in predicting users who are trolls.

Figure 28 illustrated the top 20 features in descending order of their importance to contributing to the prediction of trolls. According to the full model and the best GBM fold classifier, the number of hashtags, retweets, and tweets as well as the number of nouns and average length of users' tweets are the most predictive feature of whether users are trolls. Deception markers, including self-reference and hedges, round out the top features.

Using Partial Dependence plots, we show that the classification outcome has positive relationships with the following features: # of retweets and overall tweet counts, as well as the number of hashtags and URLs. Figure 29 visualizes these relationships with the y-axis showing its magnitude and the x-axis the distribution of the feature under examination. Figure 29 suggests moving from left to right that the number of hashtags used increases the probability of being a troll, particularly toward the end of the distribution; higher number of total tweets and retweet counts are also associated with higher likelihood of being a troll, particularly toward the end, while being flat for most of the distribution. On the other hand, we can see that the outcome has a negative relationship with the number of nouns used, word count, and average tweet length, as shown in Figure 29. This means that having fewer nouns and posting shorter tweets with fewer words are characteristics associated with higher probability of being a troll.



(a) Downward Trends

(b) Upward Trends

Figure 29: Partial Dependence plots for some of the top features considered in the full model (best preforming fold). These partial dependence plots are for the Gradient Boosting Classifier fitted to the balanced dataset. Each plot shows the dependence of the outcome variable (troll/non-troll) on the feature under consideration, marginalizing over the values of all other features (Note: x-axis values are CDF-normalized).

6.10 Conclusion

In this study, we addressed the issue of understanding Russian troll activity on Twitter in 2016. Specifically, we identified linguistic markers of deception that could be good predictors when identifying such trolls. Based on these linguistic markers, we addressed the task of automatic identification of trolls. By developing a theory-driven model working on millions of tweets of Russian trolls and legitimate users, we unfold ways to identify trolls using social media text signals of deception.

Our results showed that Russian troll accounts that discussed the 2016 U.S. election used deceptive language to influence public opinion and spread biased political information on social media. The theory-driven linguistic analysis was able to capture features of the deceptive language. For example, we found that troll accounts use significantly more persuasive language cues and less complex and specific language. We used these language cues to build a classifier that was able to identify trolls with high accuracy (average F1 score is 82% and recall is 88%). While metadata features were quite distinctive and predictive of trolls, several linguistic features were also predictive of troll accounts, particularly features related to information complexity and persuasion. We show that higher numbers of hashtags, tweets, and retweets are associated with higher likelihood of being a troll, as well as fewer usage of nouns and posting shorter tweets with fewer words.

Our work has several limitations. First, not all trolls who are identified as trolls were active in 2016, so having a full picture of all trolls' activity might be harder to achieve. Second, we lack sufficient information on how the troll list was compiled in the first place. This might be an issue, since the methodology taken to identify these trolls could include certain biases that might affect our conclusions. Third, users who are identified as non-trolls might actually be bot accounts not identified in the list. Lastly, our model might be limited by missing potential confounding variables. Despite all of these limitations, the identification of such malicious actors who are mainly responsible of the spread of misinformation is extremely important. Although the data suggest important overall differences in deceptive linguistic patterns across trolls and non-trolls, not all linguistic variables changed as a function of deception. Further investigation into discourse markers that can identify trolls is needed. Moreover, this work can be extended by including higher level interaction terms, such as syntactic constructions and discourse relations.

CHAPTER 7: *TweetChecker:* A PLATFORM TO EVALUATE CONTROVERSIAL HEALTH EVENTS IN REAL-TIME

7.1 Introduction

When online users search for information regarding health controversial issues, the presence of spam, rumors and fake content is unpreventable, which in turn reduces the value of information contained in the tweets. This chapter provides actionable, scalable and data-driven insights about controversial issues to online users and policy makers. As explained before, checking the validity of information available in Twitter is challenging, especially for controversial issues. The chapter contributes in building up real time Twitter streaming analytical pipeline from scratch using amazon AWS. The platform for evaluating controversial health events in real-time can be used effectively by security analysts, organizations, and professional agencies that want to monitor content of its interest on Twitter.

7.1.1 Platform Benefits

The implementation of such a platform can be beneficial in many ways. Firstly, the data analysis can be improved by integrating different types of analysis. Second, the increased storage that is offered by the cloud, compared to storing all of the data in a limited local machine. Third, the simplified Infrastructure that is provided in the cloud, there is no need to worry about the different components implemented and how they interact with each other since all of this now is employed in the cloud. Fourth, the cost of implementing and storing is going to be lower when using the platform since nothing is stored/ implemented in the local machine. Finally, the increased privacy and security of having all of the data stored remotely.

7.1.2 The Importance of The Platform

The tool used can answer these questions:

- How many tweets in twitter are talking about this topic?
- How many bot accounts are talking about this topic?
- What are the different types of URLs shared in the tweet?

- How much can the accurate information be found (produce %)?

7.1.3 Health Controversial Issues

The platform for *TweetChecker* is used only for health controversial topics. Health-related issues are particularly likely to become controversial topics that are plagued by poor-quality information. As the number of people seeking health information online continues to increase (Miller & Bell, 2012), the need for credible information has become more important. Previous studies have shown that in 2010, 80% of Internet users looked online for information about health topics, such as specific diseases or treatments (Fox, 2011). Moreover, previous research has stated that many patients seek and follow advice from medical websites rather than visiting doctors due to the amount of health information that is available online (Gualtieri, 2009). For this reason, ensuring that online health-related information is credible is important. Studies on the quality of online healthcare information have found that it is not always reliable (Morahan-Martin & Anderson, 2000). In a systematic meta-analysis of health website evaluations, 70% of studies concluded that quality is a problem on the Internet (Eysenbach, Powell, Kuss, & Sa, 2002), with some information being low quality, biased, misleading, or incorrect (Eysenbach, 2003).

People may make decisions based on advice found online, which may affect their health negatively when few sites contain sufficient information to support people's decision-making. Instead, many sites contain misinformation or are filled with jargon (Smart & Burling, 2001). For example, people searching for "Issels Treatment" will find a website that describes it as a "comprehensive immunotherapy for cancer"; however, the American Cancer Society considers this type of therapy to be unproven and possibly harmful (Dori-Hacohen & Allan, 2013). In this tool, the user can choose to see results out of five different health topics, the selection of these topics is explained in the next section.

7.2 Controversial Health Issues Selection

Data for this study was developed using a cross-validation approach in order to objectively and systematically represent American public opinion. In this study, I will focus only upon the US population since some controversial issues may not be as prevalent as in other countries. A two-step search of legal and journalistic publications and then medical publications was used to determine controversial health-related issues and the most common opposing viewpoints adopted regarding them. A survey was then administered to determine public opinion toward each issue.

7.2.1 Identifying Controversial topics

Identifying controversy is a difficult task. One can just simply say that a health issue is controversial or not based on intuition. However, it was needed to develop a more complete and better validated sense of American public opinion on which health-related issues are controversial. In order to accomplish this goal and to quantify public opinion about which issues are considered controversial using a survey, it was needed to first to identify controversial topics objectively and systematically. To select the topics, a three-step mixed methods/cross-validation approach was used as represented in figure 30.

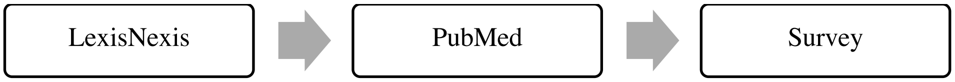


Figure 30: Topic selection approach

As shown in Figure 30 about topic selection approach, the three step mixed methods are LexisNexis, PubMed, and Survey. The steps are mentioned below in detail:

Step 1: LexisNexis

Figure 30 shows that the first step involved examining LexisNexis[28], a pioneer in electronic accessibility to legal and journalistic documents that provides easy access to a wide range of news articles examining current issues of public interest. In order to discover controversial and non-controversial issues, a bootstrapping approach seeded with a combination of keywords describing controversy (i.e. “controversial”, “controversy”) was used.

A search criteria setting was used to conduct a search of major U.S. publications only, discarding non-English publications. It was limited the search results to those publications

[28] https://www.lexisnexis.com/

published between January 1, 2016 and March 31, 2017 in order to find recent controversies exclusively, a combination of these keywords was used:

- "controversial," "controversy"
- "disagree," "disagreement," "debate"
- "well-being," "health"

After conducting the search, the following terms rose to the surface (any issue not related to health was ignored): *Abortion, AIDS, E-cigarettes, GMO, Marijuana, Vaccines.*
The frequency of the number of articles for each of these issues was recorded in order to capture its volume. Table 27 shows the frequency of each issue found in the time frame specified.

Table 27: Article Frequency for Each Controversial Issue Discovered

Controversial issue	Article frequency
Vaccines	876
Marijuana	875
AIDS	874
Abortion	851
E-cigarettes	693
GMO	477

In order to expand our understanding of the selected issues, each issue was used as a supplementary keyword in the initial search query. This approach helped determining which aspects of each issue are controversial when seen from the perspective of the public.

Step 2: PubMed

In the second step, PubMed[29] was used, which is a search engine for accessing the MEDLINE database of references and abstracts on life sciences and biomedical topics. In PubMed searches, MeSH terms (Medical Subject Headings)—the NLM-controlled vocabulary thesaurus for indexing articles— was used as part of the queries in order to expand the understanding of these issues and to construct better statements for the next step. The list of search query used are listed in table 28[30].

[29] https://www.ncbi.nlm.nih.gov/pubmed/
[30] the search was done in April 2017

Table 28: PubMed search query and the number of results found

Search query	Search results
“organisms, genetically modified”[MeSH Terms] OR (“organisms”[All Fields] AND “genetically”[All Fields] AND “modified”[All Fields]) OR “genetically modified organisms”[All Fields] OR (“genetically”[All Fields] AND “modified”[All Fields] AND “organism”[All Fields]) OR “genetically modified organism”[All Fields]	50,667
“abortion, induced”[MeSH Terms] OR (“abortion”[All Fields] AND “induced”[All Fields]) OR “induced abortion”[All Fields] OR “abortion”[All Fields]	80,990
(“measles-mumps-rubella vaccine”[MeSH Terms] OR (“measles-mumps-rubella”[All Fields] AND “vaccine”[All Fields]) OR “measles-mumps-rubella vaccine”[All Fields] OR (“mmr”[All Fields] AND “vaccine”[All Fields]) OR “mmr vaccine”[All Fields]) AND (“autistic disorder”[MeSH Terms] OR (“autistic”[All Fields] AND “disorder”[All Fields]) OR “autistic disorder”[All Fields] OR “autism”[All Fields])	463
“medical marijuana”[MeSH Terms] OR (“medical”[All Fields] AND “marijuana”[All Fields]) OR “medical marijuana”[All Fields] OR (“medical”[All Fields] AND “cannabis”[All Fields]) OR “medical cannabis”[All Fields]	5,210
(“hiv”[All Fields] AND “aids”[All Fields]) OR “hiv aids”[All Fields]	138,079
“electronic cigarettes”[MeSH Terms] OR (“electronic”[All Fields] AND “cigarettes”[All Fields]) OR “electronic cigarettes”[All Fields] OR “e cigarette”[All Fields]	3,001

Additionally, the use of other polarized MeSH terms associated with the issue-related MeSH term was investigated. As an example, when searching for the term “genetically modified organism (GMO),” it was found that two other terms co-occurred with it: “risks” and “benefits.” Another example is when “abortion” was searched, the terms “medical abortion,” “unsafe abortion,” “abortion ethics” co-occurred, among others. These co-occurring terms suggested that the discussion of each issue incorporated polarized judgements in the scientific literature, indicating that the issue is controversial or debatable. Therefore, these parts of each issue were selected:

“Abortion should be legal”

Two basic points of view were obvious from the news stream: anti-abortion and pro-abortion opinions. Also, many newspaper articles included opinions about abortion law.

“MMR vaccine causes autism”

Many articles discussed President Trump’s stated opinion in favor of the theory that the MMR vaccine causes autism and his support of anti-vaccine groups. Other streams of articles discussed how the public and pro-vaccine groups reacted to this announcement.

“Medical marijuana should be legal”

Some news articles discussed marijuana (cannabis) as a medicine that should be legalized.

"HIV causes AIDS"

Many news articles discussed the debate over whether or not HIV causes AIDS.

"E-cigarette is better than smoking"

Two sides of opinion were obvious from the news stream. One promoted e-cigarettes, arguing that they might help consumers give up smoking, while the other encouraged banning them.

"Genetically modified organisms (GMOs) are safe"

Two sides of opinion were present in the news stream. One endorsed the idea that GMOs are safe to use in producing food, while others promoted the risks associated with them.

The above statements were then validated with a native speaker of American English to confirm their meaning and interpretations. After that, the top five issues and the final set of statements chosen to be used in the survey are:

- *"Medical marijuana should be legalized"*
- *"The Measles, Mumps, and Rubella (MMR) vaccine can cause autism"*
- *"Legal abortion performed by a qualified medical professional should be a foundational right for women"*
- *"Human Immunodeficiency Virus (HIV) causes Acquired Immunodeficiency Syndrome (AIDS)"*
- *"E-cigarettes are less harmful than smoking tobacco"*

Step 3: Survey

The third and last step of the topic selection process was conducting a survey to measure the controversiality of each topic. To do that, two questions where conducted to accurately measure users' perceptions towards the controversiality of the statements provided. The first question was asking users to rate how controversial they believe each statement was, which was independent of whether they agree with the statement or not. They were asked to rate each statement on a 5-point Likert scale (from 5 = "extremely controversial" to 1 = "not at all controversial"). The second question asked about their level of agreement with each of the statements. (The full list of questions appears in Appendix C).

The survey was distributed using two main outlets. The first survey was run on Amazon's Mechanical Turk service (MTurk), an online crowdsourcing system. MTurk participants were compensated with $0.48 USD per survey which complies with the US federal minimum wage, $7.25/hr.[31]. The survey was available only to U.S. residents with at least a 95% approval rating (a screening option that MTurk provides) with number of HIT approval being above 500 HITs. These screening options allows for only recruiting qualified MTurkers. A total of 95 surveys were received from MTurkers, and 88 of them were considered valid. A response was considered invalid if the evaluation question was not answered probably. The validation question was asking respondent to answer with a specific answer. This question was used to make sure that users are actually reading the questions and not clicking randomly. See figure 32 in Appendix D for the HIT posted in MTurk.

The second outlet was to run the survey on social media, namely Facebook, Twitter and Reddit. Participants were not paid for their contribution due to the need to preserve their anonymity. For Reddit, two main subreddit was used UIUC[32] and samplesize[33]. See figure 33 and 34 in Appendix D for messages posted in these subreddits). As regards social media, Facebook ads and Twitter ads were used. For both these services, I paid a total of $40 for advertising campaign, which involved creating advertisements that appeared on the pages of the target audience meeting the criteria of location (United States), and language (English). A charge was incurred every time a user clicked on the ad. Figure 35 and 36 in Appendix D shows the advertisement used in both Facebook and Twitter.

This researcher has limited resources to continue paying for advertising, for that, another strategy was implemented for using social media for participants' recruitments, which included utilizing Facebook public pages that discusses health topics. Five random pages were chosen based on the number of members (>1000), being in the USA and the description of the page state that it consumes topics related to health. These pages were "Women's Health & Fitness Tips[34]",

[31] https://www.dol.gov/whd/minimumwage.htm
[32] https://www.reddit.com/r/UIUC/
[33] https://www.reddit.com/r/SampleSize/
[34] https://www.facebook.com/groups/AllAboutWomens/

"USA Trusted Dating & Health Product [35]", "Health Food & Weightloss Tips [36]", "Better Body Health Circle [37]", "TruVision Health Testimonies [38]".

In these surveys, the main validation strategy was to remove empty or incomplete responses besides checking the validation question. For that, empty responses and responses that did not contain complete answers were eliminated. A total of 73 responses were received, and out of those 65 were complete. For the two surveys together, the total number of valid responses was 153. The surveys were collected over a period of 17 days in October/November 2017. Completing the survey took participants 7 minutes on average.

Using the first question, the average controversy rating was calculated. Using the second question, the standard deviation (SD) was measured to quantify the spread of users' opinions towered each statement. Higher SD was assumed to mean higher controversiality. It implies that users did not have a uniform opinion regarding the issue, which make it more controversial based on the controversiality definition. Table 29 show the results of these measurements. We can see that some topics have high values in both measurements, which indicate high controversiality. For example, the topic of abortion has an average controversy rating of 4.44 and a standard deviation of 1.29. These results indicate that most participants agree that the topic of abortion is considered highly controversial and their opinions do not uniformly cluster towards one opinion (see Table 29).

Table 29: Topics and their controversiality measures

Topic	Statements	Average controversy rating (M)	Standard deviation (SD)
AIDS	*"Human Immunodeficiency Virus (HIV) causes Acquired Immunodeficiency Syndrome (AIDS)"*	1.78	0.77
E-cigarette	*"E-cigarettes are less harmful than smoking tobacco"*	3.13	0.87
Vaccine	*"The Measles, Mumps, and Rubella (MMR) vaccine can cause Autism"*	3.52	0.86
Marijuana	*"Medical marijuana should be legalized"*	3.65	0.95
Abortion	*"Legal abortion performed by a qualified medical professional should be a foundational right for women"*	4.44	1.29

[35] https://www.facebook.com/groups/Datinginformation/

[36] https://www.facebook.com/groups/572900242906760/

[37] https://www.facebook.com/groups/betterbodycircle/

[38] https://www.facebook.com/groups/TheOfficialTruVisionTestimonies/

The results in table 29 shows that *abortion, marijuana, vaccine, e-cigarette and AIDS* are controversial based on the average controversy rating. When considering the standard deviation, we can see that some topics has different values. For example, For the topic of AIDS, it has less spread, where it seems that people support the idea that HIV causes AIDS with most people surveyed (114) think it is "extremely likely" when they are asked the question of *"How likely is it that HIV (Human Immunodeficiency Virus) causes AIDS (Acquired Immunodeficiency Syndrome)?"*.

7.2.2 Building Search Queries

In order to build the search query for each topic, three strategies were used. First, Twitter was searched using a combination of keywords in each statement. For example, as regards the vaccine issue, the word "vaccine" was plugged in Twitter to capture any other words that were correlated with it, such as "shot", or similar medical words or slang words used in social media. Second, any hashtags associated with the discussions of these health issues was extracted using our original keywords; this strategy helped us to identify different aspects of each topic that users discussed. The final chosen queries are shown in Table 30.

Table 30: Twitter Search Query for each Topic

Topic	Twitter query
Abortion	("abortion" OR "reproductive choice" OR "ProChoice" OR "pro-choice" OR "pro choice" OR "PraytoEndAbortion" OR "ProLife" OR "UnbornLivesMatter") AND ("policy" OR "legal" OR "legalized" OR "right" OR "legalization") AND NOT "RT"
Vaccine	("vaccinations" OR "vaccination" OR "vaccines" OR "vaccine" OR "measles-mumps-rubella" OR "MMR" OR "mmr") AND ("autism" OR "autistic disorder") AND NOT "RT"
AIDS	("HIV" OR "human immunodeficiency virus") AND ("AIDS" OR "acquired immunodeficiency syndrome") AND NOT "RT"
E-cigarette	("e-cigarette" OR "e cigarette" OR "vaping" OR "ecig" OR "e cig" OR "vapingsaveslives" OR "vaping saves lives" OR "electronic ecigarette") AND ("smoking" OR "cigarette" OR "tobacco") AND NOT "RT"
Marijuana	("Marjuana" OR "weed" OR "420" OR "pot" OR"cannabinoid" OR "cannabis" OR "marijuana") AND ("Medical" OR "synthetic" OR "medicalmarijuana" OR "MedicalCannabis") AND ("policy" OR "legal" OR "legalized" OR "right" OR "legalization") AND NOT "RT"

7.3 Implementation and Building the Platform

In order to encourage many users to interact with TweetChecker, we provided it in an easy way to use, as a web-based application (*www.thetweetchecker.com*). The implementation includes a back-end and a front-end, which was built in Amazon AWS. Figure 31 shows the basic architecture of the platform.

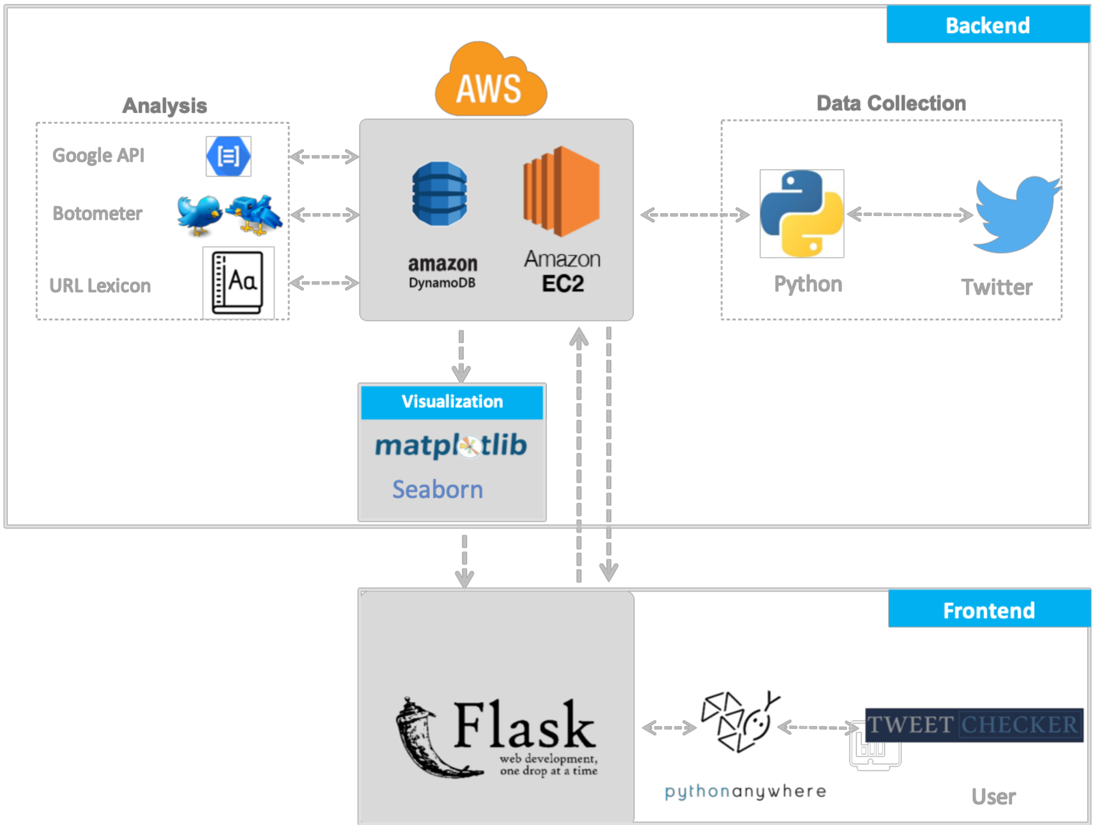


Figure 31: Platform structure

7.3.1 Back-end

The backend of the platform contains several steps.

Step 1: Data Streaming

As explained before, the analysis will be on five different health controversial issues already selected. The user can select which of these issues he/she would like to see results from. After that, Twitter streaming API pushes tweets as it happens in near real-time based on the set of keywords already established. Twitter streaming API provides only a sample of tweets that are occurring. Studies have estimated that using Twitter's Streaming API users can expect to receive anywhere from 1% of the tweets to over 40% of tweets in near real-time.

Step 2: Streaming Pipeline

Tweepy is an open-sourced, easy to use python library to access Twitter API. Tweepy supports accessing Twitter via OAuth, which is the methodology used by Twitter API to authenticate developers. The authentication keys needed to be requested before from dev.twitter.com. With Tweepy, it is possible to get any object and use any method that the official Twitter API offers. One of the main Tweepy methods is streamingAPI, which gives the capability of monitoring for tweets and doing actions when some event happens. Key component of that is the StreamListener object, which monitors tweets in real time and catches them.

Step 3: Data Storage

After collecting the tweets, they need to be stored for further analysis. The data is stored in a database offered by amazon.com. DynamoDB is a fully managed database that supports a key-value structure. The full tweets and all attached information are stored in the database. This includes the user name, country, number of followers and follows, tweet text, tweet number of likes and favorite count.

Step 4: Analysis

After data is stored in the database, the next step is to analyze the data. In this step, we do two types of analysis, bot analysis and URL analysis.

- **Bot Analysis**

The user who shares the information in social media is one of the factors in the spread of misinformation. Some of these users are not actually real users. Trolls accounts and social bots are designed to attempt to manipulate the public opinion towards different topics. One of the ways to detect bot accounts in Twitter, is to use an openly accessible solution called Botometer (a.k.a. BotOrNot) (Davis et al., 2016), consisting of both a public Web site (https://botometer.iuni.iu.edu/) and a Python API (https://github.com/IUNetSci/botometer-python), which allows for making accurate determination about the user account. Botometer is a machine learning framework that extracts and analyses a set of over one thousand features, including features as user network, the content and networkstructure, temporal and sentiment features. Typically, Botometer returns likelihood scores above 50 percent only for accounts that look suspicious to a scrupulous analysis. We adopted the Python Botometer API to systematically inspect the most active users in our dataset. The Python Botometer API queries the Twitter API to extract 300 recent tweets and publicly available account metadata, and feeds these features to an ensemble of machine learning classifiers, which produce a bot score. To label accounts as bots, we use the fifty percent threshold – which has proven effective in prior studies (Davis et al., 2016: an account is considered to be a bot if the overall Botometer score is above 0.5).

- **URL analysis**

Several studies have shown that incorporating URLs into social media texts is a main feature of credibility because the inclusion of a URL makes people more likely to believe the content of the post (Castillo et al., 2011; Kinsella et al., 2011). In the case of Twitter, others have shown that a tweet gets more retweets and shares if it has more links in its content (Tan et al., 2016). For that, we included the analysis of the type of the URL as one of the features we extract.

To identify the type of the URL, a lexicon built before, also shared in chapter 4, was used (Addawood et al.). First, the URLs are pulled from each tweet and are classified based on the created lexicon that contains 12 different categories. After that, this analysis is saved in a new table in the database to prepare them for the next step, visualization.

- **Linguistic analysis**

The language of the tweet can provide some insights into the user's opinions towards the issue discussed. by utilizing the collective opinions of online users, we can have a better understanding of user opinions towards these issues. One of the main approaches to identify opinions from text is sentiment analysis. Sentiment analysis is or Opinion Mining is the computational study of people's opinions, attitudes and emotions toward an entity. The entity can represent individuals, events or topics.

To apply this approach to the tweets text, Google Cloud Natural Language API [39] is used where it provides powerful machine learning APIs. These APIs extract entities from text, perform sentiment and syntactic analysis, and classify text into categories. The API was used mainly for sentiment analysis, where each tweet text is passed to the API and then it returns the input text's sentiment score on a scale from -1 to 1, where a score of -1 is very negative, 0 is neutral, and 1 is very positive. The sentiment is also given a magnitude score on a scale from 0 to infinity, which indicates the intensity of the emotion expressed. Sentiment scores are returned for individual sentences within a text, as well as for the document as a whole.

Step 5: Visualizations

After being done with the analysis and storing the results in new tables, the system will start showing the user the values of these results in more appealing way. These visualizations provide the user with a quick interface and big-picture about the analysis. For visualizing the results, the researcher used Seaborn[40], which is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

7.3.2 Front-end

To show these results to the user, the researcher used flask, a micro web framework written in Python. The Flask app calls the database to display the real time data to the client interface. Many applications such as Pintreset and LinkedIn use Flask framework. The formats that the browser can display include HTML, CSS and JS triple. By combining all three elements, a browser is able the render a nice looking, interactive web site, web page and web application.

Step 6: Website Deployment

To deploy the website, PythonAnywhere[41] was used. PythonAnywhere is a web hosting service based on python programming language. It is a handy tool used frequently

[39] https://language.googleapis.com

[40] https://seaborn.pydata.org/

[41] https://www.pythonanywhere.com/

by python developers to get new static sites quickly up and running. PythonAnywhere has a straightforward integration with Flask.

7.4 Conclusion

This chapter has explained the importance of the tool for evaluating controversial health events in real-time where it can be used effectively by security analysts, organizations, and professional agencies that want to monitor content of its interest on Twitter. The platform for TweetChecker is used only for health controversial topics. Health-related issues are particularly likely to become controversial topics when plagued by poor-quality information. In order to encourage users to interact with TweetChecker, a back-end and a front-end platform was built in Amazon AWS. The backend of the platform contains several steps: (1) data streaming, (2) streaming pipeline, (3) data storage, (4) data analysis: which do two types of analysis, (i) bot analysis: where troll accounts and social bots are designed to attempt to manipulate the public opinion towards different topics, and for Twitter, an openly accessible solution called Botometer is used. (ii) URL analysis: is a main feature of credibility in social media texts because the inclusion of an URL makes people more likely to believe the content of the post. (5) visualizations: provide the user with a quick interface and big-picture about the analysis. (6) front-end: a flask app, a micro web framework written in Python, is used to show the results to the user. The website is deployed through Pythonanywhere to get new static sites up and running quickly.

CHAPTER 8: CONCLUSIONS

8.1 Main Conclusions and Result Summary

This thesis aimed to understand how misinformation is manifested in social media and specifically regarding controversial health topics especially in Twitter. Having understood that online information sources are gradually replacing and supplementing traditional media outlets, it was important to understand the impacts. It's clear that real-time information via social media spreads faster to a wider audience as compared to traditional news media outlets. However, the same social media was seen to have effects on our community that needed to be discussed. Negative impacts included misleading information, spread of fake news, rumor and unsubstantiated information.

The objectives of the study identified were: (i) the ability to mitigate misinformation spread in social media by Twitter that helps to identify and spot misinformation in tweets, (ii) the ability to understand and detect people's opinions towards controversial issues that will help to detect polarized and force information, (iii) improve assessment literacy to evaluate controversial health issues discussion in real time. In the literature review, we find out that the research makes use of Natural Language Processing (NLP) that includes opinion mining, sentiment analysis, content analysis. Approaches to natural learning Processing task included machine learning based approach, lexicon-based approach, combined approach.

The main source of data in this study was by use of Twitter that was used to every public post and also the data was obtained from Crimson Hexagon13. We classified six types of evidence people use in their tweets to support their arguments. We found that Support Vector Machines (SVM) classifiers trained with n-grams and other features capture the different types of evidence used in social media and demonstrate significant improvement over the unigram baseline, achieving a macro-averaged F1 score of 82.8 %. The research tried to trace the frequency of participation in controversial discussions on social media. Through correlating users' stances with their sentiments and demographics, users' behavior online could be described. A case study about the debate towards MMR vaccines was discussed as a health controversial issue. The social movement regarding women driving in Saudi Arabia was also discussed as an example of social controversial issue.

To scale up the work done in this thesis, a web application tool, TweetChecker was implemented. Tweet Checker as a web application tool could allow online users to have a more in-depth understanding of the discussions made towards different social controversial issues. TweetChecker could help in understanding misinformation and how it is manifested in social media. The three main parts of a twitter message that can affect the spread of misinformation in social media are: the tweet text or content, the URL which allows the user to retain characters and often provides analytic measurements, and the Message writer.

The results clearly showed that Russian troll accounts that discussed the 2016 U.S. election used deceptive language to influence public opinion and spread biased political information on social media. The theory-driven linguistic analysis was able to capture features of the deceptive language in one way or the other. People with anti-vaccine attitudes linked many times to the same URL while people with pro-vaccine attitudes linked to fewer overall sources but from a wider range of resources they provided fewer total links compared to anti-vaccine. Moreover, the journalists have a huge impact on users' opinions. For women driving in Saudi Arabia the ratio of opposing tweets was lowest after the opinions were officially announced. There were more male tweeters than female tweeters in the sample. The analysis of the gender and location of tweeters showed that women were more opinionated (both, pro and against women driving) than the men, and most tweets on this topic originated from Saudi Arabia, Turkey and the USA, respectively. A Support Vector Machine (SVM) classifier trained with n-gram and additional features is capable of capturing the different forms of representing evidence on Twitter, and exhibits significant improvements over the unigram baseline, achieving a F1 macro average of 82.8%. This work can provide an estimate for how adequately the arguments have been supported. The tool which was built to assess online users with the mitigation of misinformation spread in social media could provide more information regarding tweets that discuss the different health controversial issues.

8.2 Contributions and Implications

The TwitterChecker is a tool that is capable of bringing better assessment of the information available in the social media regarding social controversial issues. More than that, it is a catalyst that would help in giving a bigger picture of controversial issues discussed online. Women driving controversial debate was one component that made headlines in this study. Thorough investigation about women driving in Saudi Arabia could help explore the relationship between a change in policy and the expression of peoples' opinions on social media. This study might also assist in the development of models of social behavior that fit the Saudi Arabic culture.

Another one is Vaccines controversial health debate. Indeed, such a controversial health debate could help us make more accurate interpretations of people's attitudes and opinions regarding controversial health topics, analyze the scientific information sharing behaviors on Twitter, and examine the usage pattern of scientific information resources by both polarized opinions. The debate let us know how people use information sources when they have different opinions towards the issue. Moreover, the controversial health debate helps us to understand the sharing of different types of information sources when discussing a controversial issue, recognize Automatic Classification of types of information sources used in controversial discussions, identify features that can help with the classification of different types of information sources, understand who online users contributing to the discussions of a controversial health issue as MMR vaccine are, figure out Automatic Classification of online users used in controversial discussions, identify features that can help with the classification of online users, identify the different opinions people have towards a controversial issue, and identify features that can help with the classification of different opinions.

Also, we have Encryption debate. The encryption debate could help us understand public opinions and attitudes towards controversial topics which could aid scholars, law enforcement officials, and policy-makers develop better policies and guidelines, know that people's attitudes and behaviors related to privacy are highly contextualized in the digital age, predict how arguments are supported, and pay close attention to information privacy and national security. While many scholars have conceptualized information privacy in various disciplines, investigations of individual users' attitudes and behaviors towards information privacy and national security remain limited.

8.3 Implications

One implication of our research is that it suggests that it is possible to understand who frequently participates in controversial discussions on social media. Moreover, correlating users' stances with their sentiments and demographics may help further describe users' behavior online. To add on this, predicting a user's stance toward a given issue can support the identification of social or political groups, help develop better recommendation systems, and/or tailor users' information preferences to their ideologies and beliefs. Additionally, it may provide engineers and designers with new ways of improving the design and users' acceptability of current technologies. Design and Build Systems could help with assessing the credibility of content on Twitter in real-time.

8.4 Recommendations

It is clearly vital to do more research in order to find out and understand how users perceive or think about the credibility of content posted on social media especially on the Twitter and internet in general with this negative impact will be eliminated. I recommend that it is very important to utilize the insights obtained from this work because we can use them to build solutions for the problem of trustworthiness of user generated content on different online social media, especially for controversial issues as discussed in this thesis, where the truthfulness of the information is hardly known and recognized and with this, we can achieve the best from our social media. Also, I suggest that it would be interesting and crucial to apply other improved methodologies on other social media websites, such as Facebook so as to obtain at least more reliable results and this would help us to solve emerging social media problems. Furthermore, I would recommend that we need to understand user perceptions that would help us develop better trust assessment tools that can aid end users to judge the quality of content in a better way. In future there is a need to build specialized real-time solutions for emergency responders and organizations that can monitor, and display crisis related tweets in an innovative way. It is advisable to look again the solutions for controversial issues because they have the functionality that can automatically identify tweets that are timely, well-written, novel, posted by reputable users, etc. Such a system would increase utilization of social media data for controversial issues.

8.5 Limitations

When it comes to data collection, dataset may fail to be representative of the overall opinions of Twitter users online as using Twitter and Crimson Hexagon as a data source and collection tool involves multiple types of potential sampling biases. It is apparent in social media data that the language and structure of tweet texts that users tend to use seem to be informal and incoherent in most occasions. Secondly, dataset may not be representative of the overall opinions of Twitter users online. Other peoples who uses other social media and who have information may be ignored and this may bring little than expected data. This is what leads to sampling biases when collecting the data.

Concerning data annotation, manual mark-up required annotators to understand culture. Annotating tweets related to a controversial topic requires annotators who not only understand the language used and its informing cultures, but who also understand the debate. Also, the TweetChecker may give unreliable data and results concerning the opinions and decisions of respondents and this results to ambiguous results. With this, recommendations may not be well given and supported in the quest to achieve the objectives and solutions to the problems. Others may give misinterpreted and exaggerated information on social controversial issues thus less that supported information. When it comes to Tweetchecker, it is not easy to maintain AWS and storage access, this means that the toll requires more maintenance. Together with that, more data points are needed to download, retrieve and be able to get pictures of the issue in the study.

8.6 Future Work

One important future work suggested is the improvement of TweetChecker tool. Automated methods and solutions based on supervised ranking techniques can be effectively used to improve the tool. More research is required to help online users identify misinformation quickly in twitter regarding social controversial issues. Also, the suggested tool can be improved to include other types of analysis such as stance analysis and user type analysis. Also building a classifier and a model to help predict the public's' acceptance of policy changes is also vital in improving Tweet Checker tool.

Additionally, this work only focused on one type of microblogging website which is Twitter to gather information, it would be more interesting to apply similar methodologies on other social media websites the likes of Facebook and Instagram. It should be noted that that more studies are also required to understand how users perceive the credibility of the whole content posted online. The important thing dwells on how to improve the Tweetchecker tool. This may include applying other types of analysis as temporal analysis and geolocation analysis, improving the design of the website for the tool to be more interactive, supervising credibility ranking techniques and complex search keywords.

8.7 Publication Targets

There are several conferences that represent the target audience of this dissertation, mainly conferences related to computational social science domains. Conferences as CSCW, CHI, WebSci, WSDM and ICWSM would be appropriate venues for the research conducted in this dissertation. Furthermore, this thesis would be appropriate at any computational linguistics conference such as Annual Meeting of the Association for Computational Linguistics (ACL) as this research is about investigating the linguistic and non-linguistics features of controversial texts.

BIBLIOGRAPHY

(CDC), C. f. D. C. a. P. (2018, October 23, 2018). Measles Cases and Outbreaks. Retrieved from https://www.cdc.gov/measles/cases-outbreaks.html

Abokhodair, N., Abbar, S., Vieweg, S., & Mejova, Y. (2016). *Privacy and twitter in qatar: traditional values in the digital world.* Paper presented at the Proceedings of the 8th ACM Conference on Web Science.

Abokhodair, N., & Vieweg, S. (2016). *Privacy & social media in the context of the Arab Gulf.* Paper presented at the Proceedings of the 2016 ACM Conference on Designing Interactive Systems.

Abu-Jbara, A., Diab, M., Dasigi, P., & Radev, D. (2012). *Subgroup detection in ideological discussions*. Paper presented at the 50th Annual Meeting of the Association for Computational Linguistics.

Ackermann, D., Chapman, S., & Leask, J. (2004). Media coverage of anthrax vaccination refusal by Australian Defence Force personnel. *Vaccine, 23*(3), 411-417.

Addawood, A. (2018). *Usage of Scientific References in MMR Vaccination Debates on Twitter.* Paper presented at the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).

Addawood, A., Alshamrani, A., Alqahtani, A., Diesner, J., & Broniatowski, D. (2018). *Women's Driving in Saudi Arabia–Analyzing the Discussion of a Controversial Topic on Twitter.* Paper presented at the SBP-BRIMS.

Addawood, A., & Bashir, M. (2016). *"What is your evidence?" A study of controversial topics on social media*. Paper presented at the 3rd Workshop on Argument Mining.

Addawood, A., Rezapour, R., Mishra, S., Schneider, J., & Diesner, J. (2017). *Developing an Information Source Lexicon*. Paper presented at the Workshop on Prioritising Online Content @ NIPS 2017.

Addawood, A., Schneider, J., & Bashir, M. (2017). *Stance Classification of Twitter Debates: The Encryption Debate as A Use Case.* Paper presented at the Proceedings of the 8th International Conference on Social Media & Society.

Aftab, A. (2015). California becomes a state divided after the controversial SB 277 vaccination bill is handed to Governor Brown. Retrieved from http://www.independent.co.uk/news/world/americas/california-becomes-a-state-divided-after-the-controversial-sb-277-vaccination-bill-is-handed-to-10355888.html

agencies, S. a. (2013, October 26, 2013). Dozens of Saudi Arabian women drive cars on day of protest against ban. Retrieved from https://www.theguardian.com/world/2013/oct/26/saudi-arabia-woman-driving-car-ban

Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008). *Finding high-quality content in social media.* Paper presented at the Proceedings of the 2008 international conference on web search and data mining.

Al-Ahmadi, H. (2011). Challenges facing women leaders in Saudi Arabia. *Human Resource Development International, 14*(2), 149-166.

Al-Dawood, A., Abokhodair, N., & Yarosh, S. (2017). *Against Marrying a Stranger: Marital Matchmaking Technologies in Saudi Arabia.* Paper presented at the Proceedings of the 2017 Conference on Designing Interactive Systems.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211-236.

Allem, J.-P., & Ferrara, E. (2018). Could social bots pose a threat to public health? *American Journal of Public Health, 108*(8), 1005.

Allem, J.-P., Ramanujam, J., Lerman, K., Chu, K.-H., Cruz, T. B., & Unger, J. B. (2017). Identifying sentiment of hookah-related posts on Twitter. *JMIR public health and surveillance, 3*(4).

Amazon, E. (2015). Amazon web services. *Available in:* http://aws. *amazon. com/es/ec2/(November 2012)*.

Anand, P., Walker, M., Abbott, R., Tree, J. E. F., Bowmani, R., & Minor, M. (2011). *Cats rule and dogs drool!: Classifying stance in online debate*. Paper presented at the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis.

Aquino, F., Donzelli, G., De Franco, E., Privitera, G., Lopalco, P. L., & Carducci, A. (2017). The web and public confidence in MMR vaccination in Italy. *Vaccine, 35*(35), 4494-4498.

Ashley, K. D., & Walker, V. R. (2013). *From Information Retrieval (IR) to Argument Retrieval (AR) for Legal Cases: Report on a Baseline Study.* Paper presented at the JURIX.

Ayers, M. S., & Reder, L. M. (1998). A theoretical review of the misinformation effect: Predictions from an activation-based memory model. *Psychonomic Bulletin & Review, 5*(1), 1-21.

Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. *arXiv preprint arXiv:1802.04291*.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science, 348*(6239), 1130-1132.

Banfield, A. (1982). Unspeakable sentences. In: Routledge and Kegan Paul, Boston.

Bao, P., Shen, H. W., Chen, W., & Cheng, X. Q. (2013). Cumulative effect in information diffusion: empirical study on a microblogging network. *PLoS One, 8*(10), e76027. doi:10.1371/journal.pone.0076027

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter, 6*(1), 20-29.

Bavelas, J. B., Black, A., Chovil, N., & Mullett, J. (1990). *Equivocal communication*: Sage Publications, Inc.

Beck, S. (2006). Scientists fear MMR link to autism. *Daily Mail UK*. Retrieved from http://www.dailymail.co.uk/news/article-388051/Scientists-fear-MMR-linkautism

Begg, N., Ramsay, M., White, J., & Bozoky, Z. (1998). Media dents confidence in MMR vaccine. *BMJ, 316*(7130), 561.

Betsch, C. (2011). Innovations in communication: The Internet and the psychology of vaccination decisions. *Euro Surveill, 16*(17), 1-6.

Betsch, C., Brewer, N. T., Brocard, P., Davies, P., Gaissmaier, W., Haase, N., . . . Reyna, V. F. (2012). Opportunities and challenges of Web 2.0 for vaccination decisions. *Vaccine, 30*(25), 3727-3733.

Betsch, C., Renkewitz, F., & Haase, N. (2013). Effect of narrative reports about vaccine adverse events and bias-awareness disclaimers on vaccine decisions: A simulation of an online patient social network. *Medical Decision Making, 33*(1), 14-25.

Bian, J., Topaloglu, U., & Yu, F. (2012). Towards Large-scale Twitter Mining for Drug-related Adverse Events. *SHB'12 : proceedings of the 2012 ACM International Workshop on Smart Health and Wellbeing : October 29, 2012, Maui, Hawaii, USA. International*

Workshop on Smart Health and Wellbeing (2012 : Maui, Hawaii), 2012, 25-32. doi:10.1145/2389707.2389713

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*: O'Reilly Media, Inc.

Bond, G. D., & Lee, A. Y. (2005). Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology, 19*(3), 313-329.

Borah, P. (2014). The Hyperlinked World: A Look at How the Interactions of News Frames and Hyperlinks Influence News Credibility and Willingness to Seek Information. *Journal of Computer-Mediated Communication, 19*(3), 576-590. doi:10.1111/jcc4.12060

Borenstein, S. (2016). Getting at the truth behind lying in politics. Retrieved from https://www.pbs.org/newshour/nation/getting-at-the-truth-behind-lying-in-politics

Borge-Holthoefer, J., Magdy, W., Darwish, K., & Weber, I. (2015). *Content and network dynamics behind Egyptian political polarization on Twitter.* Paper presented at the Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing.

Botta, A., De Donato, W., Persico, V., & Pescapé, A. (2016). Integration of cloud computing and internet of things: a survey. *Future Generation Computer Systems, 56*, 684-700.

Briscoe, E. J., Appling, D. S., & Hayes, H. (2014). *Cues to deception in social media communications.* Paper presented at the 2014 47th Hawaii International Conference on System Sciences (HICSS).

Broniatowski, D. A., Hilyard, K. M., & Dredze, M. (2016). Effective vaccine communication during the disneyland measles outbreak. *Vaccine, 34*(28), 3225-3228. doi:10.1016/j.vaccine.2016.04.044

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., . . . Dredze, M. (2018). Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. *American Journal of Public Health, 108*(10), 1378-1384. doi:10.2105/AJPH.2018.304567

Bryan, M. A., Gunningham, H., & Moreno, M. A. (2018). Content and accuracy of vaccine information on pediatrician blogs. *Vaccine, 36*(5), 765-770.

Bucchi, M., & Trench, B. (2014). *Routledge Handbook of Public Communication of Science and Technology*: Routledge.

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication theory, 6*(3), 203-242.

Buller, D. B., Burgoon, J. K., Daly, J., & Wiemann, J. (1994). Deception: Strategic and nonstrategic communication. *Strategic interpersonal communication*, 191-223.

Burgoon, J. K., Blair, J. P., Qin, T., & Nunamaker, J. F. (2003). *Detecting deception through linguistic analysis.* Paper presented at the International Conference on Intelligence and Security Informatics.

Burgoon, J. K., Buller, D. B., Guerrero, L. K., Afifi, W. A., & Feldman, C. M. (1996). Interpersonal deception: XII. Information management dimensions underlying deceptive and truthful messages. *Communications Monographs, 63*(1), 50-69.

Bybee, J. L., Perkins, R. D., & Pagliuca, W. (1994). *The evolution of grammar: Tense, aspect, and modality in the languages of the world* (Vol. 196): University of Chicago Press Chicago.

Cabrio, E., & Villata, S. (2012). *Combining textual entailment and argumentation theory for supporting online debates interactions.* Paper presented at the Proceedings of the 50th

Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2.

Carmel, D., Yom-Tov, E., Darlow, A., & Pelleg, D. (2006). *What makes a query difficult?* Paper presented at the 29th annual international ACM SIGIR conference on Research and development in information retrieval.

Carr, C. T., & Hayes, R. A. (2015). Social media: Defining, developing, and divining. *Atlantic Journal of Communication, 23*(1), 46-65.

Castillo, C., Mendoza, M., & Poblete, B. (2011). *Information credibility on twitter.* Paper presented at the Proceedings of the 20th international conference on World wide web.

Cataldi, J. R., Dempsey, A. F., & O'Leary, S. T. (2016). Measles, the media, and MMR: Impact of the 2014–15 measles outbreak. *Vaccine, 34*(50), 6375-6380.

Cavazos-Rehg, P. A., Krauss, M., Fisher, S. L., Salyer, P., Grucza, R. A., & Bierut, L. J. (2015). Twitter chatter about marijuana. *The Journal of adolescent health : official publication of the Society for Adolescent Medicine, 56*(2), 139-145. doi:10.1016/j.jadohealth.2014.10.270

Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook* (pp. 853-867): Springer

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321-357.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Editorial: Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter, 6*(1), 1-6.

Chen, Y., Conroy, N. J., & Rubin, V. L. (2015). News in an online world: The need for an "automatic crap detector". *Proceedings of the Association for Information Science and Technology, 52*(1), 1-4.

Clarke, C. E. (2008). A question of balance: The autism-vaccine controversy in the British and American elite press. *Science Communication, 30*(1), 77-107.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement, 20*(1), 37-46.

Corritore, C. L., Wiedenbeck, S., Kracher, B., & Marble, R. P. (2007). Online trust and health information websites. *SIGHCI 2007 Proceedings*, 20.

Coursey, D. (2009). Swine Flu Frenzy Demonstrates Twitter's Achilles Heel. Retrieved from http://www.pcworld.com/article/163920/swine_flu_twitter.html

Culotta, A. (2010). *Towards detecting influenza epidemics by analyzing Twitter messages*. Paper presented at the First Workshop on Social Media Analytics.

Danescu-Niculescu-Mizil, C., Cheng, J., Kleinberg, J., & Lee, L. (2012). *You had me at hello: How phrasing affects memorability.* Paper presented at the Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1.

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). *Botornot: A system to evaluate social bots.* Paper presented at the Proceedings of the 25th International Conference Companion on World Wide Web.

De Marneffe, M.-C., Manning, C. D., & Potts, C. (2012). Did it happen? The pragmatic complexity of veridicality assessment. *Computational linguistics, 38*(2), 301-333.

Deer, B. (2011a). How the case against the MMR vaccine was fixed. *Bmj, 342*, c5347.

Deer, B. (2011b). How the vaccine crisis was meant to make money. *Bmj, 342*, c5258.

Deiner, M. S., Fathy, C., Kim, J., Niemeyer, K., Ramirez, D., Ackley, S. F., . . . Porco, T. C. (2017). Facebook and Twitter vaccine sentiment in response to measles outbreaks. *Health informatics journal*, 1460458217740723.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., . . . Quattrociocchi, W. (2016). The spreading of misinformation online. *Proc Natl Acad Sci U S A, 113*(3), 554-559. doi:10.1073/pnas.1517441113

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., . . . Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences, 113*(3), 554-559.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological bulletin, 129*(1), 74.

DeStefano, F., Bhasin, T. K., Thompson, W. W., Yeargin-Allsopp, M., & Boyle, C. (2004). Age at first measles-mumps-rubella vaccination in children with autism and school-matched control subjects: A population-based study in metropolitan Atlanta. *Pediatrics, 113*(2), 259-266.

Dori-Hacohen, S., & Allan, J. (2013). *Detecting controversy on the web*. Paper presented at the 22nd ACM International Conference on Information & Knowledge Management.

Dori-Hacohen, S., Yom-Tov, E., & Allan, J. (2015). *Navigating controversy as a complex search task*. Paper presented at the ECIR Supporting Complex Search Task Workshop, Vienna, Austria.

Dube, E., Vivion, M., & MacDonald, N. E. (2015). Vaccine hesitancy, vaccine refusal and the anti-vaccine movement: influence, impact and implications. *Expert review of vaccines, 14*(1), 99-117.

Dumbrell, D., & Steele, R. (2017). Twitter and Its Role in Health Information Dissemination: Analysis of the Micro-Blog Posts of Health-Related Organisations. In *Public Health and Welfare: Concepts, Methodologies, Tools, and Applications* (pp. 372-388): IGI Global.

Dunlap, J. C., & Lowenthal, P. R. (2009). Tweeting the night away: Using Twitter to enhance social presence. *Journal of Information Systems Education, 20*(2), 129.

Eastin, M. S. (2001). Credibility Assessments of Online Health Information: The Effects of Source Expertise and Knowledge of Content. *Journal of Computer-Mediated Communication, 6*(4), 0-0. doi:doi:10.1111/j.1083-6101.2001.tb00126.x

Edwards, C., Edwards, A., Spence, P. R., & Shelton, A. K. (2014). Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter. *Computers in Human Behavior, 33*, 372-376. doi:https://doi.org/10.1016/j.chb.2013.08.013

Elamin, A. M., & Omair, K. (2010). Males' attitudes towards working females in Saudi Arabia. *Personnel Review, 39*(6), 746-766.

Etlinger, S., & Amand, W. (2012). Crimson Hexagon [Program documentation]. *Retrieved September, 15*, 2016.

Etlinger, S., & Amand, W. (2012). Crimson Hexagon [Program documentation]. Retrieved from http://www.crimsonhexagon.com/wp-cotent/uploads/2012/02/CrimsonHexagon_Altimeter_Webinar_111611.pdf

Eysenbach, G. (2003). The impact of the Internet on cancer outcomes. *CA: a cancer journal for clinicians, 53*(6), 356-371.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of medical Internet research, 13*(4).

Eysenbach, G., Powell, J., Kuss, O., & Sa, E.-R. (2002). Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review. *Jama, 287*(20), 2691-2700.

Fang, A., Ounis, I., Habel, P., Macdonald, C., & Limsopatham, N. (2015). *Topic-centric classification of Twitter user's political orientation*. Paper presented at the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval.

Faulkner, A. (2014). Automated classification of stance in student essays: An approach using stance target information and the Wikipedia link-based measure. *Science, 376*(12), 86.

Feng, S., Banerjee, R., & Choi, Y. (2012). *Syntactic stylometry for deception detection.* Paper presented at the Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2.

Feng, V. W., & Hirst, G. (2011). *Classifying arguments by scheme.* Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.

Fox, S. (2011). The social life of health information 2011. Retrieved from http://www.pewinternet.org/2011/05/12/the-social-life-of-health-information-2011/

Frank, M. G., & Feeley, T. H. (2003). To catch a liar: Challenges for research in lie detection training. *Journal of Applied Communication Research, 31*(1), 58-75.

Freed, G. L., Clark, S. J., Butchart, A. T., Singer, D. C., & Davis, M. M. (2011). Sources and perceived credibility of vaccine-safety information for parents. *Pediatrics, 127*(Supplement 1), S107-S112.

Freeley, A. J., & Steinberg, D. L. (2013). *Argumentation and debate*: Cengage Learning.

Frey, D. (1986). Recent research on selective exposure to information. *Advances in experimental social psychology, 19*, 41-80.

Fuller, C. M., Biros, D. P., & Wilson, R. L. (2009). Decision support for determining veracity via linguistic-based cues. *Decision Support Systems, 46*(3), 695-703.

Fung, I. C.-H., Tse, Z. T. H., Cheung, C.-N., Miu, A. S., & Fu, K.-W. (2014). Ebola and the social media. *The Lancet, 384*(9961), 2207.

Gabbatt, A. (2009). Fox News's hacked Twitter feed declares Obama dead. Retrieved from https://www.theguardian.com/news/blog/2011/jul/04/fox-news-hacked-twitter-obama-dead

Gallagher, C. M., & Goodman, M. S. (2010). Hepatitis B vaccination of male neonates and autism diagnosis, NHIS 1997–2002. *Journal of Toxicology and Environmental Health, Part A, 73*(24), 1665-1677.

Galland, A., Abiteboul, S., Marian, A., & Senellart, P. (2010). *Corroborating information from disagreeing views*. Paper presented at the third ACM international conference on Web search and data mining.

Gangarosa, E. J., Galazka, A., Wolfe, C., Phillips, L., Miller, E., Chen, R., & Gangarosa, R. (1998). Impact of anti-vaccine movements on pertussis control: The untold story. *The Lancet, 351*(9099), 356-361.

Gaskins, B., & Jerit, J. (2012). Internet news: Is it a replacement for traditional media outlets? *The International Journal of Press/Politics, 17*(2), 190-213.

Gawron, J. M., Gupta, D., Stephens, K., Tsou, M.-H., Spitzberg, B., & An, L. (2012). *Using group membership markers for group identification in web logs*. Paper presented at the AAAI Conference on Weblogs and Social Media (ICWSM).

Gayo-Avello, P. T. M., Eni Mustafaraj, Markus Strohmaier, Harald Schoen and Peter Gloor, Daniel, Castillo, C., Mendoza, M., & Poblete, B. (2013). Predicting information credibility in time-sensitive social media. *Internet Research, 23*(5), 560-588.

Gerber, J. S., & Offit, P. A. (2009). Vaccines and autism: a tale of shifting hypotheses. *Clin Infect Dis, 48*(4), 456-461. doi:10.1086/596476

Gerber, T. P., & Zavisca, J. (2016). Does Russian propaganda work? *The Washington Quarterly, 39*(2), 79-98.

Gil de Zúñiga, H., Jung, N., & Valenzuela, S. (2012). Social media use for news and individuals' social capital, civic engagement and political participation. *Journal of Computer-Mediated Communication, 17*(3), 319-336.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., . . . Smith, N. A. (2011). *Part-of-speech tagging for Twitter: Annotation, features, and experiments*. Paper presented at the 49th Annual Meeting of the Association for Computational Linguistics (ACL).

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering, 10*(4), 215-230.

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks, 38*, 16-27.

Gottfried, J., & Shearer, E. (2016). *News Use Across Social Medial Platforms 2016*: Pew Research Center.

Green, N., Ashley, K., Litman, D., Reed, C., & Walker, V. (2014). *Proceedings of the First Workshop on Argumentation Mining.* Paper presented at the Proceedings of the First Workshop on Argumentation Mining.

Greenhalgh, T. (2014). *How to read a paper: The basics of evidence-based medicine*: John Wiley & Sons.

Gruzd, A., & Goertzen, M. (2013). *Wired academia: Why social science scholars are using social media*. Paper presented at the 46th Hawaii International Conference on System Sciences (HICSS).

Gualtieri, L. N. (2009). *The doctor as the second opinion and the internet as the first*. Paper presented at the CHI'09 Extended Abstracts on Human Factors in Computing Systems.

Guess, A., Nyhan, B., & Reifler, J. (2018). Selective Exposure to Misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign. *European Research Council.*

Gupta, A., & Kumaraguru, P. (2012). *Credibility ranking of tweets during high impact events*. Paper presented at the 1st Workshop on Privacy and Security in Online Social Media.

Hachey, B., & Grover, C. (2005). *Sequence modelling for sentence classification in a legal summarisation system.* Paper presented at the Proceedings of the 2005 ACM symposium on Applied computing.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research, 20*(1), 98-116.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM Sigkdd Explorations Newsletter, 11*(1), 10-18.

Hampton, K., Rainie, L., Lu, W., Dwyer, M., Shin, I., & Purcell, K. (2014). Social Media and the 'Spiral of Silence'. Retrieved from http://www.pewinternet.org/2014/08/26/social-media-and-the-spiral-of-silence/

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2007). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes, 45*(1), 1-23.

Hargittai, E. (2015). Is bigger always better? Potential biases of big data derived from social network sites. *The Annals of the American Academy of Political and Social Science, 659*(1), 63-76.

Hasan, K. S., & Ng, V. (2013). *Stance classification of ideological debates: Data, models, features, and constraints*. Paper presented at the International Joint Conference on Natural Language Processing (IJCNLP).

Hilton, S., Petticrew, M., & Hunt, K. (2007). Parents' champions vs. vested interests: who do parents believe about MMR? A qualitative study. *BMC Public Health, 7*(1), 42.

Hiltzik, M. (2018). Anti-vaccine stupidity returns, as measles cases rise and California parents evade the law. *LA Times*. Retrieved from http://www.latimes.com/business/hiltzik/la-fi-hiltzik-measles-vaccine-20181030-story.html

Hyland, K. (1998). *Hedging in scientific research articles* (Vol. 54): John Benjamins Publishing.

Hyland, K. (2018). *Metadiscourse: Exploring interaction in writing*: Bloomsbury Publishing.

Ickes, W., Reidhead, S., & Patterson, M. (1986). Machiavellianism and self-monitoring: As different as “me” and “you”. *Social Cognition, 4*(1), 58-74.

Jin, F., Wang, W., Zhao, L., Dougherty, E., Cao, Y., Lu, C.-T., & Ramakrishnan, N. (2014). Misinformation propagation in the age of twitter. *Computer, 47*(12), 90-94.

Johnson, M. K., & Raye, C. L. (1998). False memories and confabulation. *Trends in cognitive sciences, 2*(4), 137-145.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260.

Junco, R., Heiberger, G., & Loken, E. (2011). The effect of Twitter on college student engagement and grades. *Journal of Computer Assisted Learning, 27*(2), 119-132.

Jurgens, D., Finethy, T., McCorriston, J., Xu, Y. T., & Ruths, D. (2015). Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice. *ICWSM, 15*, 188-197.

Kang, G. J., Ewing-Nelson, S. R., Mackey, L., Schlitt, J. T., Marathe, A., Abbas, K. M., & Swarup, S. (2017). Semantic network analysis of vaccine sentiment in online social media. *Vaccine, 35*(29), 3621-3638.

Kasperson, R. E., Renn, O., Slovic, P., Brown, H. S., Emel, J., Goble, R., . . . Ratick, S. (1988). The social amplification of risk: A conceptual framework. *Risk analysis, 8*(2), 177-187.

Kata, A. (2010). A postmodern Pandora's box: Anti-vaccination misinformation on the Internet. *Vaccine, 28*(7), 1709-1716.

Kata, A. (2012). Anti-vaccine activists, Web 2.0, and the postmodern paradigm – An overview of tactics and tropes used online by the anti-vaccination movement. *Vaccine., 30*(25), 3778-3789. doi:10.1016/j.vaccine.2011.11.112

Khan, A. S., Fleischauer, A., Casani, J., & Groseclose, S. L. (2010). The next public health revolution: Public health information fusion and social networks. *American Journal of Public Health, 100*(7), 1237-1242.

Khazaei, T., Lu, X., & Mercer, R. (2017). Writing to persuade: Analysis and detection of persuasive discourse. *iConference 2017 Proceedings*.

Kinsella, S., Wang, M., Breslin, J., & Hayes, C. (2011). Improving categorisation in social media using hyperlinks to structured data sources. *The Semanic Web: Research and Applications*, 390-404.

Kluberg, S. A., McGinnis, D. P., Hswen, Y., Majumder, M. S., Santillana, M., & Brownstein, J. S. (2017). County-level assessment of United States kindergarten vaccination rates for measles mumps rubella (MMR) for the 2014-2015 school year. *Vaccine, 35*(47), 6444-6450. doi:10.1016/j.vaccine.2017.09.080

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering, 30*(1), 25-36.

Kowitt, S. D., Schmidt, A. M., Hannan, A., & Goldstein, A. O. (2017). Awareness and trust of the FDA and CDC: Results from a national sample of US adults and adolescents. *PLoS One, 12*(5), e0177546. doi:10.1371/journal.pone.0177546

Krauss, M. J., Sowles, S. J., Moreno, M., Zewdie, K., Grucza, R. A., Bierut, L. J., & Cavazos-Rehg, P. A. (2015). Peer reviewed: Hookah-related twitter chatter: A content analysis. *Preventing chronic disease, 12*.

Kucher, K., Schamp-Bjerede, T., Kerren, A., Paradis, C., & Sahlgren, M. (2016). Visual analysis of online social media to open up the investigation of stance phenomena. *Information Visualization, 15*(2), 93-116.

Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., . . . Socher, R. (2016). *Ask me anything: Dynamic memory networks for natural language processing.* Paper presented at the International Conference on Machine Learning.

Lachlan, K. A., Spence, P. R., & Lin, X. (2014). Expressions of risk awareness and concern through Twitter: On the utility of using the medium as an indication of audience needs. *Computers in Human Behavior, 35*, 554-559.

Lampos, V., & Cristianini, N. (2010). *Tracking the flu pandemic by monitoring the social web.* Paper presented at the 2nd International Workshop on Cognitive Information Processing (CIP).

Larcker, D. F., & Zakolyukina, A. A. (2012). Detecting deceptive discussions in conference calls. *Journal of Accounting Research, 50*(2), 495-540.

Larson, H. J., Cooper, L. Z., Eskola, J., Katz, S. L., & Ratzan, S. (2011). Addressing the vaccine confidence gap. *The Lancet, 378*(9790), 526-535.

Lazer, D., Baum, M., Grinberg, N., Friedland, L., Joseph, K., Hobbs, W., & Mattsson, C. (2017). Combating fake news: An agenda for research and action. *Harvard Kennedy School, Shorenstein Center on Media, Politics and Public Policy, 2*.

Lee, D. (2016, February 18). Apple vs the FBI—A plain English guide. Retrieved from http://www.bbc.com/news/technology-35601035

Levine, T. R., Park, H. S., & McCornack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the "veracity effect". *Communications Monographs, 66*(2), 125-144.

Levitan, S. I., Maredia, A., & Hirschberg, J. (2018). *Linguistic Cues to Deception and Perceived Deception in Interview Dialogues.* Paper presented at the Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction continued influence and successful debiasing. *Psychological Science in the Public Interest, 13*(3), 106-131.

Lewis, J., & Speers, T. (2003). Misleading media reporting? The MMR story. *Nature Reviews Immunology, 3*(11), 913-918.

Li, J., & Greenhow, C. (2015). Scholars and social media: Tweeting in the conference backchannel for professional learning. *Educational Media International, 52*(1), 1-14.

Li, J. J., & Nenkova, A. (2015). *Fast and Accurate Prediction of Sentence Specificity.* Paper presented at the AAAI.

Liakata, M., Saha, S., Dobnik, S., Batchelor, C., & Rebholz-Schuhmann, D. (2012). Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics, 28*(7), 991-1000.

Liao, Q. V., & Fu, W.-T. (2013). *Beyond the filter bubble: interactive effects of perceived threat and topic involvement on selective exposure to information.* Paper presented at the Proceedings of the SIGCHI conference on human factors in computing systems.

Lin, W.-H., Wilson, T., Wiebe, J., & Hauptmann, A. (2006). *Which side are you on?: Identifying perspectives at the document and sentence levels*. Paper presented at the Tenth Conference on Computational Natural Language Learning.

Liu, Y. (2016). Mining Social Media to Understand Consumers' Health Concerns and the Public's Opinion on Controversial Health Topics.

Livingston, K. A., Rosen, J. B., Zucker, J. R., & Zimmerman, C. M. (2014). Mumps vaccine effectiveness and risk factors for disease in households during an outbreak in New York City. *Vaccine, 32*(3), 369-374. doi:10.1016/j.vaccine.2013.11.021

Llewellyn, C., Grover, C., Oberlander, J., & Klein, E. (2014). *Re-using an Argument Corpus to Aid in the Curation of Social Media Collections.* Paper presented at the LREC.

Love, B., Himelboim, I., Holton, A., & Stewart, K. (2013). Twitter as a source of vaccination information: content drivers and what they are saying. *American journal of infection control, 41*(6), 568-570.

Maddock, J., Starbird, K., Al-Hassani, H. J., Sandoval, D. E., Orand, M., & Mason, R. M. (2015). *Characterizing online rumoring behavior using multi-dimensional signatures*. Paper presented at the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing.

Magdy, W., Darwish, K., Abokhodair, N., Rahimi, A., & Baldwin, T. (2016). *# isisisnotislam or# deportallmuslims?: Predicting unspoken views.* Paper presented at the Proceedings of the 8th ACM Conference on Web Science.

Maglione, M. A., Das, L., Raaen, L., Smith, A., Chari, R., Newberry, S., . . . Gidengil, C. (2014). Safety of vaccines used for routine immunization of US children: A systematic review. *Pediatrics, 134*(2), 325-337.

Mandavilli, A. (2011). Trial by Twitter. *Nature, 469*(7330), 286.

Mann, B. W. (2018). Autism Narratives in Media Coverage of the MMR Vaccine-Autism Controversy under a Crip Futurism Framework. *Health Communication*, 1-7.

Matsa, K. E., & Mitchell, A. (2014). 8 key takeaways about social media and news. *Pew Research Journalism Project, March, 26*.

Matthews, C. (2013). How does one fake tweet cause a stock market crash. *Wall Street & Markets: Time*.

Mendoza, M., Poblete, B., & Castillo, C. (2010). *Twitter Under Crisis: Can we trust what we RT?* Paper presented at the first workshop on social media analytics.

Metzger, M. J., & Flanagin, A. J. (2011). Using Web 2.0 technologies to enhance evidence-based medical information. *Journal of health communication, 16*(sup1), 45-58.

Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics, 59*, 210-220.

Mihalcea, R., & Strapparava, C. (2009). *The lie detector: Explorations in the automatic recognition of deceptive language.* Paper presented at the Proceedings of the ACL-IJCNLP 2009 Conference Short Papers.

Miller, L. M. S., & Bell, R. A. (2012). Online health information seeking: the influence of age, information trustworthiness, and search challenges. *Journal of aging and health, 24*(3), 525-541.

Mitra, T., Wright, G. P., & Gilbert, E. (2017). *A parsimonious language model of social media credibility across disparate events.* Paper presented at the Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing.

Mochales, R., & Ieven, A. (2009). *Creating an argumentation corpus: do theories apply to real arguments?: a case study on the legal argumentation of the ECHR.* Paper presented at the Proceedings of the 12th international conference on artificial intelligence and law.

Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). Semeval-2016 Task 6: Detecting stance in tweets. *Proceedings of SemEval, 16*.

Mohammad, S. M., Zhu, X., Kiritchenko, S., & Martin, J. (2015). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management, 51*(4), 480-499.

Mohanty, S., Buttenheim, A. M., Joyce, C. M., Howa, A. C., Salmon, D., & Omer, S. B. (2018). Experiences With Medical Exemptions After a Change in Vaccine Exemption Policy in California. *Pediatrics, 142*(5). doi:10.1542/peds.2018-1051

Morahan-Martin, J., & Anderson, C. D. (2000). Information and misinformation online: Recommendations for facilitating accurate mental health information retrieval and evaluation. *CyberPsychology & Behavior, 3*(5), 731-746.

Moran, M. B., Lucas, M., Everhart, K., Morgan, A., & Prickett, E. (2016). What makes anti-vaccine websites persuasive? A content analysis of techniques used by anti-vaccine websites to engender anti-vaccine sentiment. *Journal of Communication in Healthcare, 9*(3), 151-163. doi:10.1080/17538068.2016.1235531

Morris, M. R., Counts, S., Roseway, A., Hoff, A., & Schwarz, J. (2012). *Tweeting is believing?: Understanding microblog credibility perceptions*. Paper presented at the ACM 2012 conference on Computer Supported Cooperative Work.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). *Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose.* Paper presented at the ICWSM.

Mourtada, R., & Salem, F. (2014). Citizen engagement and public services in the Arab world: The potential of social media.

Mourtada, R., Salem, F., Al-Dabbagh, M., & Gargani, G. (2011). The role of social media in Arab Women's empowerment. *Dubai, Dubai School of Government, 1*, 26.

Mukhtar, N., Khan, M. A., & Chiragh, N. (2018). Lexicon-based approach outperforms Supervised Machine Learning approach for Urdu Sentiment Analysis in multiple domains. *Telematics and Informatics, 35*(8), 2173-2183.

Murakami, A., & Raymond, R. (2010). *Support or oppose?: classifying positions in online debates from reply activities and opinion expressions*. Paper presented at the 23rd International Conference on Computational Linguistics.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin, 29*(5), 665-675.

Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L., & Konstan, J. A. (2014). *Exploring the filter bubble: the effect of using recommender systems on content diversity*. Paper presented at the 23rd international conference on World wide web.

NLM. (2013). MEDLINE Fact sheet. Retrieved from http://www.nlm.nih.gov/pubs/

Novin, A., & Meyers, E. (2016). *Controversial Search Engine Results: An Exploratory Study of Information Presentation and Use*. Paper presented at the Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior, 32*(2), 303-330.

Odlum, M., & Yoon, S. (2015). What can we learn about the Ebola outbreak from tweets? *American journal of infection control, 43*(6), 563-571.

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). *Finding deceptive opinion spam by any stretch of the imagination.* Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.

Palau, R. M., & Moens, M.-F. (2009). *Argumentation mining: the detection, classification and structure of arguments in text.* Paper presented at the Proceedings of the 12th international conference on artificial intelligence and law.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up?: Sentiment classification using machine learning techniques*. Paper presented at the Conference on Empirical Methods in Natural Language Processing.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. New York: Penguin UK.

Park, H., Reber, B. H., & Chon, M.-G. (2016). Tweeting as health communication: health organizations' use of Twitter for health promotion and public engagement. *Journal of Health Communication, 21*(2), 188-198.

Park, J., & Cardie, C. (2014). *Identifying appropriate support for propositions in online user comments.* Paper presented at the Proceedings of the First Workshop on Argumentation Mining.

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological science, 8*(3), 162-166.

Pennebaker, J. W. (2011). The secret life of pronouns. *New Scientist, 211*(2828), 42-45.

Pennebaker, J. W., & Francis, M. E. (1996). Cognitive, emotional, and language processes in disclosure. *Cognition & Emotion, 10*(6), 601-626.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology, 77*(6), 1296.

Pennebaker, J. W., Mayne, T. J., & Francis, M. E. (1997). Linguistic predictors of adaptive bereavement. *Journal of personality and social psychology, 72*(4), 863.

Pennycook, G., & Rand, D. G. (2018). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking.

Poland, G. A., & Jacobson, R. M. (2011). The age-old struggle against the antivaccinationists. *New England Journal of Medicine, 364*(2), 97-99.

Popescu, A.-M., & Pennacchiotti, M. (2010). *Detecting controversial events from twitter*. Paper presented at the Proceedings of the 19th ACM international conference on Information and knowledge management.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2017). Automatic Detection of Fake News. *arXiv preprint arXiv:1708.07104*.

Qiu, M., Yang, L., & Jiang, J. (2013). *Modeling interaction features for debate side clustering*. Paper presented at the 22nd ACM International Conference on Information & Knowledge Management.

Radzikowski, J., Stefanidis, A., Jacobsen, K. H., Croitoru, A., Crooks, A., & Delamater, P. L. (2016). The measles vaccination narrative in Twitter: A quantitative analysis. *JMIR public health and surveillance, 2*(1).

Rajadesingan, A., & Liu, H. (2014). *Identifying users with opposing opinions in Twitter debates*. Paper presented at the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction.

Ramanathan, R., Voigt, E. A., Kennedy, R. B., & Poland, G. A. (2018). Knowledge gaps persist and hinder progress in eliminating mumps. *Vaccine, 36*(26), 3721-3726. doi:10.1016/j.vaccine.2018.05.067

Rao, A., Spasojevic, N., Li, Z., & DSouza, T. (2015). *Klout score: Measuring influence across multiple social networks*. Paper presented at the Big Data (Big Data), 2015 IEEE International Conference on.

Raskin, D. C., & Esplin, P. W. (1991). Statement validity assessment: Interview procedures and content analysis of children's statements of sexual abuse. *Behavioral Assessment*.

Reed, C., & Rowe, G. (2004). Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools, 13*(04), 961-979.

Reyes, A., Rosso, P., & Buscaldi, D. (2012). From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering, 74*, 1-12.

Rezapour, R., & Diesner, J. (2017). *Classification and detection of micro-level impact of issue-focused documentary films based on reviews*. Paper presented at the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing.

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the Web. *Journal of the American society for information science and technology, 53*(2), 145-161.

Rieke, R. D., & Sillars, M. O. (1984). *Argumentation and the decision making process*: Addison-Wesley Longman.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013). *Sarcasm as contrast between a positive sentiment and negative situation*. Paper presented at the 2013 Conference on Empirical Methods in Natural Language Processing.

Rinott, R., Dankin, L., Perez, C. A., Khapra, M. M., Aharoni, E., & Slonim, N. (2015). *Show me your evidence-an automatic method for context dependent evidence detection.* Paper presented at the Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.

Robinson-García, N., Torres-Salinas, D., Zahedi, Z., & Costas, R. (2014). New data, new possibilities: Exploring the insides of Altmetric. com. *El profesional de la información, 23*(4), 1386-6710.

Roobaert, D., Karakoulas, G., & Chawla, N. V. (2006). Information gain, correlation and Support Vector Machines. In *Feature Extraction* (pp. 463-470): Springer.

Ross, C., Terras, M., Warwick, C., & Welsh, A. (2011). Enabled backchannel: Conference Twitter use by digital humanists. *Journal of Documentation, 67*(2), 214-237.

Rubin, V. L. (2017). Deception detection and rumor debunking for social media. *The SAGE Handbook of Social Media Research Methods*, 342-363.

Rubin, V. L., Chen, Y., & Conroy, N. J. (2015). *Deception detection for news: three types of fakes.* Paper presented at the Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community.

Rubin, V. L., & Conroy, N. J. (2011). Challenges in automated deception detection in computer-mediated communication. *Proceedings of the American Society for Information Science and Technology, 48*(1), 1-4.

Rubin, V. L., Liddy, E. D., & Kando, N. (2006). Certainty identification in texts: Categorization model and manual tagging results. In *Computing attitude and affect in text: Theory and applications* (pp. 61-76): Springer.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science, 346*(6213), 1063-1064.

Salathe, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., . . . Mabry, P. L. (2012). Digital epidemiology. *PLoS computational biology, 8*(7).

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

Savolainen, R. (2011). Requesting and providing information in blogs and internet discussion forums. *Journal of Documentation, 67*(5), 863-886.

Schneider, J., Samp, K., Passant, A., & Decker, S. (2013). *Arguments about deletion: How experience improves the acceptability of arguments in ad-hoc online task groups.* Paper presented at the Proceedings of the 2013 conference on Computer supported cooperative work.

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

Shaw, C., & Tomljenovic, L. (2013). Aluminum in the central nervous system (CNS): Toxicity in humans and animals, vaccine adjuvants, and autoimmunity. *Immunologic research, 56*(2-3), 304-316.

Shema, H., Bar-Ilan, J., & Thelwall, M. (2015). How is research blogged? A content analysis approach. *Journal of the Association for Information Science and Technology, 66*(6), 1136-1149.

Sherf, E. N., Tangirala, S., & Weber, K. C. (2017). It Is Not My Place! Psychological Standing and Men's Voice and Participation in Gender-Parity Initiatives. *Organization Science, 28*(2), 193-210.

Shuy, R. W. (1998). *The language of confession, interrogation, and deception* (Vol. 2): Sage.

Sikdar, S. K., Kang, B., O'Donovan, J., Hollerer, T., & Adal, S. (2013). Cutting through the noise: Defining ground truth in information credibility on twitter. *Human, 2*(3), 151-167.

Simon, T., Goldberg, A., Aharonson-Daniel, L., Leykin, D., & Adini, B. (2014). Twitter in the cross fire—the use of social media in the Westgate Mall terror attack in Kenya. *PloS one, 9*(8).

Smart, J. M., & Burling, D. (2001). Radiology and the internet: a systematic review of patient information resources. *Clinical radiology, 56*(11), 867-870.

Smith, L. M., Zhu, L., Lerman, K., & Kozareva, Z. (2013). *The role of social media in the discussion of controversial topics*. Paper presented at the 2013 International Conference on Social Computing (SocialCom).

Sobhani, P., Inkpen, D., & Matwin, S. (2015). *From argumentation mining to stance classification*. Paper presented at the 2nd Workshop on Argumentation Mining.

Somasundaran, S., & Wiebe, J. (2010). *Recognizing stances in ideological on-line debates*. Paper presented at the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.

Son, I., Lee, D., & Kim, Y. (2013). *Understanding the Effect of Message Content and User Identity on Information Diffusion in Online Social Networks.* Paper presented at the PACIS.

Speers, T., & Lewis, J. (2004). Journalists and jabs: Media coverage of the MMR vaccine. *Communication & Medicine, 1*(2), 171-181.

Stab, C., & Gurevych, I. (2014). *Annotating argument components and relations in persuasive essays.* Paper presented at the Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers.

Starbird, K. (2017). *Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter*. Paper presented at the International Conference on Web and Social Media.

Statistic. (2018). Saudi Arabia: number of internet users 2022. Retrieved from https://www.statista.com/statistics/462959/internet-users-saudi-arabia/

Steller, M., & Koehnken, G. (1989). Criteria-based statement analysis.

Sudau, F., Friede, T., Grabowski, J., Koschack, J., Makedonski, P., & Himmel, W. (2014). Sources of information and behavioral patterns in online health forums: observational study. *Journal of medical Internet research, 16*(1), e10.

Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010). *Want to be retweeted? large scale analytics on factors impacting retweet in twitter network.* Paper presented at the 2010 IEEE Second International Conference on Social Computing.

Sullivan, S. J., Schneiders, A. G., Cheang, C.-W., Kitto, E., Lee, H., Redhead, J., . . . McCrory, P. R. (2012). 'What's happening?'A content analysis of concussion-related traffic on Twitter. *British Journal of Sports Medicine, 46*(4), 258-263.

Sunstein, C. R. (2009). *Republic. com 2.0*: Princeton University Press.

Taboada, M. (2016). Sentiment analysis: an overview from linguistics.

Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016). *Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions.* Paper presented at the Proceedings of the 25th international conference on world wide web.

Tanaka, Y., Sakamoto, Y., & Honda, H. (2014). *The impact of posting URLs in disaster-related tweets on rumor spreading behavior.* Paper presented at the System Sciences (HICSS), 2014 47th Hawaii International Conference on.

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

Taylor, L. E., Swerdfeger, A. L., & Eslick, G. D. (2014). Vaccines are not associated with autism: An evidence-based meta-analysis of case-control and cohort studies. *Vaccine, 32*(29), 3623-3629.

Teufel, S., & Moens, M. (2002). Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational linguistics, 28*(4), 409-445.

Thomas, M., Pang, B., & Lee, L. (2006). *Get out the vote: Determining support or opposition from congressional floor-debate transcripts*. Paper presented at the 2006 Conference on Empirical Methods in Natural Language Processing.

Toma, C. L., & Hancock, J. T. (2012). What lies beneath: The linguistic traces of deception in online dating profiles. *Journal of Communication, 62*(1), 78-97.

Tomeny, T. S., Vargo, C. J., & El-Toukhy, S. (2017). Geographic and demographic correlates of autism-related anti-vaccine beliefs on Twitter, 2009-15. *Social science & medicine, 191*, 168-175.

Toulmin, S. E. (2003). *The uses of argument*: Cambridge university press.

Tsikerdekis, M., & Zeadally, S. (2014). Online deception in social media. *Communications of the ACM, 57*(9), 72-80.

Undeutsch, U. (1989). The development of statement reality analysis. In *Credibility assessment* (pp. 101-119): Springer.

UNICEF. (2013). Tracking anti-vaccination sentiment in eastern European social media networks. *New York: UNICEF.*

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

Veletsianos, G. (2012). Higher education scholars' participation and practices on Twitter. *Journal of Computer Assisted Learning, 28*(4), 336-349.

Villalba, M. P. G., & Saint-Dizier, P. (2012). Some Facets of Argument Mining for Opinion Analysis. *COMMA, 245*, 23-34.

Vrij, A. (2000). Detecting Lies and Deceit: The Psychology of Lying and Implications for Professional Practice.(Wiley Series on the Psychology of Crime, Policing and Law).

Wagner, L. (2016). How "Zika-Proof" Hope Solo Became the Biggest Villain of the Rio Olympics. *Slate*. Retrieved from https://slate.com/culture/2016/08/how-zika-proof-hope-solo-became-the-biggest-villain-of-the-rio-olympics.html

Wakefield, A. J., Murch, S. H., Anthony, A., Linnell, J., Casson, D., Malik, M., . . . Harvey, P. (1998). RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children.

Walker, M. A., Anand, P., Abbott, R., & Grant, R. (2012). *Stance classification using dialogic properties of persuasion*. Paper presented at the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Walton, D., Reed, C., & Macagno, F. Argumentation schemes. 2008. In: Cambridge University Press.

Wang, W., Hernandez, I., Newman, D. A., He, J., & Bian, J. (2016). Twitter analysis: Studying US weekly trends in work stress and emotion. *Applied Psychology, 65*(2), 355-378.

Wang, X., Wei, F., Liu, X., Zhou, M., & Zhang, M. (2011). *Topic sentiment analysis in Twitter: A graph-based hashtag sentiment classification approach*. Paper presented at the 20th ACM international conference on Information and knowledge management.

Wiebe, J. (2000). Learning subjective adjectives from corpora. *Aaai/iaai, 20*(0), 0.

Wikgren, M. (2001). Health discussions on the Internet: A study of knowledge communication through citations. *Library & Information Science Research, 23*(4), 305-317. doi:https://doi.org/10.1016/S0740-8188(01)00091-3

Wilson, T., Wiebe, J., & Cardie, C. (2017). Mpqa opinion corpus. In *Handbook of Linguistic Annotation* (pp. 813-832): Springer.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). *Recognizing contextual polarity in phrase-level sentiment analysis.* Paper presented at the Proceedings of the conference on human language technology and empirical methods in natural language processing.

Witteman, H. O., & Zikmund-Fisher, B. J. (2012). The defining characteristics of Web 2.0 and their potential influence in the online vaccination debate. *Vaccine, 30*(25), 3734-3740.

Wyner, A. Z., Schneider, J., Atkinson, K., & Bench-Capon, T. J. (2012). Semi-Automated Argumentative Analysis of Online Product Reviews. *COMMA, 245*, 43-50.

Xiao, L. (2018). *A Message's Persuasive Features in Wikipedia's Article for Deletion Discussions.* Paper presented at the Proceedings of the 9th International Conference on Social Media and Society.

Xu, W. W., Chiu, I.-H., Chen, Y., & Mukherjee, T. (2015). Twitter hashtags for health: applying network and content analyses to understand the health knowledge sharing in a Twitter-based community of practice. *Quality & Quantity, 49*(4), 1361-1380.

Yancheva, M., & Rudzicz, F. (2013). *Automatic detection of deception in child-produced speech using syntactic complexity features.* Paper presented at the Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Yang, K. C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, e115.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine, 13*(3), 55-75.

Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., & Blackburn, J. (2018). Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. *arXiv preprint arXiv:1801.09288*.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group decision and negotiation, 13*(1), 81-106.

Zollo, F., Novak, P. K., Del Vicario, M., Bessi, A., Mozetič, I., Scala, A., . . . Quattrociocchi, W. (2015). Emotional dynamics in the age of misinformation. *PloS one, 10*(9), e0138740.

APPENDIX A. LEXICONS

Table 31. Feature types used in the Model

Type	Feature	Description
Basic Features	Unigram	Word count for each single word that appears in the tweet
	Bigram	Word count for each two words that appears in the tweet
Psychometrics Features	Perceptual process	Percentage of words that refers to multiple sensory and perceptual dimensions associated with the five senses.
	Biological process	Percentage of words related to body, health, sexual and Ingestion
	Core Drives and Needs	Percentage of words related to personal drives as power, achievement, reward and risk
	Cognitive Processes	Percentage of words related to causation, discrepancy, tentative, certainty, inhibition and inclusive.
	Personal Concerns	Percentage of words related to work, leisure, money, death, home and religion
	Social Words	Percentage of words that are related to family and friends
Linguistic Features	Analytical Thinking	Percentage of words that captures the degree to which people use words that suggest formal, logical, and hierarchical thinking patterns
	Clout	Percentage of words related to the relative social status, confidence, or leadership that people display through there writing or talking.
	Authenticity	Percentage of words that reveals people in an authentic or honest way, they are more personal, humble, and vulnerable
	Emotional Tone	Percentage of words related to the emotional tone of the writer which is a combination of both positive emotion and negative emotion dimensions.
	Informal Speech	Percentage of words related to informal language markers as assents, fillers and swears words
	Time Orientation	Percentage of words that refer to Past focus, present focus and future focus.
	Grammatical	Percentage of words that refer to personal pronouns, impersonal pronouns, articles, prepositions, auxiliary verbs, common adverbs, punctuation
	Positive emotion	Percentage of positive words in a sentence
	Negative emotion	Percentage of negative words in a sentence
	Subjectivity type	Subjectivity type derived by Wilson et al. (2005) lexicon
	Punctuation	Percentage of punctuation in text including periods, commas, colons, semicolons etc.
Twitter-specific Features	RT	1.0 if the tweet is a retweet
	Title	1.0 if the tweet contains a title to the article title
	Mention	1.0 if the tweet contains a mention to another user '@'
	Verified account	1.0 if the author has a 'verified' account
	URL	1.0 if the tweet contains a link to a URL
	Followers	Number of people this author is following at posting time
	Following	Number of people following this author at posting time
	Posts	Total number of user's posts
	hashtag	1.0 if the tweet contains a hashtag '#'
	WC	Word count of the tweet
	Words>6 letters	Count of words with more then six letters
	WPS	Count of words per sentence

Table 31 (cont.)

Type	Feature	Description
	QMark	Percentage of words contains question mark
	Exclam	Percentage of words contains exclamation mark

Table 32. User Categorization Code Book

	Dimension	Factor	Definition	Location
Individuals	Occupation	Health Professional	An individual who provides/is involved in health care services (doctor, physician, nurse, etc.)	Bio (Linked URL), Username
		Blogger	A personal log of thoughts published on a Web page or on the Twitter account itself and is maintained by one person.	Bio (Linked URL), Username
		Celebrity	An individual notable for their fame (musician, actor, entrepreneur, writer etc.)	Bio (Linked URL), Verification Status
		Health Writer	Writers who can be considered experts from taking part in published works such as books, papers, magazines, etc. Should write about health topics i.e. vaccines	Bio (Linked URL), Verification Status
		Company Representative	An individual authorized to act on behalf an organization (publicist, advertiser, promoter, etc.)	Bio (Linked URL)
		Health Activist	An individual who campaigns to bring about political or social change in the health field.	Bio (Linked URL)
		Lay Person	An individual who discusses health topics with no occupation involving the topic discussed. Does not have any other online content beside their tweets. i.e. do not have a blog	Bio (Linked URL)
		Other	Cannot be identified in a category within occupation.	
Organizations	**Dimension**	**Factor**	**Definition**	**Location**
	Media	Official	Well-known, credible news media sources such as news channels (BBC, CNN) and health magazines.	Username, Verification Status, Bio
		Unofficial	Private news outlets run by individuals/small groups of people. (Health Standards, personal blogs, etc.)	Username, Verification Status, Bio

Table 32 (cont.)

	Dimension	Factor	Definition	Location
		Blog	A regularly updated website or web page, run by a group of individuals, usually written in an informal or conversational style.	Username, Bio (Linked URL)
		Social Media Profile	An account that serves to provide tips and information related to health but is not linked to any other sources such as a blog. The only source of content comes from the account itself.	Username, Bio (Linked URL), Timeline
	Educational	Universities	Institution of Higher Education.	Username, Verification Status, Bio
	Hospitals	Hospitals	Health Care Institution.	Username, Verification Status, Bio
	Governmental	Government	Organization active in the field of global health research; dependent of the government	Username, Verification Status, Bio, URL domain name
	Nonprofits	Societies	A non-profit organization that specializes in improving healthcare.	Username, Verification Status, Bio
		Foundations	An organization designed to strive towards new research opportunities within a field in order to bring about improvement.	Username, Verification Status, Bio
		NGO’s	Organization active in the field of global health research; independent of the government.	Username, Verification Status, Bio, URL domain name
	Health Corporations	Health-Care Companies	A profitable organization that provides health programs and services for all. (CVS Health, WebMD, etc.)	Username, Verification Status, Bio, URL domain name
	Other	Other	Cannot be identified in a category.	

Table 33. URL Categorization Code Book

Category	Description	Examples	Count
News	An official broadcast or published report of current events, important information, etc.	Nytimes.com, Cnn.com, Theguardian.com	1068
Blogs	A regularly updated website or web page, typically one run by an individual or small group, that is written in an informal or conversational style.	Blogspot.com, Wordpress.com, Tumblr.com	62
Scientific	Scholarly publications, research, articles or journals that contain credible scientific data/information.	Bmj.com, Elsevier.com, Nature.com	1858
Federal, State or Local Government agencies	A government or state agency, often an appointed commission, is an organization in the machinery of government that is responsible for the oversight and administration of specific functions.	Nih.gov, Cdc.gov, medlineplus.gov, Any .gov URL	291
Commercial content	Websites that are available to list and search for properties at no charge but may offer additional services on a fee basis.	Amazon.com eBay.com	65
National or state professional medical societies and associations	An official national or state organization representing a particular group of medical professionals.	Mayoclinic.org kaiserpermanente.org Apa.org	133
Commercial subsidized health websites	For-profit organization or corporation. It usually includes advertising	webmd.com	1
Health magazine websites	Magazines that cover a variety of topics including physical fitness and well-being, nutrition, beauty, strength, bodybuilding, weight training, etc.	Health.com Prevention.com Self.com	46
Health Insurance	An insurance that covers the whole or a part of the risk of a person incurring medical expenses.	molinahealthcare.com assuranthealth.com unicare.com	22
Videos	Video sharing services where users can watch, share, upload their own videos, etc.	YouTube.com Vimeo.com	13
Social Media	Websites that enable users to create and share content or to participate in social networking.	Facebook.com Twitter.com Reddit.com Pinterest.com	87
Educational institutions	Websites that represents a place where people of gain an education.	Berkeley.edu Colorado.edu Sdsu.edu Any .edu URL	-
Fake News	A type of journalism or propaganda that consists of deliberate misinformation or hoaxes.	100percentfedup.com 21stcenturywire.com	834

APPENDIX B. WEBSITE CREDIBILITY QUESTIONS

To identify if a website contains accurate information, here are five questions that needs to be answered:

1. **Who runs or created the site? (look for the "about us" page)**
    a. For-profit organization/person
    b. Not for profit organization
    c. Not clear who created the website

2. **When was its information written or reviewed?**
    a. The information is up to date
    b. The information is not up to date
    c. Not sure when was the information posted.

3. **Where does the information come from?**
    a. Fact-based source (based on scientific research)
    b. Opinion- based source
    c. Not clear where did the information come from

4. **Why does the site exist? Who pays for the site?**
    a. Advertisements with disclosure
    b. Advertisements without disclosure
    c. No Advertisements

5. **Is there a contact information?**
    a. Yes, there is.
    b. No, there is not.

6. **Is this a health-related website?**
    a. Yes, it is
    b. No, it is not
    c. Not clear what the website is providing

7. **What is the sentiment of the website?**
    a. Pro-vaccine
    b. Anti- vaccine
    c. Neutral
    d. Not clear/ unrelated

APPENDIX C. DATA COLLECTION INSTRUMENTS AND MATERIALS

Consent Form (MTurk participants)

Survey on Health Controversial Topics

Background:

In our daily social interactions, we discuss different topics. Some are considered to be controversial and some are not. In this survey, we are trying to understand and learn more about people's thoughts and opinions regarding what constitutes a controversial health issue. Your participation in this survey will help in understanding what issues the public considers to be controversial vs. not. This will help policy makers make accurate decisions regarding these issues. Moreover, researchers need more insights from the public to understand what aspects of a controversy need more exploration.

Please read the following consent form and click submit if you agree with it.

Voluntary Consent:

You are invited to participate in a research study on understanding the public opinions regarding controversial health topics. This study is conducted by Professor Jana Diesner and her research team. We are working at the School of Information Sciences at the University of Illinois at Urbana-Champaign. Benefits of participating include approximately $0.48-$0.96 US dollar payment from Amazon. At the end of this study, you will receive a code that you need to obtain to receive the payment. Your decision to participate or decline participation in this study is completely voluntary and you have the right to terminate your participation at any time without penalty. You may skip any questions you do not wish to answer. If you do not wish to complete this survey, please just close your browser.

About the Survey:

This study will take approximately **4-8 minutes** of your time. You will be asked to complete an online survey about your opinion on controversial topics. Your participation in this research will be completely anonymous and the data will be averaged and reported in aggregates. Possible outlets of dissemination may be academic journals or research conferences. We will use all reasonable efforts to keep your personal information confidential, but we cannot guarantee absolute confidentiality. When this research is discussed or published, no one will know that you were in this study. But, when required by law or university policy, identifying information (including your signed consent form) may be seen or copied by:

1. The Institutional Review Board that approves research studies;
2. The Office for Protection of Research Subjects and other university departments that oversee human subjects research;
3. University and state auditors responsible for oversight of research.

Your participation in this research may not benefit you personally, but it will help in understanding the public opinions regarding controversial health topics. To the best of our knowledge, participating in this survey have no more risk of harm than you would experience in everyday life.

<u>Questions?</u>

If you have any questions regarding the survey please contact the research team via email at aaddaw2@illinois.edu. If you have any questions about your rights as a participant in this study or any concerns or complaints, please contact the University of Illinois Office for the Protection of Research Subjects at 217-333-2670 or via email at irb@illinois.edu. Please print a copy of this consent form for your records, if you so desire.

I have read and understand the above consent form, I certify that I am 18 years old or older and, by clicking the submit button to enter the survey, I indicate my willingness voluntarily take part in this study.

SUBMIT

Consent Form (Social media participants)

Survey on Health Controversial Topics

Background:

In our daily social interactions, we discuss different topics. Some are considered to be controversial and some are not. In this survey, we are trying to understand and learn more about people's thoughts and opinions regarding what constitutes a controversial health issue. Your participation in this survey will help in understanding what issues the public considers to be controversial vs. not. This will help policy makers make accurate decisions regarding these issues. Moreover, researchers need more insights from the public to understand what aspects of a controversy need more exploration.

Please read the following consent form and click submit if you agree with it.

Voluntary Consent:

You are invited to participate in a research study on understanding the public opinions regarding controversial health topics. This study is conducted by Professor Jana Diesner and her research team. We are working at the School of Information Sciences at the University of Illinois at Urbana-Champaign. Your decision to participate or decline participation in this study is completely voluntary and you have the right to terminate your participation at any time without penalty. You may skip any questions you do not wish to answer. If you do not wish to complete this survey, please just close your browser.

About the survey:

This study will take approximately **4-8 minutes** of your time. You will be asked to complete an online survey about your opinion on controversial topics. Your participation in this research will be completely anonymous and the data will be averaged and reported in aggregates. Possible outlets of dissemination may be academic journals or research conferences. We will use all reasonable efforts to keep your personal information confidential, but we cannot guarantee absolute confidentiality. When this research is discussed or published, no one will know that you were in this study. But, when required by law or university policy, identifying information (including your signed consent form) may be seen or copied by:

4. The Institutional Review Board that approves research studies;
5. The Office for Protection of Research Subjects and other university departments that oversee human subjects research;
6. University and state auditors responsible for oversight of research.

Your participation in this research may not benefit you personally, but it will help in understanding the public opinions regarding controversial health topics. To the best of our knowledge, participating in this survey have no more risk of harm than you would experience in everyday life.

<u>Questions?</u>

If you have any questions regarding the survey please contact the research team via email at aaddaw2@illinois.edu. If you have any questions about your rights as a participant in this study or any concerns or complaints, please contact the University of Illinois Office for the Protection of Research Subjects at 217-333-2670 or via email at irb@illinois.edu. Please print a copy of this consent form for your records, if you so desire.

I have read and understand the above consent form, I certify that I am 18 years old or older and, by clicking the submit button to enter the survey, I indicate my willingness voluntarily take part in this study.

SUBMIT

Survey Questions

Q1: For the purpose of this survey and according to previous research, **Controversial topics** are those that generate disagreement or different opinions among large groups of people [1].

Example of controversial topics include:

1. Puppies are cuter than kittens.
2. Existence of climate change.

Please keep this definition in mind while answering the questions in this survey.

[1] Dori-Hacohen, S., Yom-Tov, E., & Allan, J. (2015). Navigating Controversy as a Complex Search Task. In SCST@ ECIR.

Q2: In your opinion, list one to three health-related topics that you think are **controversial.** These topics can be personal and/or public health issues.

____

____

____

____

____

Q3: In your opinion, list one to three health-related topics that you think are **NOT controversial.** These topics can be personal and/or public health issues.

____

____

____

Q4: Based on your opinion and using the scale below, please rate **how controversial** you believe the following statements to be. (*This is independent of whether you agree or not to the sentences*)

	Not at all controversial	Barely controversial	Moderately controversial	Very controversial	Extremely controversial
“Medical marijuana should be legalized”	❍	❍	❍	❍	❍
“The Measles, Mumps, and Rubella (MMR) vaccine can cause autism”	❍	❍	❍	❍	❍
"Legal abortion performed by a qualified medical professional should be a foundational right for women"	❍	❍	❍	❍	❍
“Human Immunodeficiency Virus (HIV) causes Acquired Immunodeficiency Syndrome (AIDS)”	❍	❍	❍	❍	❍
"Exposure to sunlight and/or taking vitamin D supplements is important for human health"	❍	❍	❍	❍	❍
“Adequate sleep at night is important for maintaining a healthy body and mind”	❍	❍	❍	❍	❍

Q5: Based on your opinion and using the scale below, please rate <u>how controversial</u> you believe the following statements to be. (*This is independent of whether you agree or not to the sentences*)

	Not at all controversial	Barely controversial	Moderately controversial	Very controversial	Extremely controversial
“Peoples' health can benefit from moderate exercise”	❍	❍	❍	❍	❍
For validation purposes: Please select Slightly Controversial button	❍	❍	❍	❍	❍
“A balanced and nutritional diet is beneficial for peoples' well-being"	❍	❍	❍	❍	❍
“E-cigarettes are less harmful than smoking tobacco”	❍	❍	❍	❍	❍
“Proper oral hygiene (i.e. flossing and brushing) is key to keeping a healthy and bright smile throughout adulthood”	❍	❍	❍	❍	❍

Q6: The following set of questions is not about the controversiality of the issues, it is about **your opinion**.

Q7: How much do you support or oppose legalizing the use of marijuana by adults for medical purposes, if a doctor prescribes it?

- ○ Strongly support
- ○ Moderately support
- ○ Neither support nor oppose
- ○ Moderately oppose
- ○ Strongly oppose

Q8: How important is proper oral hygiene (i.e. flossing and brushing) in keeping a healthy and bright smile throughout adulthood?

- ○ Extremely important
- ○ Very important
- ○ Moderately important
- ○ Slightly important
- ○ Not at all important

Q9: How likely do you think it is that the Measles, Mumps, and Rubella (MMR) vaccine can cause autism in children?

- ○ Extremely likely
- ○ Very likely
- ○ Moderately likely
- ○ Not very likely
- ○ Not at all likely

Q10: How much can peoples' health benefit from moderate exercise?

- ○ A great deal
- ○ A lot
- ○ A moderate amount
- ○ A little
- ○ Not at all

Q11: How much can peoples' health benefit from exposure to sunlight and/or taking vitamin D supplements?

❍ A great deal

❍ A lot

❍ A moderate amount

❍ A little

❍ Not at all

Q12: How important is getting adequate sleep at night for adults to maintain a healthy body and mind?

❍ Extremely important

❍ Very important

❍ Moderately important

❍ Slightly important

❍ Not at all important

Q13: How likely is it that HIV (Human Immunodeficiency Virus) causes AIDS (Acquired Immunodeficiency Syndrome)?

❍ Extremely likely

❍ Somewhat likely

❍ Neither likely nor unlikely

❍ Somewhat unlikely

❍ Extremely unlikely

Q14: How harmful is smoking e-cigarettes compared to smoking tobacco?

❍ E-cigarettes are much more harmful
❍ E-cigarettes are somewhat more harmful
❍ E-cigarettes are about as harmful
❍ E-cigarettes are somewhat less harmful
❍ E-cigarettes are much less harmful

Q15: To what extent do you support or oppose the idea that getting a legal abortion performed by a qualified medical professional should be a basic right for women in this country?

❍ Strongly support
❍ Moderately support
❍ Neither support nor oppose
❍ Moderately oppose
❍ Strongly oppose

Q16: How much can peoples' health benefit from a balanced and nutritious diet?

❍ A great deal
❍ A lot
❍ A moderate amount
❍ A little
❍ Not at all

Q17: **Background Information**

Q18: What is your age in years?

____

Q19: What is your gender?

❍ Male
❍ Female
❍ Other
❍ I prefer not to say

Q20: What is your primary employment status?

❍ Employed for wages

❍ Self-employed

❍ Unemployed

❍ A homemaker

❍ A student

❍ Retired

❍ Other: ____

❍ I prefer not to say

Q21: What is the highest level of education you have completed?

❍ Less than high school degree

❍ High school graduate

❍ Some college or Associate’s degree

❍ 4 year degree or Bachelor's degree

❍ Master’s or Professional degree

❍ Doctorate

❍ I prefer not to say

Q22: Generally speaking, would you describe your political view as:

❍ Very conservative

❍ Somewhat conservative

❍ Moderate

❍ Somewhat liberal

❍ Very liberal

❍ None of these

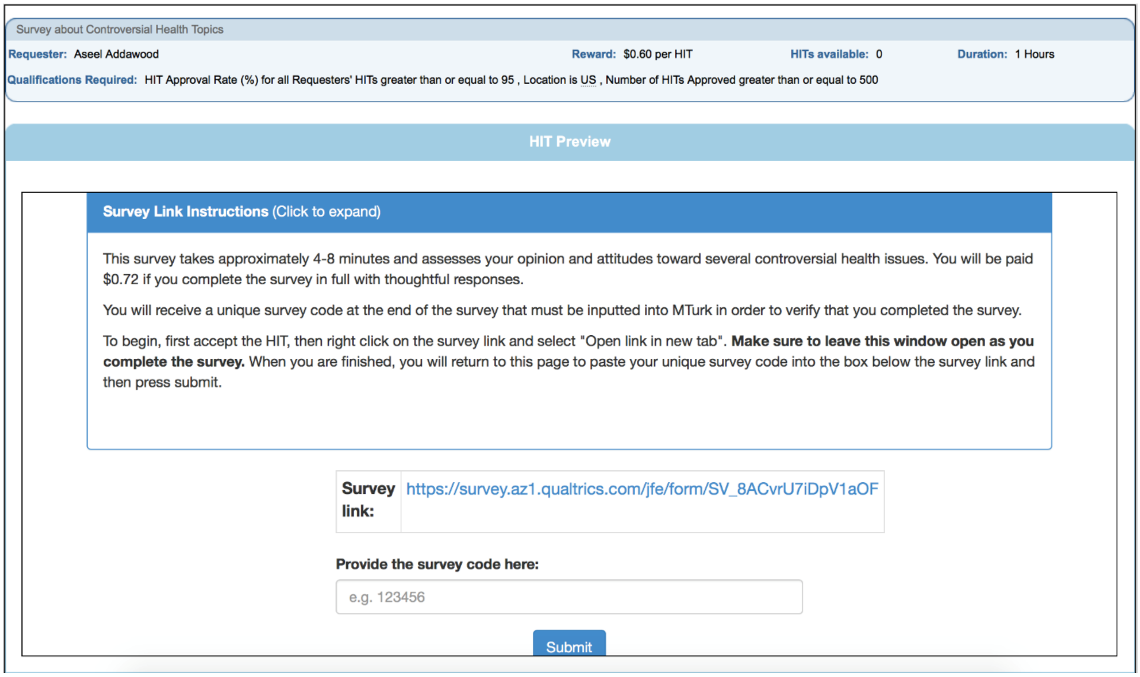
❍ I prefer not to say

Q23: In which state do you currently reside?

❍ I do not reside in the United States
❍ Alabama
❍ Alaska
❍ Arizona

❍ Arkansas
❍ California
❍ Colorado
❍ Connecticut
❍ Delaware
❍ District of Columbia
❍ Florida
❍ Georgia
❍ Hawaii
❍ Idaho
❍ Illinois
❍ Indiana
❍ Iowa
❍ Kansas
❍ Kentucky
❍ Louisiana
❍ Maine
❍ Maryland
❍ Massachusetts
❍ Michigan
❍ Minnesota
❍ Mississippi
❍ Missouri
❍ Montana
❍ Nebraska
❍ Nevada
❍ New Hampshire
❍ New Jersey
❍ New Mexico
❍ New York
❍ North Carolina
❍ North Dakota
❍ Ohio
❍ Oklahoma
❍ Oregon
❍ Pennsylvania
❍ Rhode Island
❍ South Carolina
❍ South Dakota
❍ Tennessee
❍ Texas
❍ Utah
❍ Vermont
❍ Virginia
❍ Washington
❍ West Virginia
❍ Wisconsin
❍ Wyoming

APPENDIX D. MESSAGES USED FOR DATA COLLECTION



Survey about Controversial Health Topics

Requester: Aseel Addawood **Reward:** $0.60 per HIT **HITs available:** 0 **Duration:** 1 Hours

Qualifications Required: HIT Approval Rate (%) for all Requesters' HITs greater than or equal to 95 , Location is US , Number of HITs Approved greater than or equal to 500

Survey Link Instructions (Click to expand)

This survey takes approximately 4-8 minutes and assesses your opinion and attitudes toward several controversial health issues. You will be paid $0.72 if you complete the survey in full with thoughtful responses.

You will receive a unique survey code at the end of the survey that must be inputted into MTurk in order to verify that you completed the survey.

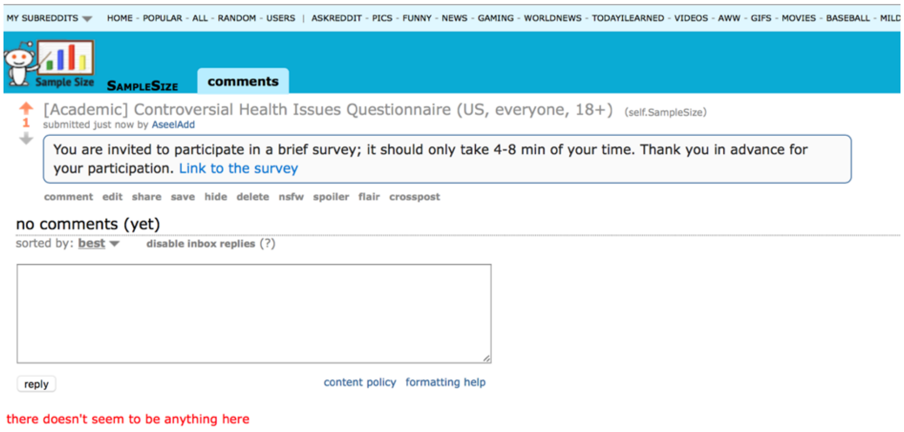
To begin, first accept the HIT, then right click on the survey link and select "Open link in new tab". **Make sure to leave this window open as you complete the survey.** When you are finished, you will return to this page to paste your unique survey code into the box below the survey link and then press submit.

Survey link: https://survey.az1.qualtrics.com/jfe/form/SV_8ACvrU7iDpV1aOF

Provide the survey code here:

e.g. 123456

Figure 32: MTurk HIT



MY SUBREDDITS
HOME - POPULAR - ALL - RANDOM - USERS | ASKREDDIT - PICS - FUNNY - NEWS - GAMING - WORLDNEWS - TODAYILEARNED - VIDEOS - AWW - GIFS - MOVIES - BASEBALL - MILD
Sample Size
SampleSize
comments
1

[Academic] Controversial Health Issues Questionnaire (US, everyone, 18+) (self.SampleSize)

submitted just now by AseelAdd

You are invited to participate in a brief survey; it should only take 4-8 min of your time. Thank you in advance for your participation. Link to the survey

comment edit share save hide delete nsfw spoiler flair crosspost

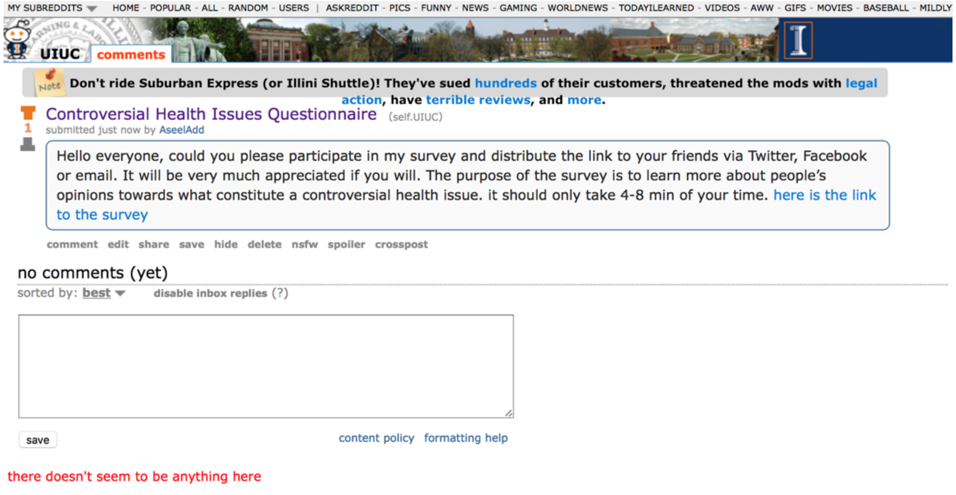
no comments (yet)

sorted by: best disable inbox replies (?)

reply
content policy formatting help

there doesn't seem to be anything here

Figure 33: Sample size subreddit message


MY SUBREDDITS ▾ HOME - POPULAR - ALL - RANDOM - USERS | ASKREDDIT - PICS - FUNNY - NEWS - GAMING - WORLDNEWS - TODAYILEARNED - VIDEOS - AWW - GIFS - MOVIES - BASEBALL - MILDLY
UIUC
comments
I

Don't ride Suburban Express (or Illini Shuttle)! They've sued hundreds of their customers, threatened the mods with legal action, have terrible reviews, and more.

Controversial Health Issues Questionnaire (self.UIUC)

submitted just now by AseelAdd

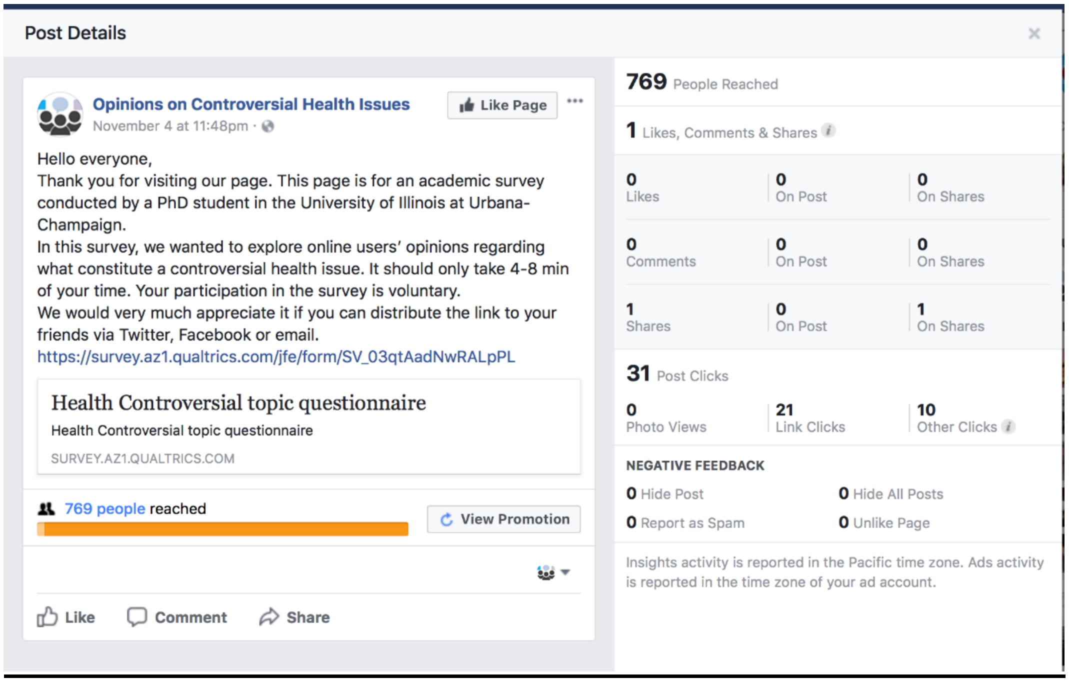
Hello everyone, could you please participate in my survey and distribute the link to your friends via Twitter, Facebook or email. It will be very much appreciated if you will. The purpose of the survey is to learn more about people's opinions towards what constitute a controversial health issue. it should only take 4-8 min of your time. here is the link to the survey

no comments (yet)

sorted by: best ▾ disable inbox replies (?)
save
content policy formatting help

there doesn't seem to be anything here

Figure 34: UIUC subreddit message



Opinions on Controversial Health Issues

November 4 at 11:48pm ·

Hello everyone,
Thank you for visiting our page. This page is for an academic survey conducted by a PhD student in the University of Illinois at Urbana-Champaign.
In this survey, we wanted to explore online users' opinions regarding what constitute a controversial health issue. It should only take 4-8 min of your time. Your participation in the survey is voluntary.
We would very much appreciate it if you can distribute the link to your friends via Twitter, Facebook or email.
https://survey.az1.qualtrics.com/jfe/form/SV_03qtAadNwRALpPL

Health Controversial topic questionnaire

Health Controversial topic questionnaire

SURVEY.AZ1.QUALTRICS.COM

769 people reached

769 People Reached

1 Likes, Comments & Shares

0 Likes	**0** On Post	**0** On Shares
0 Comments	**0** On Post	**0** On Shares
1 Shares	**0** On Post	**1** On Shares

31 Post Clicks

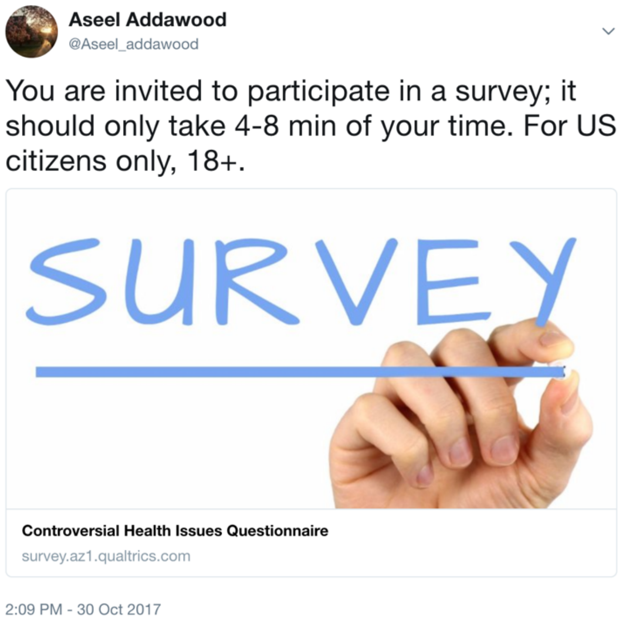
0 Photo Views	**21** Link Clicks	**10** Other Clicks

NEGATIVE FEEDBACK

0 Hide Post	**0** Hide All Posts
0 Report as Spam	**0** Unlike Page

Insights activity is reported in the Pacific time zone. Ads activity is reported in the time zone of your ad account.

Figure 35: Facebook advertising and its results



Aseel Addawood
@Aseel_addawood

You are invited to participate in a survey; it should only take 4-8 min of your time. For US citizens only, 18+.

Controversial Health Issues Questionnaire
survey.az1.qualtrics.com

2:09 PM - 30 Oct 2017

Figure 36: Twitter advertising

APPENDIX E. PUBLICATIONS

The research reported in this dissertation has contributed to the following publications:

1. **Addawood, A**. (2018). Scientific Credibility Behind MMR Vaccination Debates on Twitter. In the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2018). Barcelona, Spain.
2. **Addawood, A**., Alshamrani, A., Alqahtani, A., Diesner, J. & Broniatowski, D. (2018). Women's Driving in Saudi Arabia – Analyzing the Discussion of a Controversial Topic on Twitter. In International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS 2018). Washington, DC.
3. **Addawood A**, Schneider J, Bashir M. (2017). Stance Classification of Twitter Debates: The Encryption Debate as A Use Case. In the Proceedings of the International Conference on Social Media & Society (#SMSociety17). Toronto, Canada, July 28-30.
4. **Addawood A**, Bashir M. (2016). What is Your Evidence? A Study of Controversial Topics on Social Media. In the Proceedings of the third workshop on argumentation mining. ACL.

APPENDIX F. ADDITIONAL MATERIALS

Code

All the code used in this thesis is written in Python and can be found in this GitHub repository (https://github.com/aseelad).

Datasets

All datasets will be available through my website:
https://sites.google.com/view/aseeladdawood/