ASSESSMENT OF INFORMATION CREDIBILITY IN TWITTER

by

KYLE MAURICE THOMPSON

(Under the Direction of Ismailcem Budak Arpinar)

ABSTRACT

In today's society, social media has become one of the primary sources of obtaining information. Citizens utilize social media platforms to familiarize themselves with and learn about, current events around the world. There is an increase in individuals that depend on social media to be valid and provide correct information on newsworthy topics. This research provides a unique automated approach that incorporates human intelligence and text mining techniques. The phrase "What vs. Who" is used to separate the past research from this approach. The automated approach includes techniques similar to text mining which uses Python's Natural Language Toolkit to incorporate natural language processing and search through individual tweets for certain keywords. To provide credible information, news articles, verified Twitter accounts, and criminal and domestic release records, which possess official power, were used as the golden standard and deemed credible. Crowd sourcing was then used to compare results with the automated approach.

INDEX WORDS: Social Media, Twitter credibility

ASSESSMENT OF INFORMATION CREDIBILITY IN TWITTER

by

KYLE MAURICE THOMPSON

BS, Hampton University, 2012

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2014

ASSESSMENT OF INFORMATION CREDIBILITY IN TWITTER

by

KYLE MAURICE THOMPSON

Major Professor:	Ismailcem Budak Arpinar
Committee:	E. Rodney Canfield
	Don Potter

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
December 2014

ACKNOWLEDGEMENTS

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of this project. I am thankful for their guidance, constructive criticism and friendly advice during the project.

I express my warm thanks to Dr. I. Budak Arpinar, Mr. Nilayan Bhattacharya, and Mr. Sanmit Tatoba Desai for their support at The University of Georgia. I also want to express my warm thanks to Cortney Seaborne and Kamau Hull for proofreading this thesis.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF EQUATIONS

Page

INTRODUCTION

In the 21st century, social media networks have emerged and supplements traditional media. Traditional media includes television and print media as the primary source for obtaining news and information. Journalists and news agencies have evolved from printing news to posting or tweeting news on social networks: "*Social media provides a platform for users to create content, but also to discuss that content in a collaborative effort to create better content, and to come to a shared understanding of the content that is created"* (O'Reilly & Battelle, 2009). Social networks provide a platform for citizens to express themselves.

As a result of the emergence of news via social platforms, there has been an increased concern pertaining to the reliability of social networks such as Twitter, Facebook, and blogs. There is a need for users and journalists to quickly identify false information, fake pictures, tampered videos, and false evidence from users. Jason Whittaker states, "*Without accuracy, verification, fairness, evidence-based statements and proper attribution citizen sourcing is not accomplishing the goals of journalism, which is to report the truth*" (Mabweazara, Mudhai, & Whittaker, 2014). Society must establish a way to determine the truth without biases and provide that truth to the public.

Since Twitter has gained popularity as a major news source and information dissemination agent, State and National Representatives use Twitter to ensure that the United States has an open government. For example, immediately after Arizona's Governor Jan Brewer vetoed bill SB 1062 (Legislature & Session, 2011) on February 26, 2014, she posted her decision on Twitter. This information spread until everyone knew the outcome and reports dispersed regarding the decision. Like Governor Jan Brewer, many officials have Twitter accounts to

provide valid information to the public as means of transparency. Figure 1 displays elected officials Barack Obama and Jan Brewer who have Twitter accounts to provide communication to the public. The emergence of officials on social media has become prevalent as the government makes its information visible to the public.

Figure 1: Barack Obama and Jan Brewer's Twitter Profile Overview

TWITTER BACKGROUND INFORMATION

Twitter is a micro-blogging service that accounts for millions of users from around the world. It allows users to post a 140-character-long message, which is called a tweet (Chernin, 2014). Tweets can be published via email, SMS message, or third-party applications. Twitter website explains that, "*Twitter can be used through various clients: 78% of active users correspond with Twitter via mobile phones. Users tweet an average of 500 million tweets per day and all information is real-time dispersion of information"* (Twitter, 2014c). Twitter has characteristics of high response time, high connectivity speeds, and the ability to direct message users. Over 650 million users across the world are able to connect to this social media platform at high speeds, rapidly refresh for updated tweets, and message between users in seconds. Therefore, Twitter is the ideal application to disseminate breaking news from the news source or

the news agency. Nearly all news agencies and journalists have a Twitter account. These users provide insight into their opinions about news reports and provide clarity on the subjects.

Having a Twitter account promotes popularity and disperses information rapidly throughout the world. According to Pew Research Center in 2013, over 30% of the general population gets their news from social networks. 16% (50,240,000) of United States citizens have Twitter accounts and 8% (25,120,000) get their news from Twitter (Jesse Holcomb, 2013). These numbers are continuously increasing as more people join Twitter and rely on social networks as their primary source of news. Social networks have emerged as a way to obtain information about the world: 65% of social media users only use one platform when reading the news and 35% use more than one site to verify the story (Gahran, 2011). Since the public generally uses one site to get their information, there must be an automated way to deem whether this information credible or not.

RELATED WORKS

Due to the increase in utilizing social media for news, credibility has become a common interest in research. There is existing research pertaining to trusting sources and determining the credibility of a tweet. Existing methods have been used to rank, verify, or predict the credibility of a user or tweet. This related works section is categorized into two sections: (1) credibility prediction research, and (2) crowd sourcing heuristics.

Prediction

Twitter's characteristics make credibility an ideal topic to conduct research. Researchers generally create complex feature-based equations using meta-data to determine the validity of a tweet. Canini presents research to compute credible information in Twitter. Topic-specific

credibility was used to define a ranking strategy for users based on their relevance and expertise (Canini, Suh, & Pirolli, 2011). The author concluded that there is great potential for automatic identification and ranking of credible users for any given topic. Later in Canini's research, he studied the effects of text variance on perceived credibility, claiming that using text similarity would provide good credibility results. Mendoza also evaluated trust when disseminating information through Twitter (Mendoza, Poblete, & Castillo, 2010). The research focused on Chile earthquakes in 2010 and determined that through the method of aggregate analysis of tweets, rumors can be successfully detected.

Castillo frequently conducts research on credibility of Twitter's newsworthy topics (Castillo, Mendoza, & Poblete, 2011). In his research, Castillo provides a method that predicts tweet's credibility by using a complex set of features that employ messages, topics, propagation and users. Tweets were collected that had a short burst of activity, specifically trending topics, and mined word co-occurrences from the tweets. The classifier analyzed credibility and calculated a 60-70% increase in precision versus manually labeled credibility data.

Kang presents two models of determining credibility using Twitter, which are the social model and the content model (Kang, O'Donovan, & Höllerer, 2012). The social model utilizes a weighted combination of positive credibility indicators from the underlying social network, or "re-tweets". The content model uses a probabilistic language-based approach for identifying patterns of terms and other tweet properties that lead to positive feedback, for example sentiment, the number of swear words, the number of URLs, or whether a tweet has question marks or emoticons, etc. Results show that the social model outperforms the content model by a significant margin. An equation regarding re-tweets using the social model predicted results of

88.17% as opposed to 62% from the content model. Therefore, Kang concludes that number of re-tweets is the predominant feature for a credible tweet.

Crowd Sourcing Using Heuristics

When defining a tweet's credibility score, there must be a way to check if the content of a tweet is correct. Some researchers score credibility using heuristics instead of undergoing the many steps of determining the credibility of a tweet. Researchers trust users to determine the correctness of information by using common sense. Shariff uploaded 400 tweets to the application CrowdFlower to utilize crowd sourcing techniques (Shariff, Zhang, & Sanderson, 2014). The author concluded that the crowd is a good way to verify credibility, but most tweets lack links. The lack of links makes it difficult to prove truthfulness and provides inconsistent scores while judging tweets. Display name and presence of links in a tweet were two features that were given by the workers in CrowdFlower that factored into credibility assessment.

Schmierbach displayed college students' articles on The New York Times website and tweets from the verified The New York Times Twitter page that presented the same information (Schmierbach & Oeldorf-Hirsch, 2012). Students rated Twitter's information less credible than from the website despite there being no clear indication of what led to such determination. Morris conducted similar experiments to determine why users felt Twitter information was not credible (Morris, Counts, Roseway, Hoff, & Schwarz, 2012). The researcher's objective was to determine what features were more prevalent in determining the credibility of a tweet. Accounts were made and features were manipulated to display different types of usernames, links, and pictures with similar content. A survey was then conducted to determine the amount of concern for credibility for each feature. Although the researcher found what features were prevalent, the survey revealed that users are poor judges of truthfulness based on content alone, and that users

were influenced by heuristics relying on usernames when making a decision. Westerman conducted related research but concentrated on friend and follower counts (Westerman, Spence, & Van Der Heide, 2012). Three Twitter accounts were made with various friend and follower count information. These friends and follower count were manipulated to determine the most believable account. Results illustrated that users found Twitter accounts with a narrow ratio of friends and followers were credible, and that large Twitter communities don't operate correctly, and raise suspicion in credibility.

CITIZEN JOURNALISM

Citizen journalism is defined as the "*collection, dissemination, and analysis of news and information"(Dictionary, 2014).* This type of journalism includes "*practices such as current affair-based blogging, photo and video sharing, and posting eyewitness commentary on current events, as well as activities such as re-posting, linking, tagging and modifying, or commenting on news materials posted by other users or by professional news outlets*" (Goode, 2009). Today, many people participate and rely on citizen journalism by simply posting on their social networks. Opinions are voiced and moods are expressed while posting on particular topics. Pictures and videos are uploaded to report information or crimes that are occurring near the user. With the emergence of smart phones with video recorders, all types of situations are being recorded to ensure firsthand accounts are captured without biases. When an incident happens, reporters aren't there to record the occurrence. Reporters rely on surveillance cameras and citizens to take pictures and provide information to the reporter. Citizens have the power to report and disperse information when traditional journalists are not present. Additionally, citizen journalism is also used to determine the public's opinion or sentiment of topics in the media.

Any group of users who support a position on a topic have the potential to gain attention through social media.

Wikipedia has become the largest example of citizen journalism (License, 2014). The web application facilitates many-to-many communication among users editing articles that work towards maintaining a truthful, neutral point of view. It allows users to add, modify, or delete information in collaboration with others. Information and pictures are dispersed while users edit this information and therefore increasing the credibility. This, in turn, increases the knowledge and evidence users possess about the topic or incident. With Wikipedia, citizens have come to an overall consensus of the information that is essential and credible. J.D. Lasica (Lasica, 2003) classified citizen journalism into six types:

1. Audience participation: The audience attaches comments to news stories, social networks, personal blogs, photos, videos, or local news written by residents of a community;
2. Independent news and information websites: News agencies such as Consumer Reports or CNET that give opinions on topics or merchandise;
3. Full-fledged participatory news sites: Websites that only rely on users articles and don't have any editors or community to aid in editing the information;
4. Collaborative and contributory media sites: Sites that allow users to post articles and gives users the ability to work as a community to provide correct information;
5. Other kinds of "thin media": Release of newsletters or any email mailing list to disperse information; and
6. Personal broadcasting sites: Personal video podcast or website that allows citizens to record video on any topic.

Citizen sourcing is a broad topic, but all categories are based on the public's opinion. Although citizen sourcing implements ways for the citizen to express him or herself, it also gives vulnerabilities of providing false information. If majority believes a false claim, rumors can be dispersed as the truth. In this research, we will study information credibility aspect of citizen sourcing in Twitter.

USING SOCIAL MEDIA AS A PRIMARY SOURCE OF INFORMATION

Using the Internet for news is a very tedious task. There are a lot of misleading blogs, comments, articles, and social media posts. One must carefully read and understand the source of information. Biases come involuntarily and sometimes cannot be controlled. Huckeryby stated:

> *"The web is not some orderly library, tended with care by diligent scholars; sometimes it's more akin to a garbage heap. As a reporter, you need to be an expert at sifting out the rubbish, and at detecting the dubious claims and biased presentations that abound"* (Huckerby, 2006).

Information provided on the Internet is not always true, and Huckerby highlighted the issue of credibility of social networks, news agencies, and blogs with his statement. The trustworthiness of events has become a big issue. Certain news companies have hired teams that sift through social media platforms to identify trends. By having employees to monitor trends, they are able to determine if news companies have a story that have caught the public's attention, and discover breaking news with specific grounds of credibility. Webster's dictionary defines credible as "able to be believed, reasonable trust or believe, offering reasonable grounds for being believed" (Dictionary, 2014). According to the definition, credibility methods must present reasonable grounds for a certain message or tweet to be considered true. In order for a story to be created and dispersed through news agencies, it must be credible. Journalists from the agency must

interview the primary source and report the story. The news agency will also begin to investigate by interviewing witnesses to the event for multiple perspectives. Police officials receive information regarding an event and make the information available to the public. Journalists, news reporters, and news agencies take advantage of primary sources. The primary source is usually a person who has witnessed the mishap or event firsthand. Mabweazara emphasized, "Sources of information are the lifeblood of journalist" (Mabweazara et al., 2014). Without proper sources, the event lacks trustworthiness, thus making the reporting less credible. Journalists usually prefer more than one primary source in order to corroborate the report or determine any inconsistencies in the story.

REASON FOR USING SOCIAL MEDIA

Citizen sourcing exercises a First Amendment right of the United States of America Constitution. The First amendment

> *"Prohibits the making of any law respecting an establishment of religion, impeding the free exercise of religion, abridging the freedom of speech, infringing on the freedom of press, interfering with the right to peaceably assemble or prohibiting the petitioning for a governmental redress of grievances"* ("Constitution," 1789).

Citizens are given the freedom of speech and press to express their opinions. The difference between past generations and modern society is the emergence of social media platforms. Today, even in third world countries, except Iran and China, users can express their opinions and make them visible to everyone in the world via Twitter. If the user does not want to make their identities public out of fear for the consequences, they can post anonymously. Expressing ones opinion or providing news anonymously helps ensure the safety of users.

Social media account holders wish to connect and reconnect with potential or past friends. Friends may relocate or contact information may get lost. With social media, names can be searched and friends can be found through networks (school, class, company, etc). Some users wish to be a part of a community or organization on a social media platform. Groups or organizations can be created to give the social media platform more of a community effect. This provides a community inside of a community where it offers an enhanced select group of friends. Some users desire to promote themselves in various ventures. For artists or entertainers, social media allows users to release information to assist in popularity and success. These users can post their latest endeavors and material to ensure exposure to the public (Mueller, 2014). Social media gives users the ability to promote their talents and market their product.

Companies and organizations use social media to showcase their brands. They create social media accounts to ensure that their company is keeping up with the latest technology, and technology appeals to all consumers. By using social media accounts, the company is promoting themselves, increasing digital exposure, awareness, and profit for the company (Angelova, 2013). Instead of allocating funds for marketing, companies can save money by advertising on social media. By utilizing social media, it develops a loyal community of prospects. Companies that have social media accounts provide the feeling of open communication to the company. Organizations have the ability to post activities highlighting their participation in the community, allowing consumers to connect with the company on a personal level.

LAYERS OF SOCIAL MEDIA

Steve Outing notes 11 layers that citizen journalism utilizes (Outing, 2005). These layers are the outline of social media, and the purpose of social media platforms. Of the 11 layers of citizen journalism, the relevant layers of social networks concerning Twitter are:

1. The ability to allow the public to comment: Twitter gives users the option to comment on any tweet of choice. Any user is able to create a new tweet and mention a person to ensure the message appears in their notifications. Another way to comment is simply clicking the reply button to directly answer, or add information, to any tweet.
2. Add-on to the report/tweet: Users can reply and add supporting information if they choose. More information can be shown to the originator of the tweet to provide better details of the event. Tweets can be edited in order to correctly report an event. A link can be provided where users have more characters to publish information.
3. Practice open-source reporting: Twitter is an open-source social media platform in which everyone can view your micro-blog if the user enables the account to be public. Anyone is authorized to have a Twitter account and can report on their topic of choice. Users can help provide tips on crimes, eyewitness information, and personal opinion.
4. Blog: Twitter is a micro-blog that allows up to 140 characters to post information. Anyone with an account can comment or create a story on any subject. A blog is a discussion or informational site where a writer or group of writers can discuss a topic.
5. Enable transparency: Government and company officials use social media to increase the transparency of the company or government. This is keeps the public aware of events happening on a daily or weekly basis. Stockholders take advantage of the transparency to help sell and buy stocks in specific companies.

With these avenues of citizen sourcing, all social networks have taken advantage of the preceding layers. These layers serve as the foundation for social media platforms operations and goals assessments. Layers are used to better understand what social media should be

accomplishing. Without the layers, a social media platform would not excel and would be used solely for communicating with friends.

WHY DO WE TWEET

There are millions of posts on social networks daily. There are 8 reasons why users post, write, or communicate via social networks (Schweidel, 2014). These reasons are:

1. Concern for other consumers: This occurs when users want to warn other consumers of a bad product or event so others will not be put in the same situation. For example, when a consumer dislikes a product they are inclined to spread their opinion about their experience with the product.
2. Platform assistance: Users feel as if social platforms are an advantage to expressing opinions about a company. Customers believe that higher management in the company will read their post. Using social media has the potential to gain support of others that have the same issue. Social media platforms are more convenient than writing to or calling the company. Getting in contact with the company to express one's frustration is a lengthy process. The consumer must spend several minutes on the phone just to leave a message. With social media platforms such as Twitter and Facebook, users can voice his/her opinion and concerns in seconds. Along with the opinion, other users can support the position and re-tweet or repost.
3. Helping the company or cause: Supporting the company generally occurs when the consumer has a very positive sentiment about a certain product or company and decides to express their satisfaction. These are loyal customers who are enthusiastic about the company's progression and products.

4. Advice seeking: Advice seeking is similar to viewing reviews for a product or a movie. Others hope to receive advice from these posts to help resolve an issue. A user might post an open-ended question hoping that someone replies to his or her tweet. For example, a user posts “Is Batman a good movie?” The user then waits for someone to reply to him with advice or information regarding his or her question.
5. Venting negative feelings: This reason is one of the most prevalent. People feel the need to express their negative sentiment on a topic through social media. People often disagree with daily occurrences. Instead of keeping their opinion to themselves, social media is a tool that lets everyone express their feelings. General users use social networks for venting purposes more than any other reason.
6. Positive extraversion: These individuals are satisfied when they can tell others about their buying success. The consumer becomes exhilarated announcing their experience of buying a product.
7. Economic incentives: Users can receive a reward for posting on certain topics that the company mentions. This could be a monetary reward, product, or discount from the company. Companies sometime support a promotion in which you can tweet the company and receive various rewards. For the company it provides publicity, more activity, and sometimes an increase in sales.
8. Social Benefits: Social media was invented for social benefits. Social platforms are a fun way to communicate with other people in the community. Users can communicate with each other without exchanging phone numbers. Social media account holders can also meet people that are in their community before physically meeting them. If someone does not have the luxury of traveling nationally or internationally, they can meet or converse

with someone globally via social media. Social benefits are the initial reason why users post, tweet, or connect using social media.

METHODOLOGY

The research and method concentrates on Twitter. To assess reliability, a method was created to determine if a tweet is credible or not from a given corpus of tweets. There are several steps to determine if the tweet is true. The steps are collection, filtering, feature scoring, parsing, expansion, text mining, Amazon Mechanical Turk (Citizen Sourcing), and results. Refer to Figure 2 for the architecture of the experiment.

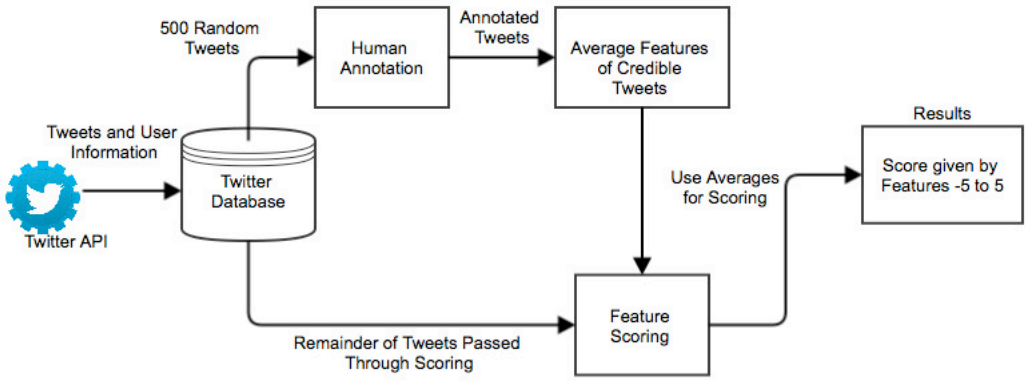


Figure 2: Architecture of Scoring a Tweet using Features Credibility

The Architecture

Figure 2 gives the overview of how tweets are analyzed using features to determine the validity. Tweets were collected using Twitter's API and placed in a file. 500 random tweets were annotated and determined if the tweet was credible or not. From the tweets, the credible tweets were collected and the features were averaged. The remaining tweets were passed through the

scoring method to retrieve a score between -5 and 5. From the 500 random tweets, the average features of the annotated tweets that were credible are used to compute the score.

Collection

The programming language Python (Foundation, 1995) was used because of the various libraries and services it provides the users. Tweepy (Roesslein, 2014) is a library that was provided and used to capture tweets. Tweepy is open-sourced and hosted on GitHub. It allows users to access and communicate with Twitter's entire RESTful API methods (Twitter, 2014b) without any threshold request of tweets per day. Tweepy can be downloaded, installed, and imported in minutes.

The main Twitter method that was used is the filter method by Tweepy. This filter method requires a string of characters to filter traffic. The string or hash tag is used as a parameter for the filter method in Tweepy. This method utilizes strings to improve the relevance of the tweets collected. If the string were null, Tweepy would return all traffic of Twitter and will return around 6,000 tweets per second. Twitter's API returns a JSON object for each tweet in which it provides all information pertaining to the tweet and the user's information. Figure 3 shows how the filter method is utilized using Tweepy. After the filter method, the information is saved to a text file and later used to analyze the credibility of the tweet. The data collected includes creation time, id number, text, source (web/phone), response, users tweet responded to, username, full name, location provided, geo-location of tweet, biography, time zone, number of followers, number of people following user, number of times re-tweeted, if user tweet was a re-tweet, number of times favored, etc. Figure 4 shows how tweets are collected with the supporting information of the tweet. Each tweet included with its raw data is classified into three different components: the tweet (actual message posted), the metadata (number of followers, geo-location,

number of re-tweets, and number of favorites) and the source (user who posted tweet). From these three components, a method was created to determine the credibility.

```
auth = OAuthHandler(ckey,csecret)
auth.set_access_token(atoken,asecret)
twitterStream = Stream(auth,listener())
twitterStream.filter(track = ["sb1062"])
```

Figure 3: Snippet of Code Using Tweepy to Collect Tweets. Filter method takes in string of an event


```
{"created_at":"Tue Feb 25 20:22:21 +0000 2014","id":438408649392656384,"id_str":"438408649392656384","text":"RT @SenJohnMcCain: I hope Governor Br
{"created_at":"Tue Feb 25 20:22:25 +0000 2014","id":438408662701191169,"id_str":"438408662701191169","text":"Blockade #SB1062 this Saturday! Direc
{"created_at":"Tue Feb 25 20:22:30 +0000 2014","id":438408687217283072,"id_str":"438408687217283072","text":"Every day is \"Throw Back Thursday\"
{"created_at":"Tue Feb 25 20:22:31 +0000 2014","id":438408688252878848,"id_str":"438408688252878848","text":"Arizona takes a beating in national p
{"created_at":"Tue Feb 25 20:22:32 +0000 2014","id":438408693479002113,"id_str":"438408693479002113","text":"RT @herophoenix: Now Intel has come o
{"created_at":"Tue Feb 25 20:22:44 +0000 2014","id":438408745832710144,"id_str":"438408745832710144","text":"@nfl to Arizona &amp;amp; @GovBrewer: Pas
{"created_at":"Tue Feb 25 20:22:45 +0000 2014","id":438408749586587648,"id_str":"438408749586587648","text":"Advocating for ALL their 11K #AZ work
{"created_at":"Tue Feb 25 20:22:45 +0000 2014","id":438408749599166464,"id_str":"438408749599166464","text":"Advocating for ALL their 11K #AZ work
```


Figure 4: Tweets that were collected and format in which they are collected

5 days after the incident, the media had begun investigating the mishap in Ferguson, Missouri, and social media had continuous bursts of comments about this topic. After seeing the response from the public, tweets were collected starting at 2:37 pm on August 14th, 2014 and ending at 8:17 pm on August 27th, 2014. 1,150,889 tweets were collected in this timeline. From the 1,150,889 tweets, there were 122,411 unique tweets. The size of the file with the features pertaining the users profile information was 5.28 GBs.

Bill SB 1062 had a shorter timeline because the bill was resolved in days. The timeline for this issue was from 8:22 pm on February 25th to 6:50 pm on February 27th, 2014. 28,837 tweets were collected and from these tweets there were 9,967 unique tweets. The reason for the vast amount of re-tweets is because Twitter users were concerned with dispersing the decision of the bill. The size of the file is 134.1 MBs. The file is divided into two parts with the first half

containing tweets that were published before the result of the bill, and the later half containing tweets that were published after the Governor made the decision.

Filtering Using Metadata/Features

After collecting tweets, there is garbage and unnecessary information that was filtered for each feature. One feature is utilized at a time so that ambiguity of the method is reduced. This was done using a Python script to filter and retrieve different parts of the tweet. JSON parser is a Python import that uses the load method to allow users to easily extract objects in JSON format. Figure 5 shows Python's import JSON and how it is implemented. This method extracts the id of a tweet from a tweet in a JSON object.

```
data = json.loads(line)
id_t = data["retweeted_status"]["id"]
```

Figure 5: Example of JSON method load to decode JSON objects

FEATURE SCORING

500 Tweets were randomly selected to extract the average amount of the features of a credible tweet. These tweets were annotated to reflect the characteristics of a credible tweet. Ranges were set to reflect these averages of features that were categorized as credible by the annotator. Twitter features that were used are verified account holders, re-tweets, favorites, friend and follower count, SocialBaker's fake friends application, profile picture, Klout score, and biography.

Verified Account Holders

Verified account holders were used in the method of credibility. Twitter states that "*We concentrate on highly sought users in music, acting, fashion, government, politics, religion, journalism, media, sports, business, and other key areas*. "[This is to help with] *verification and to establish authenticity of identities of key individuals and brands on Twitter"* (Twitter, 2014a). Twitter has problems with credibility, and combats the issue by verifying users to ensure that citizens are getting trusted and credible information. Due to the numerous false statements and fake profiles, Twitter has used verified accounts to give a sense of credibility when searching through topics. Consequently, this verification method utilized by Twitter will be implemented in this research to assist with the official power.

When the Michael Brown incident occurred, numerous verified news agencies dispersed information regarding the shooting and the city of Ferguson, MO. The feature scoring method treats these users as if they are official power because Twitter has verified that these users are higher officials in society. This feature is the largest component of scoring using features. For all tweets that have a verified user, 3 points were added towards the overall credibility of the tweet. For users that were not verified accounts, the credibility of the tweet was left unchanged because there was no information in this feature to determine the credibility. Figure 6 shows the twitter page of Apple's CEO, Tim Cook. Notice the verified check beside his name that signifies that his account is verified and is the correct page for Tim Cook.

Figure 6: Tim Cook's Verified Twitter Page, (b) Twitters Verified Badge

Re-tweets

If a user does not have adequate followers, in order for their story to spread and be believed, it must be re-tweeted. The user can reply to certain users' tweets, direct message users, or favor a user's tweet to have them re-tweet the message. Re-tweeting is the easiest way to share information or a tweet on Twitter. Once a Twitter user has re-tweeted a post, it is available to all of the users' followers. The objective of re-tweeting is to get the message to more people that are not in the original user's social circle, and to get other users to re-tweet as well. Without the power of re-tweeting, a user's story will continue to be deemed invalid because it does not have an official that verifies the information. The tweet will simply stay in the user's network and eventually lose exposure through the timeline. The number of re-tweets factor into the credibility of a certain tweet because of the number of people supporting the view and wanting to spread the information. If a tweet has been re-tweeted numerous times, the probability of the tweet being credible increases as it continues to be re-tweeted (Kang et al., 2012).

Using the 500 tweets that were annotated, the average number of re-tweets that were credible tweets was 5. Tweets the annotators categorized as not credible average less than 3 re-

tweets. A range was created to score the tweets on their amount of re-tweets. For tweets that received less than 3 re-tweets, a score of -1 was given toward the credibility of the tweet. All tweets that have more than 3 re-tweets received a score of +1 that was calculated toward the credibility score.

Favorites

Twitter's favorite button is another function that aids the credibility method. The reason users choose to favor a tweet is to "express support or approval" (Stadd, 2013). This is very similar to Google's +1 button, Facebook's like button, Instagram's like button, and YouTube's thumbs up button. These buttons express support for a certain post or position that the user has taken. If a post has received numerous favorites, the public has spoken to support the post. This is also a way to show like-mindedness because the user agrees with the tweet. Instead of replying to a tweet to express support, it is easier to press the favorite button.

By utilizing the random 500 annotated tweets, the average number of favorites was 2. This information is used to score the tweet by the feature of favorites. There were a number of tweets that were credible that did not possess any favorites. Therefore, the range for a credible tweet deserving a +1 towards the credibility score was 2 favorites and above. Tweets that have below 2 favorites are disregarded because this feature is difficult to incorporate with credibility. Favorite scores of 2 and below will receive a 0 towards the credibility score, neither increasing nor decreasing the score.

Friend and Follower Count

The size of a social network increases the probability of credibility. Having followers implies that these users want to see what a person publishes on their micro-blog. The follower has the option to un-follow the user if they feel as if the information is not worth viewing on

their timeline. The number of friends and followers gauges the user's activity and is constantly changing because of difference in opinions. Many people will not follow a user if they do not have sufficient friends and followers. Therefore, friends and followers are considered in the method.

Kang conducted research and developed a range of followers and friends that are correlated with credibility (Kang et al., 2012). Figure 7 displays a plot in which the shaded areas present suspicion in the validity of a tweet. The cold start area classifies users that do not possess enough friends for a reasonable assessment of their account. Celebrities and large corporations are generally verified, have a high volume of followers, and are not following many people. Users that were classified as low credibility possess many followers and are following many people. These accounts have a high possibility of being a fake account because of the high amount of followers and friends. Users that have a high number of friends and a low number of followers, generally re-tweet rather than composing a new tweet. Therefore, Kang, unable to make a determination on the credibility, classifies the area as suspicious.

Based on the annotated tweets, an average of 512 followers per user was considered to be credible. An average of 521 friends was considered to be credible. From the averages, a range of 500 and greater friends and followers was selected to increase the credibility score of a tweet. A follower to following ratio is implemented to increase the believability. Researchers and the general public view profiles with a high follower to following ratio as an account that cannot be trusted (Morris et al., 2012). Twitter has many users that have a number of followers and little friends. Government officials have this high ratio because they choose not to view a random person's blog. A user that follows numerous people without becoming friends with the user is viewed as incredible (Westerman et al., 2012). With Kang's research and the manual

annotations, a ratio of 0.8 was introduced as a standard of credibility. Profiles that possess greater than a ratio of 0.8 and less than 50.0 will have 1 added to the credibility score, and all users that have less than 0.8 and greater than 50.0 will have 1 decreased from the credibility.

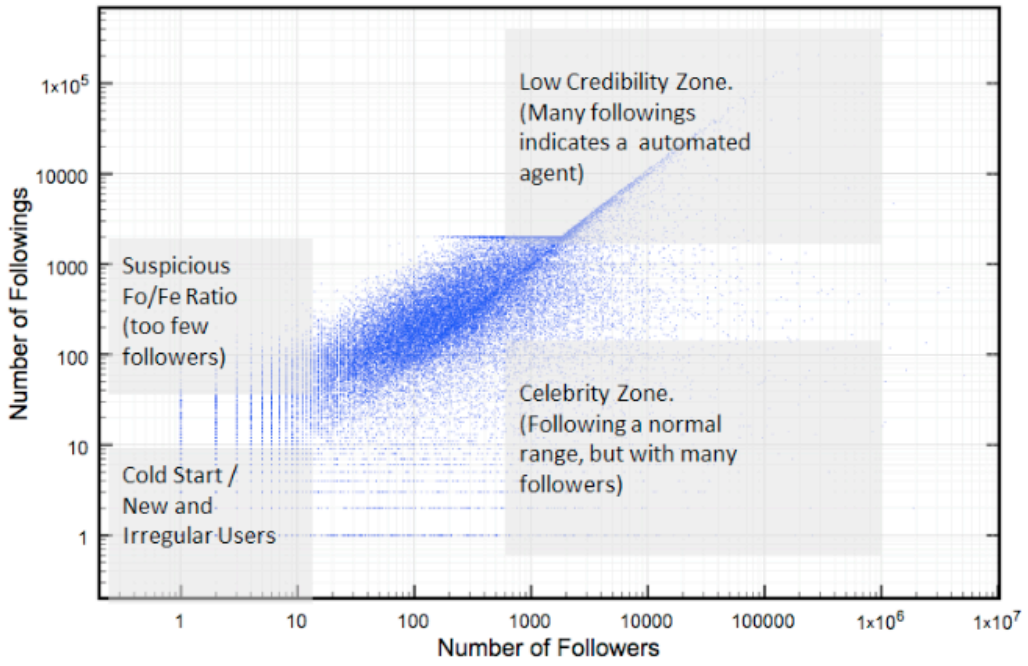


Figure 7: Plot showing the number of followers to number of following profiles for the Libya dataset. Shaded areas are of concerns of credibility and displays users of Twitter (Kang et al., 2012).

Socialbaker' Fake Followers Application

SocialBakers explains the application as *"a user friendly social media analytics platform which provides a leading global solution that allows brands to measure, compare, and contrast the success of their social media campaigns with competitive intelligence"* (SocialBakers, 2014a). This company provides information from Twitter, Facebook, YouTube, LinkedIn, and Google+ to assist campaigns. One fifth of Global Fortune 500 companies are taking advantage of social analytics to aid in comparing their competitors. SocialBakers provide information such as

top brands, amount of fans/active followers, average engagement rate, and rating on all social media platforms. The company also shows the age range percentage for each social platform in various countries. With a subscription fee, a user can access analytics in the given social media platforms.

SocialBakers provide a free fake followers application to determine the amount of followers that seem fake or empty. This application provides the percentage of followers that are good, suspicious, and inactive. The category suspicious has many flags because there is not a definite algorithm to determine if an account is false. Suspicious criteria pertains to (SocialBakers, 2014b):

- The account has a follower ratio of 50:1;
- More than 30% of the users tweets are spam phrases such as “diet”, “make money”, and “work from home”;
- The same tweets are repeated more than three different times, even when posted to different accounts;
- More than 90% of the account tweets are re-tweets;
- More than 90% of the account’s tweets are links;
- The account has never tweeted;
- The account is more than two months old and still has a default profile picture; and
- The user simultaneously didn’t fill in the bio or location and, at the same time, is following more than 100 accounts.

Figure 8: (a) Results from SocialBakers on President Obama. (b) List of Suspicious followers hovering over user it shows why it has been categorized as suspicious.

Each of the criteria has a given number of point valuations and when an account exceeds a certain number, SocialBakers identify the account as suspicious. If the account does not identify as suspicious, they test for two things to determine if the user is inactive: whether the account has posted less than three tweets, and whether the last tweet posted is more than 90 days old. If the account is not classified as suspicious or inactive, it is a legitimate account and will be categorized as good. Figure 8 is an example of Barack Obama's result when his username was entered into the search bar.

The fake follower application is currently in the Beta stage, and the company is dedicated to enhancing the algorithm. The application takes a 100 random set of followers and implements the method. For accounts with a large number of followers, this might cause a problem, but the results can scale toward the actual number of followers. SocialBakers does not provide an API for the application so retrieving information was a lengthy process.

A manual API was created to retrieve the amount of suspicious and inactive accounts that were associated with a user. The made API simply concatenates the username on the end of the URL and gets the results of the page. To get the information regarding a user, Python's urlib2 import was used to obtain the content from a particular webpage (Foundation, 2014). Beautiful Soup was also imported to aid in acquiring tags from the html text (Soup, 2014). Simple filtering of the output was implemented to reduce ambiguity and give clean results as if it was from SocialBakers. If the account cannot be found, it will provide all 0's on the results. Figure 9 displays a snippet of the API created to obtain the necessary information about the user.

For scoring, accounts were searched to find a high follower to friend ratio and higher than usual followers. All followers to friend ratios that were more than 30 and accounts that had over 2,500 followers were flagged for potentially having fake followers. These accounts were recorded and passed through the application to determine the percentage of followers that are fake or empty. If an account has less than 50 percent of its followers categorized as good, then the account gets a credibility score of -1. The decision to decrease the account credibility score by 1 is because a point was increased in the previous friends and followers scoring method by the amount of friends. By deducting a point, the credibility score will not be affected by friends and followers.

```
array = []
with open("BakersUserNameLinks.txt") as r:
    for line in r:
        line = line.split(";")
        response = urllib2.urlopen(line[1])
        html = response.read()
        soup = Soup(html)
        x = soup.findAll('strong')
        for element in x:
            for y in element:
                if "%" in y:
                    array.append(y)
        print line[0] + ":" + array.pop() + ":Good"
        print line[0] + ":" + array.pop() + ":Inactive"
        print line[0] + ":" + array.pop() + ":Suspicious"
```

Figure 9: Snippet of java code to retrieve information from SocialBaker's fake friend application

Profile Picture

The profile picture is one of the first items viewed when determining the validity of the user. If a user has not changed their picture from the default picture, a flag is raised for a fake page. By having the default picture, the user could be inactive, a spam user, or someone that does not want their picture to become public. Default profile pictures are a feature that researchers have concentrated on. Julia Schwarz's research states that participants gave the default picture a 2.87 of 5 for credibility impact (Morris et al., 2012). Extracting the picture is simple and takes minimal time. Twitter provides a Boolean tag that appears "True" if a user has a default profile picture. Figure 10 shows the various Twitter default profile pictures. When using the default profile picture, it is difficult for an account to be perceived as credible.

From the 500 annotated tweets that have been deemed credible, no users were credible that had the default profile picture. Therefore, the credibility score was decreased by 1 if an account has the default profile picture. If the user changed the picture to a personal photo, the credibility score remained the same. Searching for a default profile picture serves as a flag for spam, inactive users, users that only re-tweet, and users that have not taken the time to make their profile unique.

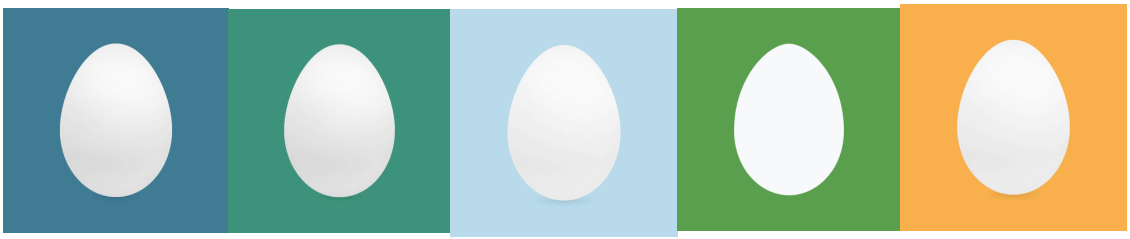


Figure 10: Default Twitter Profile Picture

Klout Score

Klout is a company that uses social media analytics to rank users based on their social influence (Klout, 2014b). They give users a numerical score between 1 and 100. According to Klout, influence is "*the ability to drive action,*" or *"the ability to share something and get responses to your post*" (Klout, 2014a). Klout uses an algorithm that measures the size of a user's social media network, and how others interact with the user. This company links Facebook, Google+, LinkedIn, Instagram, YouTube, Flickr, Blogger, Tumblr, Bing, Foursquare, and Wikipedia with Twitter to give a valid and detailed score.

Klout is a unique application because it searches through a network to determine the number of users that have low activity. This algorithm determines how influential the people who re-tweet the user are and the user's amount of activity. Klout incorporates three measures to achieve the score called true reach, amplification, and network impact. True reach is scored by the size of the network that is actively engaged in the user messages. This neglects the friends and followers that are not active with the user that rarely post. Amplification predicts the possibility that the user's tweet/message will generate actions which includes re-tweets, favorites, and comments according to their past posts. Network impact is the computed influence of the user's audience that is engaged. This is calculated by over 400 signals from eight different social networks. The algorithm has combinations of features that increase the score if certain signals such as the percentage of reactions you receive are found.

Klout provides an API that allows developers to call various methods, which generates Klout score, usernames, user id, show influencers, and accounts that influence the user. In this research the score is the only feature that will be used. To obtain the Klout score, the developer must pass in a username as the parameter to acquire the Klout id. The Klout id is the unique id

for each user's information of their influence to their network. The id is passed into the score method from Klout and returns the score. A list of users were filtered from the corpus of tweets and used to determine the Klout score. Figure 11 displays a snippet of code to get the score.

According to Klout, the average score of all users is 40 and any score above 63 is in the top 5% of users. The scores provided are dynamic because scores update frequently. The score gets updated daily and enables developers to create graphs on the progression of each user's Klout score. Scores that were above 34 were given a credibility score of +1, while any score below 30 was given the credibility score of -1. Tweets with Klout scores between 34 and 30 were classified as non-determined and disregarded leaving the credibility score unaffected.

```
kloutId = k.identity.klout(screenName = users).get('id')
score = k.user.score(kloutId=kloutId).get('score')
```

Figure 11: Klout score code snippet

Biography

Biography is used in Twitter to tell users about themselves. There is no limit on characters that can be used in the biography. Users usually state where their hometown, school, age, affiliation of club or organization, quotes, religious background, political background, etc. When setting up an account, this is one of the first things that users complete. Users tell about themselves so that the public will have background information of the user. Twitter users without a biography introduce a sense of doubt as to whether the user is correct or intended person. If there is doubt when viewing a profile then the credibility of the text is also questioned. Therefore, the biography is featured in the method of credibility through features.

From the annotated tweets, all tweets that were categorized as credible had completed a biography. There were users that did not have a biography, and these accounts were closely

related with their profile picture. 73% of users that had the default profile picture also had a null biography. Thus, 1 is subtracted from the credibility score if a user's biography is null. If a biography is completed, the credibility score is left unaffected. Increasing the score for having a biography would provide incorrect credibility scores. Analyzing the biography's text would provide unnecessary information and would be ambiguous. Since having a null biography is a characteristic for a fake user, this feature was used to determine the credibility of an account or user.

Features that were not used in the Credibility Score

Grammar wasn't used in scoring the credibility of a tweet or user. In a tweet, a post is restricted to 140 characters, so the information and grammar will be compacted. Links to websites and short acronyms such as lol, smh, and hash tags, would cause false errors when correcting grammar. Natural Language Tool Kit has a Python import for grammar but wasn't selected to score the profile or tweet. Since majority of Twitter users write shorthand, grammar was neglected when computing the credibility score.

Tweet length isn't factored into the feature credibility scoring of a particular tweet. Some researchers incorporate tweet length in prediction, but while analyzing the annotated tweets, length had little correlation with credibility. Credible tweets were short and long. Some tweets had headlines of a topic with a link, while other tweets used more characters to explain the position or topic. Instead of using the length of a tweet, natural language processing was used later to score the credibility of a tweet. Therefore, tweet length wasn't used to determine the credibility of a tweet.

The presence of links was disregarded because it requires a human to visit all individual links in a tweet. This would require additional work to determine if the link is correct and

credible. While annotating, tweets that had links were directed to personal websites that were blogs. In other instances, it was a link to a picture in which a user would try to validate the post by the image. If links were included into the method, image processing would be needed to verify if the image is credible. Although links can provide supporting information regarding the topic, it has the potential to be from a blogger or picture that is not credible. Links can have an invalid addresses where it has been taken down, changed, or the server was down. Consequently, links were not included in the feature credibility score.

Calculation of Features

A Python file was created to produce a score for each feature. A for loop was used to go through each file and match the unique tweet ID to compute the score for each tweet. Score for biography, Klout score, profile picture, Socialbaker's fake follower application, friend and follower count, re-tweets, and favorites were all used to provide the final credibility score using features of the account. The maximum credibility score using features is 5 and the minimum score is -5. Using these various features works as a Swiss cheese model for a credible tweet. Figure 12 shows how the Swiss cheese model is implemented using these features and applications to sift through features to get a credible account. Lastly, error handling was used to handle all situations that would prevent a inconsistent scores. The output simply returns the tweet ID and the overall credibility score.

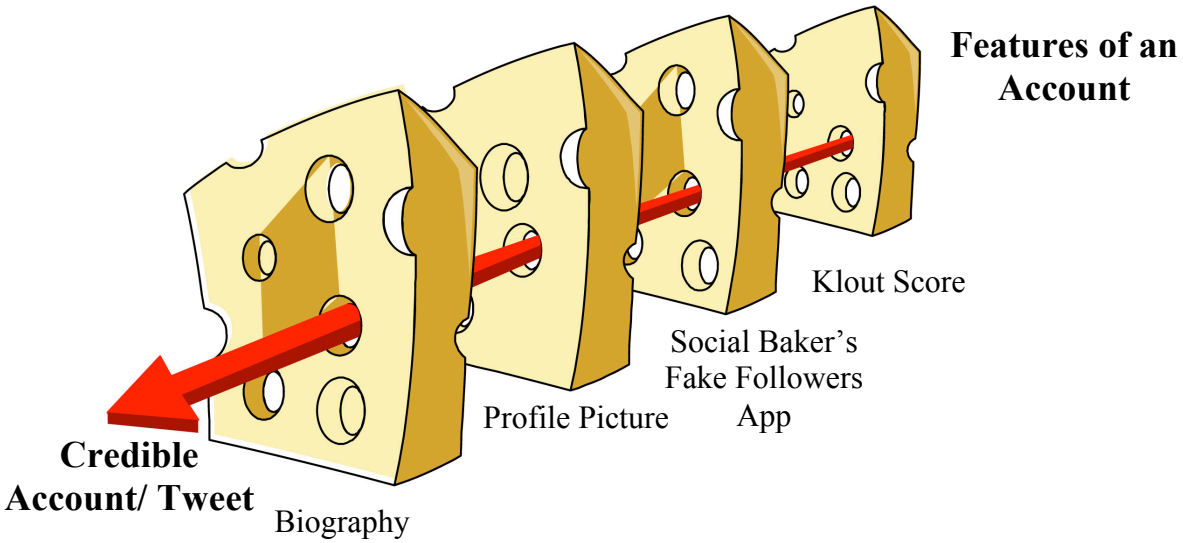


Figure 12: Swiss cheese model to filter out the characteristics of a fake/incorrect tweet using features

NATURAL LANGUAGE PROCESSING APPROACH

Many researchers have used features to obtain credibility but issues arise when solely using features. This assessment of credibility approach attempts to accommodate all areas of credibility. This unique process uses natural language processing from Natural Language Toolkit to utilize the English language by searching for words rather than obtaining the sentiment. A word search was used instead of sentiment analysis because it was more logical to get tweets regarding information that has been deemed true. It is difficult to determine credibility by a user's sentiment. Sentiment changes frequently and sentiment analysis for credibility does not utilize the true meaning of a word. Using text similarity does not provide flexibility that is needed in an everyday process of determining credibility. Researchers attempted to place credible tweets into a file and score tweets based on the physical construction of the tweet. Using text similarity the determination of credibility has a fixed sentence that does not change. The

natural language processing method in this research expands words so that the sentence has flexibility of changing and has the possibility of being credible if related words are used.

Using features may show the number of supporter but features cannot be used independently to determine the validity of a tweet. Each tweet has different information regarding a topic and using its background information to determine overall credibility is difficult. The features are used and factored into the final decision or credibility score of a tweet. Credibility scores should be determined by the content of the tweet. Many users aren't popular, have little re-tweets, and have a low number of favorites but possess valid statements. The text is analyzed to make a decision about whether the content is correct.

The goal of the natural language processing approach is to retrieve similar tweets that possess certain keywords. If information has similar keywords then we can conclude that tweets are related and have a higher possibility of being credible. By giving the tweets flexibility of changing, results are expanded by retrieving the same keywords and similar keywords.

Architecture of Natural Language Processing

The same process of collecting tweets using Twitter's API was used to collect and store tweets in a text file. Tweets are divided by the status of verification through Twitter's verified account feature. Verified tweets from news agencies are manually entered into a file. Newspapers from news agencies, criminal records, and domestic records are placed into this file pertaining to the event. The information from the verified file is tagged with a parts of speech tagger. A list of nouns, verbs, adjectives, and adverbs is created from the Deemed Credible file of keywords passed through the parser. The parts of speech files are combined to make one file of credible words. The words are expanded to provide synonyms of the words that were from the credible file. A list of stop words was created to reduce ambiguity from the credible words.

From the parts of speech text (credible words) the process searches through the corpus of tweets for similar words pertaining to the event. An equation was implemented to provide the score of the individual tweet depending on the number of words found. Random tweets are selected to use crowdsourcing through Amazon Mechanical Turk for humans to analyze the tweets. Tweets were selected that possessed elevated high scores and lower scores to determine if there is any correlation. Results are analyzed and the automated approach is compared with human intelligence.

Previous Methods	Current Method
Used Crowd Sourcing	Used Crowd Sourcing
Only uses tweets for the golden standard	Uses credible sources and news agencies for the golden standard
Analyzes text using frequent words that crowd deems credible	Analyzes text using articles, criminal release records, and tweets that are credible
Only uses words that are frequent to search through corpus of tweets	Expands words by obtaining the synonyms from articles and other reliable sources to search throughout the corpus of tweets
Features utilized was number of unique characters, presence of swear words, number of pronouns, and emoticons.	Utilized friend count, follower count, re-tweet count, favorite count, friend to follower ratio, swear words, emoticons, and Klout score.
Nothing was done to combat the presence false followers	Used a flag to determine if user potentially brought followers. Viewed the amount of Suspicious and inactive followers user has.
Used BM25 algorithm to determine tweet similarity	Used text expansion using NLTK and search for tweet similarity

Table 1: Difference in the assessment and previous methods for obtaining credibility of tweets

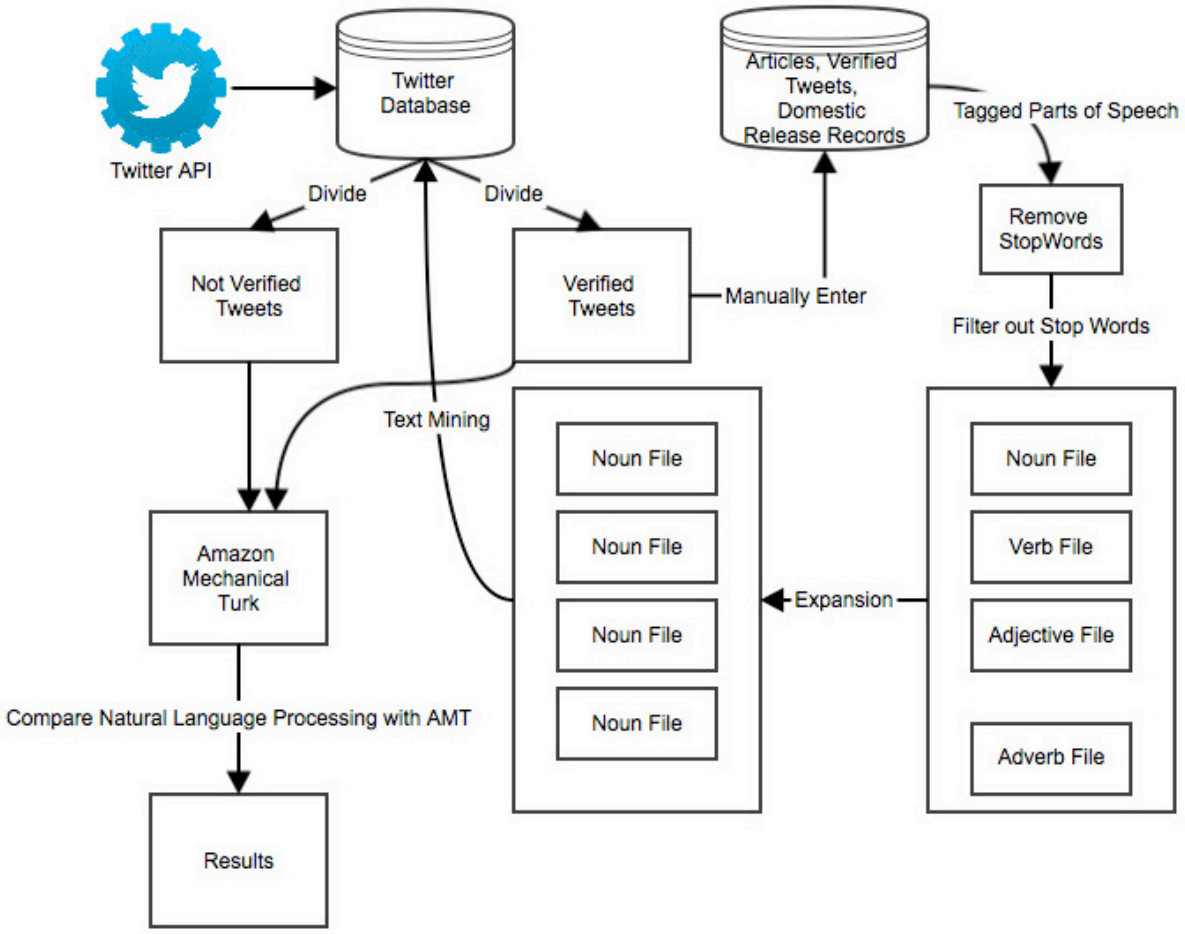


Figure 13: Architecture of the Natural Language Search using Natural Language Toolkit

News Agency Dependency

With the information obtained from the metadata, there is an issue with solely using background information to determine if a tweet or user is credible. “Fact of the matter is that ordinary people are usually neglected as credible sources. The general population is not deemed authoritative like official sources such as news a gencies” (Mabweazara et al., 2014). For example, when an event occurs, the public does not believe the first person that posts about the event. Even though the information may be true, society disregards the information until it has

been dispersed by a reputable source. Thus, without an official source, one cannot validate any information on Twitter or any social network.

Information coming from the news has been deemed credible because of the process it takes to make it truthful. A text analysis comparison method was chosen because when an event occurs, the source reports to a news agency. The news agency conducts an interview with the primary and secondary sources that have witnessed the event, and quickly disseminates this information as breaking news, making the information from the news companies valid.

One issue that may arise is the event that happen before a reporter is notified. What if someone witnesses a murder and post a tweet before contacting an official or being interviewed by police officials? The approach to this issue is simple. The time between the event occurring and reporting of the story by the news agency has a very short timeframe. National and local news agencies have employees that relay information quickly to the public via social media. The time frame for a person tweeting about an event within 20 minutes after the occurrence is unlikely depending on the severity. Figure 14 (Becker, 2014) displays the timeline of the dispersion of information by different ways. The timeframe for social media and protest activity is almost immediate but traditional media sifts through the rumors and waits to confirm that the facts are reported. The reason for the delay is interviews and investigations to ensure that the information is credible. Text mining and search relies on news agencies' reports/articles, criminal and domestic released records, and verified twitter accounts holders to provide credible information.

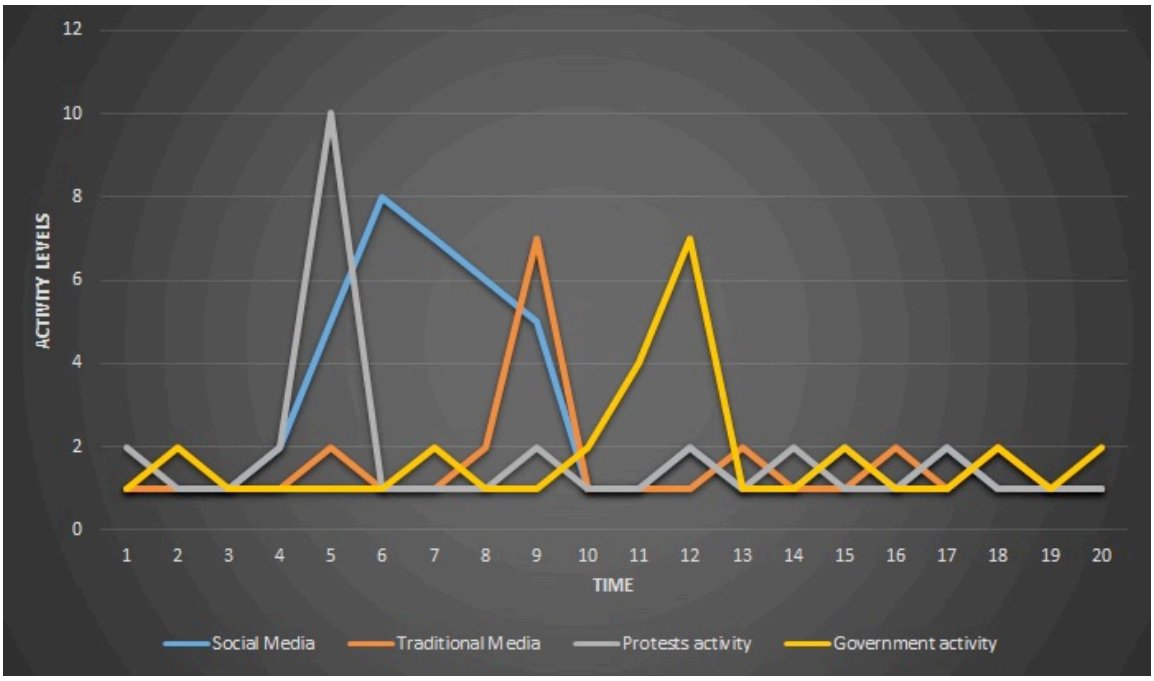
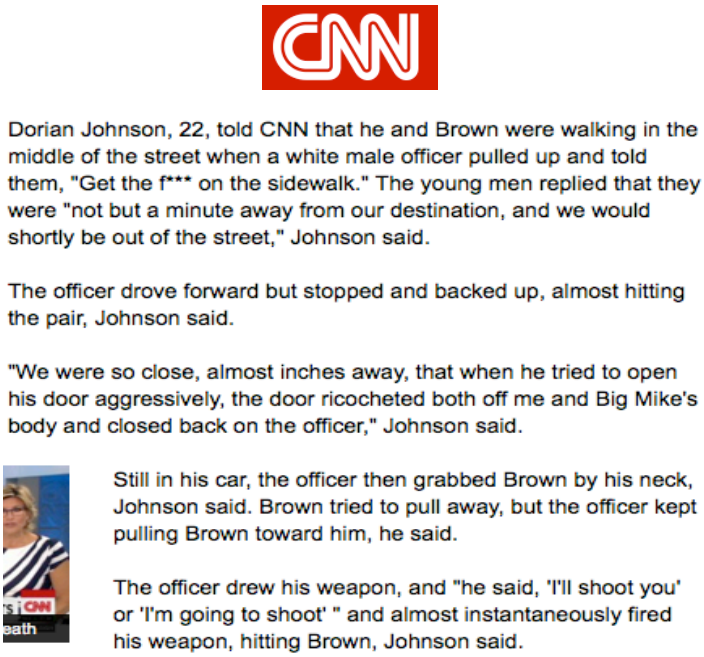


Figure 14: Timeline of information/news dispersion on events

Creating the Golden Standard

From the corpus of tweets, tweets were partitioned into verified and unverified to differentiate tweets between users that have and have not been deemed credible by Twitter. They were saved into two different files while preserving the original corpus of tweets called Deemed Credible file. Tweets were selected manually that are credible by conducting a quick Google search to ensure that each tweet is credible. Along with the tweets in the Deemed Credible text file, the method requires a user to manually insert sections of news articles on the topic provided. The selection of tweets and articles is the foundation and the most important part for determining the validity of the tweets through text mining. The credible text file is reliable and the information has been deemed truthful by prior interviews or Twitter. The public believes the news coming from news agencies more than they believe news coming from social media. Therefore, news agencies' articles are incorporated into the text mining method. Figure 15 displays a CNN article on Mike Brown's incident.



Dorian Johnson, 22, told CNN that he and Brown were walking in the middle of the street when a white male officer pulled up and told them, "Get the f*** on the sidewalk." The young men replied that they were "not but a minute away from our destination, and we would shortly be out of the street," Johnson said.

The officer drove forward but stopped and backed up, almost hitting the pair, Johnson said.

"We were so close, almost inches away, that when he tried to open his door aggressively, the door ricocheted both off me and Big Mike's body and closed back on the officer," Johnson said.

Still in his car, the officer then grabbed Brown by his neck, Johnson said. Brown tried to pull away, but the officer kept pulling Brown toward him, he said.

The officer drew his weapon, and "he said, 'I'll shoot you' or 'I'm going to shoot' " and almost instantaneously fired his weapon, hitting Brown, Johnson said.

Figure 15: Excerpt from CNN about the facts of the Michael Brown shooting

Power of Twitter and Flaws

Twitter can be used to report and receive tips without any verification. Users can follow tips given by users and examine the story closely to get more information in order to verify the integrity. In certain countries where freedom of speech is not exercised, Twitter provides them an avenue to express themselves. Some individuals believe individuals should never believe information from the general public because it is best to retrieve information from a credible source. Mabweazara states, “The public is perceived as being too biased or personally entrenched in their situation to be neutral (Mabweazara et al., 2014)”. In politics, this usually occurs because individuals are emotionally involved in the situation and cannot create an opinion from a neutral perspective. On social networks, the public is not attentive to the actual facts, but address rumors. Users express politics as it pertains to their life and support the position that is beneficial to them. Individuals express how laws and politics affect them while posting tweets.

Most individuals generate their opinion from word of mouth conversation, blogs, social media, and experiences. This credibility method helps with the believability of certain tweets by text mining and through its metadata.

Twitter can be used for dispersing the truth but sometimes it disperses myths as well. On many occasions, users have started a myth and it became "viral". Users on Twitter have spread misinformation about celebrity's relationships, financial status, and morality. For example, Miley Cyrus was believed to be dead for ten minutes when a journalist, Peaches Geldof, tweeted that she died (3am, 2009). The journalist claimed that the information regarding the death came from a close friend of Miley Cyrus and was deemed true. The death was spread over Twitter by re-tweets in a matter of minutes. At the time of the increase activity about Miley Cyrus, she was performing at a concert and was very much alive. Thus, proving the statements from journalist Geldof incorrect, and she immediately deleted the tweet. Figure 16 shows journalist Charles Johnson who dispersed false information regarding Michael Brown. Johnson's reputation drastically decreased once the public found out he was dispersing information from his personal beliefs that had no truth. This tweet was the top re-tweet that was false on the Michael Brown incident.

Figure 16: Mike Brown's Top Rumor that was Re-tweeted over 1000 times

False information gets dispersed about our government official's activity and health daily. In 2012, President Barack Obama was believed to be dead by Fox News Politics Twitter account on the fourth of July (Greene, 2011). It was stated that Obama had suffered two gunshot wounds and had died. Fox News account had sent their condolences to the President Joe Biden and wished him luck with being our new President. After the tweets were posted, Fox News found out that their Twitter account had been compromised by hackers. Following the hoaxes, Fox announced that the information was false and none of their employees were involved in the compromise of the account.

Twitter's security system uses the simple login with the email address and password. Hackers can easily breach the account and get into various accounts and make tweets with false information. Making passwords difficult and changing passwords frequently decreases the possibility of an account being hacked. News accounts hold official power and influence millions of people to believe their posts. Hacking and reporting false information has a possibility of occurring on Twitter. Time is the only factor that disproves myths. Time is needed to object to the false information with valid information that disproves the myth. This is the only drawback and reason the general public does not use Twitter as their primary source of information.

With Twitter, everyone has an opportunity to express their frustrations and make up their own story. With a micro-blog such as Twitter, users have the power to post about anything. The information has the potential to spread throughout various social circles whether it is true or false. In the method used, tweets were collected and given time to withstand the myths.

Tagging & Stemming

Natural Language Toolkit (NLTK) was used to tag the sentences and words of the verified tweets. NLTK is a "leading platform for building Python programs to work with human

language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning" (Project, 2013a). Natural Language Processing with Python, written by the creators of NLTK, guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, and analyzing linguistic structure. NLTK is intended to support research and teaching in Natural Language Processing or areas including empirical linguistics, cognitive science, artificial intelligence, information retrieval, and machine learning.

Sentences from the Deemed Credible file (Credible words), which have all verified tweets and articles, are tokenized and tagged by NLTK's part of speech tagger. NLTK categorizes nouns, verbs, adjectives, and adverbs and files were created to store each word in. Words are tagged and given their part of speech and saved in their respective part of speech text file. Table 2 gives the tags for the parts of speech that will be used in the research and provides examples of various parts of speech given by NLTK. After the execution of a sentence, NLTK provides an array of tagged words. The sentence (e.g., "United States is increasing its military presence in Iraq, ordering 300 more troops to the violence-ravaged nation, the Pentagon announced Monday") is used as the input. The result is:

```
[('United', 'NNP'), ('States', 'NNPS'), ('is', 'VBZ'), ('increasing',
'VBG'), ('its', 'PRP$'), ('military', 'JJ'), ('presence', 'NN'), ('in',
'IN'), ('Iraq', 'NNP'), (',', ','), ('ordering', 'VBG'), ('300', 'CD'),
('more', 'JJR'), ('troops', 'NNS'), ('to', 'TO'), ('the', 'DT'), ('violence-
ravaged', 'JJ'), ('nation', 'NN'), (',', ','), ('the', 'DT'), ('Pentagon',
'NNP'), ('announced', 'VBD'), ('Monday', 'NNP'), ('.', '.')].
```

Refer to Table 2 for tag description. All categories of adjectives, nouns, adverbs, and verbs are grouped in one file for their part of speech.

For the verb file and adjective file a stemmer is used to reduce the word to its root word for easier processing while preserving the original word. Table 3 shows examples of input word and output stemmed word of Snowball Stemmer from NLTK. There are many stemmers, but NLTK's stemmer has the largest corpus of words and the better algorithm for difficult words.

POS Tag	Description	Example
JJ	Adjective	Bright
JJR	Adjective, Comparative	Brighter
JJS	Adjective, Superlative	Brightest
NN	Noun, singular or mass	Seat
NNS	Noun plural	Seats
NNP	Proper noun, singular	Kyle
NNPS	Proper noun, plural	Lakers
RB	Adverb	Naturally, brightly,
RBR	Adverb, comparative	Quicker, faster
RBS	Adverb, superlative	Best
VB	Verb, base form	Take
VBD	Verb, past tense	Took
VBG	Verb, present tense	Taking
VBN	Verb, past tense	Taken
VBZ	Verb, 3rd person present	Takes

Table 2: Examples of the Parts of Speeches that are will be utilized in the research

Stemming is the process of reducing inflected or derived words to their stem, base, or root from a certain word. Snowball Stemmer from NLTK performed the best through testing and provided accurate results. NLTK provides an understandable algorithm that makes it feasible to use. A

stemmer was used to increase the amount of words in the Deemed Credible file to further find similar words in the mass corpus of tweets.

Original Word	Stemmed Word
Running	Run
Crying	Cry
Vetoed	Veto
Cars	Car
Compromised	Compromise
Naivety	Naïve
Academically	Academic
Presumably	Presume
Literally	Literal

Table 3: Examples of NLTK's Snowball Stemmer. Stemmer utilizes a mass corpus of words to find the root or base of the word.

Expansion

In the parts of speech file, the words must be expanded to get the synonyms for the search method. NLTK has an import Wordnet from its large corpus of words. Wordnet is a "lexical database for the English language. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms" (Project, 2013b). These synonyms are grouped into synsets, which are interlinked by means of conceptual-semantic and lexical relations. This import provides a combination of dictionary and thesaurus that makes it feasible to support text analysis and artificial intelligence.

Each word in each file is selected and passed into Wordnet for similar words of the same meaning. These synonyms are recorded and also saved into its respective file. For the word best, Wordnet gives these results:

```
[Synset('best.n.01'), Synset('outdo.v.02'), Synset('best.a.01'),
Synset('good.a.01'), Synset('full.s.06'), Synset('estimable.s.02'),
Synset('beneficial.s.01'), Synset('adept.s.01'), Synset('dear.s.02'),
Synset('dependable.s.04'), Synset('effective.s.04'),
Synset('better.r.02'), Synset('well.r.01')].
```

Expanding the original word is necessary because the assumption if the user is tweeting similar content then it has the potential to be credible is used. Without the expansion or synonyms of words, it would be a simple search of the word occurrence in the corpus of tweets without any flexibility of words changing. After the expansion, a script is run to remove redundancy throughout each part of speech file.

Removal of Stop Words

Stop words were removed to reduce the amount of words that are irrelevant. Stop words causes problems when computing natural language by providing false positives in text. There are numerous stop words files but all words were combined to make one stop word file. Removal of stop words is common in all researches in query expansion, text mining, and natural language processing. By using this file of words, stop words were identified and removed from the parts of speech file.

Manual Editing

WordNet is not flawless because it does not provide consistent correct synonyms for all words. It has some erroneous words in which it will not provide correct synonyms humans would use. For example, the word “was” in WordNet would include Washington, be, exist, equal, constitute, embody, and cost. The words Washington, equal, and cost are synonyms that are incorrect. The user must go through the noun and verb part of speech file to delete the incorrect

synonyms that WordNet has provided. Manual editing takes minutes, without this step it would provide false positives and increase computation time.

In the noun file, words containing the topic are deleted. While gathering the tweets, the topic or subject was already filtered. For example, the subject is Boston bombing. Boston should be deleted because every tweet is about Boston by using Twitter's API method of filter.

Text Mining

A python script was made to search through the corpus of tweets to find certain keywords from the credible words file. The corpus of tweets is reduced to words within the tweet and returns the tweets that match the words in the parts of speech file. For hash tags (#) and mentions (@) at the beginning of a word, python's strip method was used to remove the special character. For special characters such as punctuation at the end of the word, python's strip method was also used. This is essential because some words contain special characters and without stripping the special characters, information in the search would not be accurate. Also a set of vulgar words was created to weed out invalid tweets. In this assessment, vulgarity is associated with lack of credibility, lack of proper use of words, and was deemed false. Word frequency was recorded for statistical data within the corpus. These tweets that match the words are then saved into a Verified text file. The tweets in the file have been deemed somewhat credible from the credible words using the automated method of text mining. While analyzing tweets, question words were used to determine if a user is credible. Words such as who, what, when, where, and how present questions and cannot be considered true. A user present a question to receive answers or asks a rhetorical question. In both instances, these are not considered credible because it isn't stating any information. Equation 1 was used to find the credibility and relevance score. The equation divides the number of words by the length disregarding the stop words.

$$CredScore = \frac{\#\ ofWordsFound}{\#\ ofWordsFoundInTweet - \#\ ofStopWords} * \alpha$$

Equation 1: Equation to Determine the Credibility Score of an Individual Tweet

Amazon Mechanical Turk

Today, computer scientists rely on computers to do majority of their work. Computers are made to make everyday living easier and complete tasks faster than humans with higher accuracy. With certain tasks researchers can depend on computers to complete assignments without questioning the results. However, in some cases humans outperform the most powerful computers in simple tasks. Some simple tasks include determining colors and shapes, detecting sarcasm, providing sentiment, identifying objects or words in a picture, and credibility. The validity of a statement is difficult for a computer to complete with high accuracy. Amazon Mechanical Turk provides helps humans to complete certain tasks by utilizing crowdsourcing.

> "Crowdsourcing is an act of outsourcing tasks, traditionally performed by an employee or contractor, which are now performed by a large group of people (the crowd) (Kuikkaniemi, 2010)". "Amazon Mechanical Turk provides a web services API for computers to integrate Artificial Artificial Intelligence directly into their processing by making requests of humans (Amazon Web Services, 2014)".

Amazon Mechanical Turk is the ideal way to incorporate crowdsourcing into our credibility research. In this approach, humans are required to conduct a Google search pertaining to a tweet

and make the decision. This slight change boosts accuracy and makes the approach unique. In previous research, users make the decision based on their knowledge of the user and topic. Users/Turkers understand reliable sources and certain sources disperse rumors as news. Using a crowd gives you a broader perspective of views and topics on a task. Having turkers from different regions gives results that translate to a larger scale. The data received from the crowd is not depended on demographics instead from the general public.

Using the crowd for services such as Amazon Mechanical Turk gives the requester flexibility. In the past, companies would have to sign long-term contracts in order to get tasks done. Companies would have to commit to an amount of money charged for the entire project of research. Today, contracts are not needed to get tasks completed. Developers can get the crowd to complete tasks for money agreed by the Requester.

The quality of the research increases when you depend on the crowd. Amazon Mechanical Turk allows the requester the ability to quiz the Turker. Requirements for the quiz can be used to ensure that the developer is collecting information from the people who have knowledge on the topic or task. Broadening the idea is another advantage of having the crowd complete tasks instead of receiving information in-house. When obtaining information from one group of people, the opinions and information would be similar.

The crowd's motivation to complete these tasks is money. Money compensates the users for their time and their efforts in completing the task for helping the developer's research. Turker's primary goal is to receive the most amount of money for their services in research. Another incentive is the pleasure of helping with tasks that they have interest in. Some users complete tasks because it is fun and truly enjoy the feeling of providing assistance for a good cause.

The government out-sources (crowd-sources) its issues to receive information, insight, and solutions from different perspectives. For tasks that are difficult and require experience the task has a higher reward. A reward for these difficult tasks can also be an employment opportunity. The requester can post tasks that are related to the job and serve as competition for a job. By correctly completing the difficult task, the turker builds his reputation by displaying his knowledge in the field.

Requesters of the service or "developers" submits tasks on the website for the crowd to complete. Later the requester approves the task and uses the data from users to provide insight on their overall task. Individual tasks, referred to as HITs (Human Intelligence Tasks), are tasks used to complete a larger problem.

In this research, crowd sourcing was utilized to determine if the statement is credible. The crowd on Amazon Mechanical Turk analyzed each tweet in the random set of 400 tweets that had a good score from the equation for words found in a tweet. The workers have 3 choices to categorize a tweet, credible, not enough information, and not credible. Each tweet will be completed 2 times to ensure optimal accuracy of credibility for each tweet. To ensure money is being spent effectively, a python script eliminated redundancy by removing re-tweets and similar tweets. Figure 17 shows the web page for the AMT task. The task is completed by selecting credible, not enough information, or not credible.

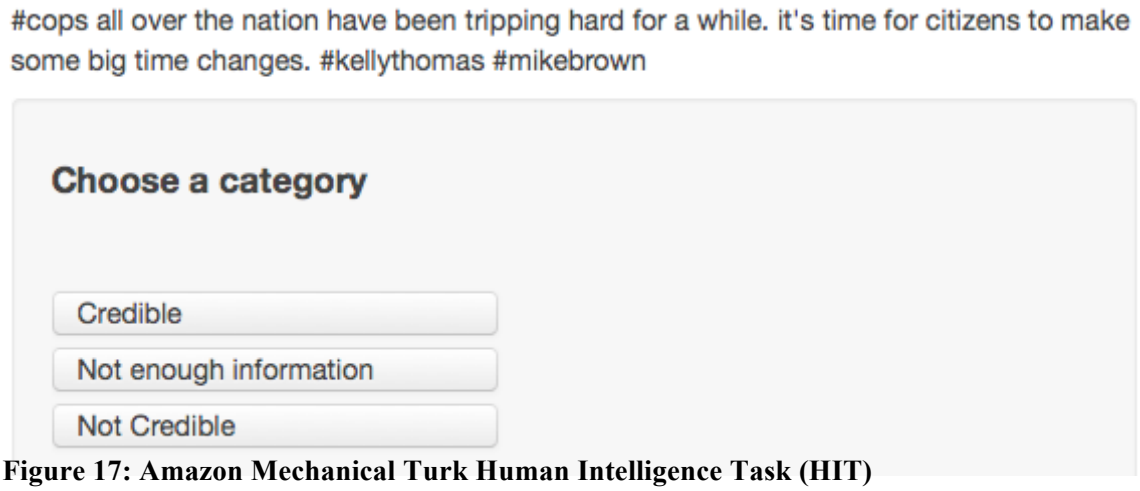


Figure 17: Amazon Mechanical Turk Human Intelligence Task (HIT)

OVERALL RESULTS

The top scores of tweet relevance from the equation,

$$CredScore = \frac{\#\,ofWordsFound}{\#\,ofWordsFoundInTweet - \#\,ofStopWords} * \alpha$$

, was used to sift through tweets. Tweets with higher scores was used and passed through Amazon Mechanical Turk for human computation. Tweets that were false with lower scores were also passed into Amazon Mechanical Turk. Users agreed on certain tweets and users divided tweets based on the crowd consensus of credible and tweets with false information.

Results from AMT

Results from each tweet were analyzed and "majority rules" concept was implemented to determine if the tweet was credible. Users agreed on 61% of tweets that were credible showing word searches have the potential to determine credibility. The remaining 39% of tweets that had optimal scores was determined false or not enough information. Users were told to use Google searches before making a judgment to aid in accuracy of their results. Results were accurate with

minor issues of selecting tweets in a shorter timeframe. This is from users creating robots to provide answers or users answering for the incentive to make the most money in a short time frame. Some results were incorrect but information was reviewed and correctly annotated to provide accurate information. AMT was a good way to judge the crowd on the credibility of a tweet. While reviewing the results, it still shows biases on certain positions. Humans tend to be biased unintentionally because of how strong users feel on certain subjects.

Feature Scoring Results

From the 122,411 unique tweets on the Michael Brown incident, results are shown on Table 4. Figure 18 displays a pie chart of all the unique tweet's score from features. The results were evenly distributed but no tweet or users obtained a positive 5 or negative 5. No accounts were calculated to have a perfect account that has ideal characteristics for all features to be considered credible. This calculation is the sum of verified account score, re-tweet score, favorite score, biography score, default profile picture score, fake follower score, Klout score, and friend and follower score.

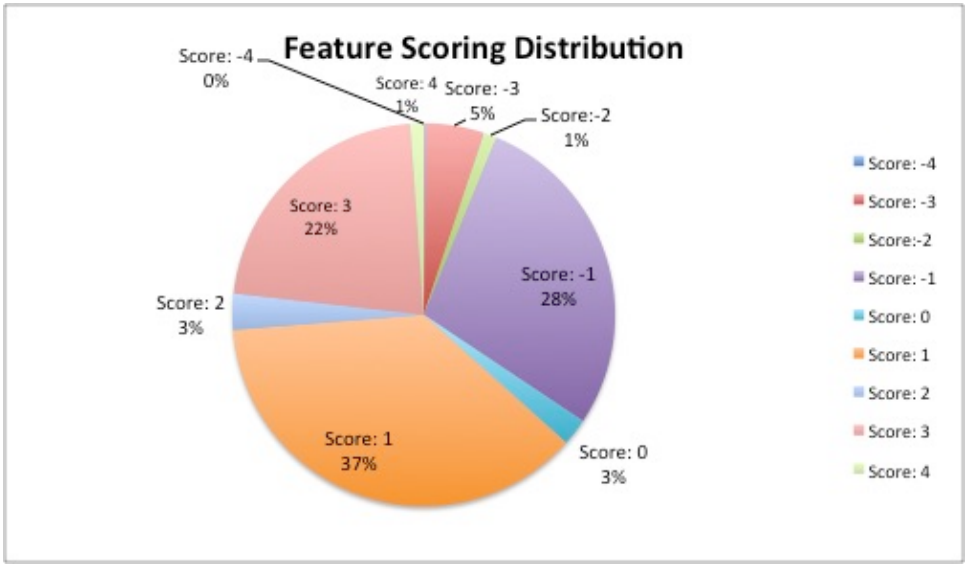


Figure 18: Pie chart of scoring based on features for all unique tweets.

-4	-3	-2	-1	0	1	2	3	4
189	6071	1336	34467	2809	45390	3728	27024	1397

Table 4: Raw data from Figure 11. The amount of tweets that has a score from -5 to 5

Text Mining Scoring

An equation was used to determine the relevance of a tweet by searching for words throughout a tweet. Figure 19 shows that majority of the tweets were not related to the information inputted into the Deemed Credible file or was filtered out by certain characteristics found. Certain characteristics such as vulgarity, presence of question mark, and length automatically disregarded the tweet. The length of tweets below 10 was filtered because there were many short tweets that had minimal information but had a good score. Table 5 displays the number of tweets that were in certain ranges.

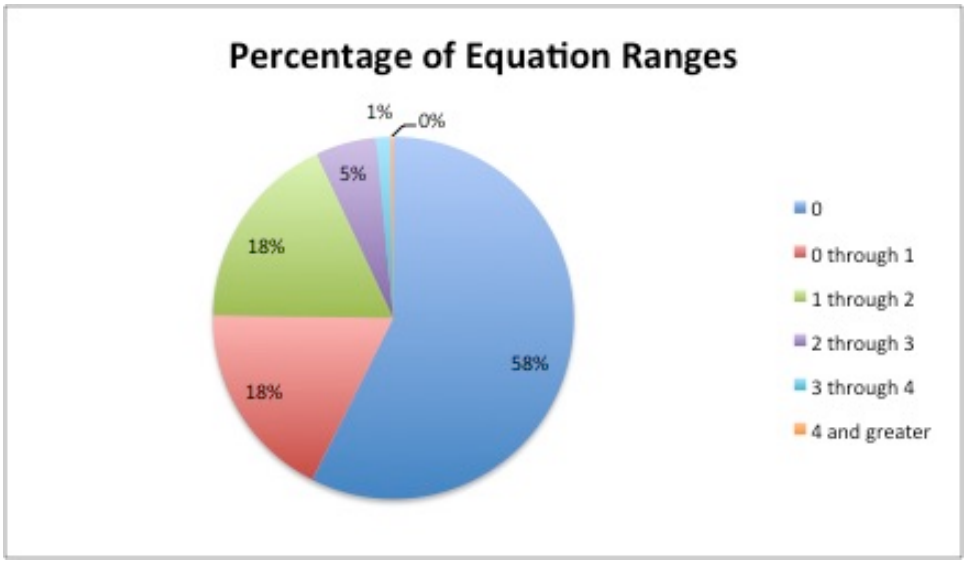


Figure 19: Pie chart of ranges for text mining scoring

0	0-1	1-2	2-3	3-4	4 <
70305	21790	21794	6627	1552	343

Table 5: Raw data from Figure 12. The score was calculated by the equation for text mining

Correlation of Methods

Figure 20 shows 35 credible tweet's scores that consist of feature, text mining, and a hybrid score, which is an average of the scores. The tweets that had optimal scores reflected that they were relevant to the topic and possessed information. Some information was truthful but most were opinions. For the tweets in the graph, some tweets show correlation between both scoring methods. The average of both scores was taken to use both scoring techniques to determine if both can determine credibility. For credibility using the search score, the scores were higher and showed that higher search scores present relevant tweets and have a higher possibility of being credible. Highest scores don't always predict correctly but it provides tweets that are about the topic and are relevant. There was a negative correlation of -0.280 between feature and search scoring, positive correlation of 0.907 between hybrid and feature scoring, and positive correlation of 0.148 between hybrid and search scoring with credible tweets.

Figure 21 shows 35 tweet that were deemed not credible scores, which are feature, text mining, and hybrid score. It was found that tweets that were not credible possessed lower scores. From the results in Figure 20, tweets that were credible had higher scores opposed to tweets that were deemed not credible. Figure 21 has a negative correlation of -0.255 between feature and search scoring, positive correlation of 0.895 between hybrid and feature scoring, and positive correlation between of 0.202 between hybrid and search scoring.

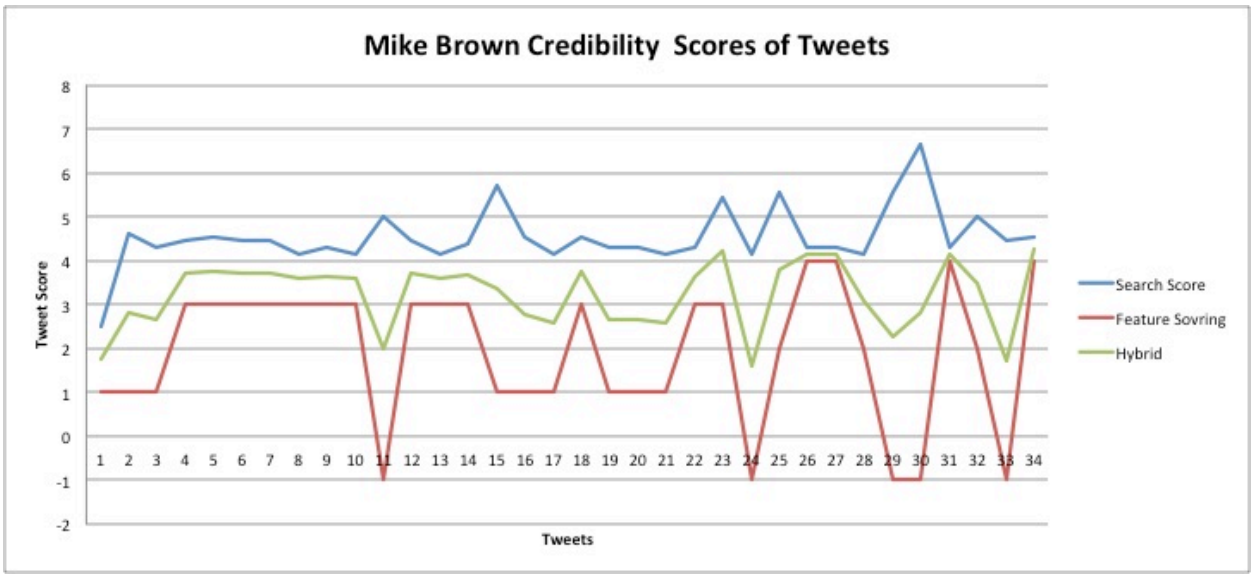


Figure 20: Scores of 35 Tweets that were Deemed Credible by AMT

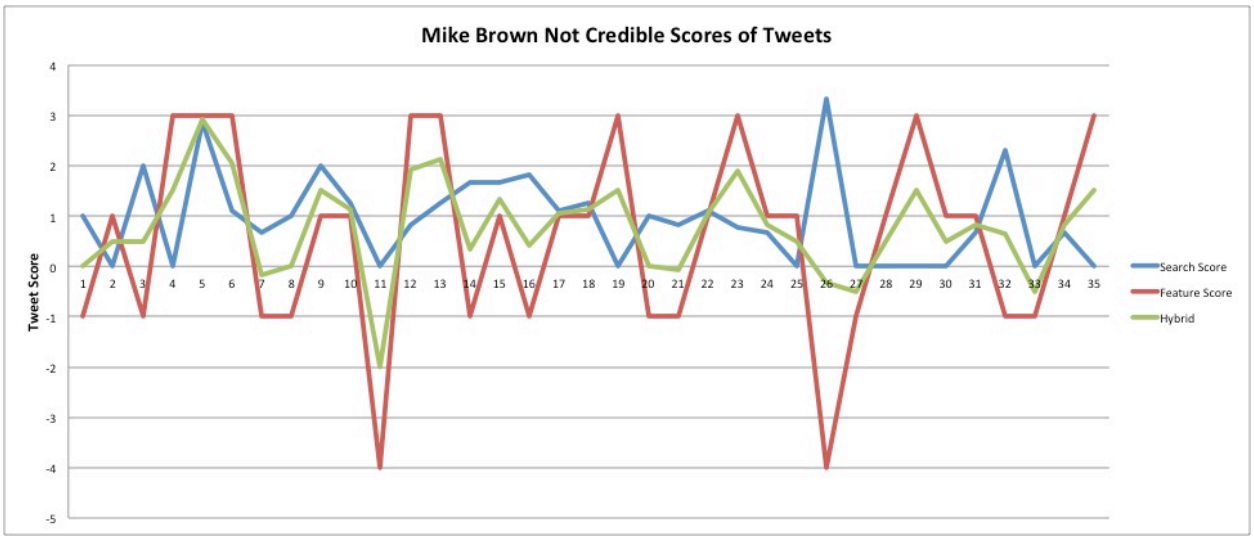


Figure 21: Scores of 35 Tweets that were Deemed Not Credible by AMT

CONCLUSION

Many researchers have used equations to score a tweet's credibility using features and text similarity. This approach utilizes all major features to provide the ideal user to receive information from. Users that have fake followers are prevalent in today's society because of the emergence of user's need to feel popular. Fake profiles and inactive users provide false information regarding a person's network. Fake friends were addressed by an application that detects fake friends. Biography was addressed by viewing if it has been updated, and the default profile picture was viewed as if it was a fraud account. Klout score was used to determine the activity level of the accounts. Many features were addressed in order to make an ideal profile that would be considered credible like similar researchers have done.

Text mining for certain keywords was used to get relevant tweets that could possibly be credible. With the information pertaining to the lack of believability of Twitter, the approach incorporates outside information from credible sources. News agencies, criminal and domestic release records, and verified twitter profiles were used as the golden standard. Keywords were chosen and manually annotated to provide accurate terms to provide a score.

From the results of the scores, it was found that tweets that had higher text mining scores had the higher potential of being credible. Tweets that had optimal scores were relevant to the topic and provided more information regarding the event. For select tweets the score of feature and text mining scoring were very correlated by having users that had suspicious features and false information. Overall, the results show that the account features do not affect the credibility of a tweet and the text mining approach works independently.

Tweets that had a lower score were more likely to possess information not regarding the event and more likely to be false. Therefore, this method can provide tweets that have a higher

possibility of the information being correct. This method can also deliver tweets that are irrelevant and opinions. The relevancy is determined from the news articles placed in the file. There is a chance that new information arises or truthful information was not added to the file to search through tweets. If this occurs, the file must be updated to reflect the new information that was not added in the initial search. By expanding the keywords in each search, the credibility and relevance increases.

FUTURE WORK

This research can be expanded to comments and blogs to determine credibility. For larger text like blogs, a method requires a machine language approach. Sentiment analysis can also be implemented to incorporate all studies into one method. By using all methods, it will provide more information on the text to make an accurate prediction of credibility. Bigrams can be utilized to provide an increase in accuracy of text searching. Bigrams will provide more ways to retrieve relevant and possibly increase the credibility text mining method.

SocialBakers Fake Followers application is a beta application and is not a finished product. The application only uses 100 followers to analyze therefore, the application is limited in the amount of data it processes. The application will continue to get better as it increases the amount of followers it implements. Twitter Counter is an application by Twitter to plot the amount of friends added in a particular time. The application shows how fast friends are added and to determine if fake followers were brought. Twitter Counter does not provide an API and requires users to manually view each profile graph of friends to determine if the followers' are suspicious. A monetary fee is required to view profile information regarding friend and follower activity from the creation date. The information is updated hourly and provides accurate information. If this is implemented in the future, researchers can determine if users have

purchased fake friends. A method must be created to find information that has not come from news agencies. Using this research, information that was credible was searched for rather than searching for new information that was credible. Overall, Twitter credibility will always be an ongoing research topic. As users rely more on social media to get their information, they need to know what is true.

WORK CITED

3am. (2009). Peaches starts Miley Cyrus internet death hoax. What a clever girl. *Mirror*.

Amazon Web Services, I. (2014). Amazon Web Services FAQs. Retrieved October 22, 2014, 2014, from http://aws.amazon.com/mturk/faqs/

Angelova, V. (2013). 7 Reasons Why Socia Media is an Absolute Must for Business Branding. *Likeable Media*.

Becker, L. B. (2014). *Social Media vs. Traditional Media*. School of Journalism. UGA.

Canini, K. R., Suh, B., & Pirolli, P. L. (2011). *Finding credible information sources in social networks based on content and social structure.* Paper presented at the Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom).

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.

Chernin, P. (2014). Twitter. from https://twitter.com/

Constitution, 1 (1789).

Dictionary, M. W. s. (Ed.) (2014) Merriam-Webster.

Foundation, P. S. (1995). Python Launguage Reference, version 2.7.

Foundation, P. S. (2014). HOWTO Fetch Internet Resources Using urllib2. Retrieved October 23, 2014, 2014, from https://docs.python.org/2/howto/urllib2.html

Gahran, A. (2011). Most Twitter users access service from their phones, study says. *CNN*.

Goode, L. (2009). Social news, citizen journalism and democracy. *New Media & Society*.

Greene, R. A. (2011). Secret Service to probe Fox News' Twitter hack. *CNN*.

Huckerby, M. (2006). *The Net for Jounalists: A practical guide to the Internet for Journalists in Developing countries*. Paris, France: Thomson Foundation.

Jesse Holcomb, J. G., and Amy Mitchell. (2013). News Use Across Social Media Platforms. *Pew Research*.

Kang, B., O'Donovan, J., & Höllerer, T. (2012). *Modeling topic specific credibility on twitter.* Paper presented at the Proceedings of the 2012 ACM international conference on Intelligent User Interfaces.

Klout, I. (2014a, October 22, 2014). The Klout Score. from https://klout.com/corp/score

Klout, I. (2014b). Klout: Be Known for What You Love. 2014, from https://klout.com/home

Kuikkaniemi, K. (2010). White paper: Crowdsourcing in media industry. *Helsinki Institute for Information Technology, 16*.

Lasica, J. D. (2003). What is Participatory Journalism? *Online Journal Review*.

Legislature, F.-e., & Session, F. R. (2011). SENATE BILL 1062.

License, C. C. A.-S. (2014). Wikipedia The Free Encylopedia. Retrieved November 1, 2014, 2014, from http://en.wikipedia.org/wiki/Main_Page

Mabweazara, H. M., Mudhai, O. F., & Whittaker, J. (2014). *Online Journalism in Africa: Trends, Practices and Emerging Cultures*: Routledge.

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

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

Mueller, K. (2014). 11 Reasons Why People Love and Use Social Media. *Inkling Media*.

O'Reilly, T., & Battelle, J. (2009). *Web squared: Web 2.0 five years on* (Vol. 20): O'Reilly Media.

Outing, S. (2005). The 11 Layers of Citizen Journalism. *Poynteronline*.

Project, N. (2013a). NLTK 3.0 documentation. Retrieved November 4, 2014, 2014, from http://www.nltk.org/

Project, N. (2013b). WordNet Interface. Retrieved November 4, 2014, 2014, from http://www.nltk.org/howto/wordnet.html

Roesslein, J. (2014). Tweepy:An easy-to-use Python library for accessing the Twitter API. Retrieved October 22, 2014, 2014, from http://www.tweepy.org , https://github.com/tweepy/tweepy

Schmierbach, M., & Oeldorf-Hirsch, A. (2012). A little bird told me, so I didn't believe it: Twitter, credibility, and issue perceptions. *Communication Quarterly, 60*(3), 317-337.

Schweidel, W. W. M. a. D. A. (2014). *Social Media Intelligence*. Cambridge, England: Cambridge University Press.

Shariff, S. M., Zhang, X., & Sanderson, M. (2014). User Perception of Information Credibility of News on Twitter *Advances in Information Retrieval* (pp. 513-518): Springer.

SocialBakers. (2014a). Company Overview. Retrieved October 23, 2014, 2014, from http://www.socialbakers.com/company-overview

SocialBakers. (2014b). Methodology. Retrieved October 23, 2014, 2014, from http://www.socialbakers.com/twitter/fakefollowercheck/methodology

Soup, B. (2014). Beautiful Soup Documentation. Retrieved October 23, 2014, 2014, from http://www.crummy.com/software/BeautifulSoup/bs4/doc/

Stadd, A. (2013). What To Use Twitter's Favorite Button For. *All Twitter*.

Twitter, I. (2014a). FAQ'S about verified accounts. Retrieved October 22, 2014, 2014, from https://support.twitter.com/articles/119135-faqs-about-verified-accounts#

Twitter, I. (2014b). Twitter Developers. from https://dev.twitter.com/rest/public

Twitter, I. (2014c). Twitter:About. Retrieved October 24, 2014, from https://about.twitter.com/company

Westerman, D., Spence, P. R., & Van Der Heide, B. (2012). A social network as information: The effect of system generated reports of connectedness on credibility on Twitter. *Computers in Human Behavior, 28*(1), 199-206.