DATA-MINING SOCIAL MEDIA FOR SPATIOTEMPORAL

PATTERNS OF NEGATIVE OPINION

By

Chris Cantey

A thesis in partial fulfillment of the
Requirements for the degree of
Master of Science
(Cartography and GIS)

At the

UNIVERSITY OF WISCONSIN-MADISON

2013
Acknowledgements

I would first like to thank my academic advisor Dr. A-Xing Zhu of the University of Wisconsin-Madison Geography Department. He had the unenviable task of keeping me focused and did it masterfully. He kept me motivated and pushed me exactly when I needed his encouragement. I will always be grateful for his expertise, guidance and support throughout this process. I would also like to thank my committee members Dr. Robert Roth and Dr. Jim Burt. Dr. Burt, Jamon Van Den Hoek, Jing Gao, Fei Du, and Jing Liu all provided critical feedback throughout the early stages of my thesis and helped me workshop my ideas. Dr. Roth is my cartographic mentor; the University’s rich tradition of cartographic excellence is in good hands.

I would also like to thank my cousin-in-law, Stefan Nelson-Lindall of Google and Steven Hemmy of UW DoIt for their technical expertise, as well as all of my former cartography students, and my cohort and friends in the Geography Department. This research was supported in part by the Trewartha Research Grant.

Most of all thank my wife and son, Kristine and Aiden. Boy have we come a long ways. I couldn’t have achieved anything without them.
Table of Contents

Chapter One: Introduction .................................................................1
  1.1: Research Question ......................................................................2
  1.2: The Significance .........................................................................2
  1.3: Research Approach ....................................................................5
  1.4: Organization of Thesis ..............................................................6

Chapter Two: Methodology .................................................................8
  2.1: Overview of Methodology ..........................................................8
    2.1.1: The Social Media Platforms .............................................8
  2.2: Selection of Social Media: Twitter ..............................................12
    2.2.1: Reasons for Twitter .........................................................12
    2.2.2: Provisions of Twitter Data ...............................................14
  2.3: Discovering Context .................................................................15
  2.4: Case Study .................................................................................18
    2.4.1: Hurricane Sandy and the Government’s Response ...........19
    2.4.2: The 2012 Election ............................................................22
  2.5: Data Analysis .............................................................................23
    2.5.1: Database Schema ............................................................23
    2.5.2: Filter by Keyword ............................................................24
    2.5.3: Frequency Analysis .........................................................26
    2.5.4: Socioeconomic Overlay ..................................................27

Chapter Three: Results and Discussion .............................................30
  3.1: Spatio-Temporal Patterns of Negative Opinions .......................30
  3.2: Median Household Income .......................................................34
  3.3: Median Age ..............................................................................37
3.4: Race ........................................................................................................39

Chapter Four: Conclusion and Future Direction .................................42
  4.1: Conclusion ..........................................................................................42
  4.2: Future Direction ..............................................................................44

Works Cited ................................................................................................46
Appendix ......................................................................................................50
1. Introduction

Communication over the Internet has evolved from a static entity, where people passively retrieve information from websites, to something more collaborative. Today we share information through Web 2.0. This new communication model can be partially defined as interactive websites and applications that allow for public participation in the sharing of information along with the creation of personal web pages and rich user experiences (O’Reilly, 2005).

We are now active participants in the spread of information over the internet. We navigate to social media sites, blogs, wikis, and other news threads to receive and distribute content that was, up until a few years ago, found exclusively within the domain of traditional print and television media outlets.

This expansion of sources of information presents new opportunities to share knowledge and to collect data, which previously could only be compiled using exhaustive surveys to gather opinion data. Using web application programming interfaces (API) along with modern geographic information systems, it is now possible to understand the spread of information, and determine who is able to receive and pass along this new wealth of information.

Social media behavior leaves a particularly deep digital footprint. Our online actions, conversations, social networks, and every keystroke or mouse-click is recorded digitally. And whether it is an implicit mention of a geographical place or time, or an explicit geo-referenced message, every conversation we have on social media has a unique geographical tag and time stamp attached to the message. With this
information, we can analyze, and map with reasonable confidence events\(^1\) and themes as they happen.

1.1 Research Question

Given that political and natural events cause unrest in social media, the research of this thesis serves to address the following question:

*Using census demographics and the geographic signature of social media content, can we determine what effect major political and natural disaster events have on public attitudes towards the American government?*

1.2 Significance

Although the body of literature is growing on social media, much less is known about the spatial and temporal components of this information. A location-based social network affords geographic information scientists the opportunity to collect real-time spatiotemporal data, provided they can set-up workflows to retrieve, validate, and organize this information (De Longueville et al., 2009). In their paper, De Longueville and colleagues study how social media can be used as a source of spatiotemporal data and demonstrate its ability to possibly serve a role of support in emergency planning, risk assessment, and damage assessment activities.

\(^1\) For the purpose of this paper, the term "event" will refer to any theme, topic, crisis, or any other item of significance or happening that forms a geographical pattern in the social network.
Prior research has investigated social media’s potential to locate areas of impact in a natural disaster (Li et al., 2011, Palen et al., 2012, Starbird & Palen, 2010, Pozdnoukhov, et al., 2011) or to identify natural disaster events, as they happen in near real-time (Lee & Sumiya, 2010, Sakaki, 2010, ). It has yet to be determined how these types of natural disaster events affect public discourse, especially as it applies to those who have an impact on response and the policies that directly affect the population.

In the past, before the rise in popularity of social media networks, it was difficult to gather general public opinion data. In many cases, opinions were hidden in long forum posts, blogs, or anonymous news threads. Social networks have dramatically changed the way that people express their views and opinions. With an overwhelming volume of social media content to analyze, it can be a difficult task for a human reader to find relevant sources, extract related sentences with opinions, read them, summarize them, and organize them into usable forms (Liu, 2010). To efficiently analyze this data, an automated opinion indexing system is needed.

Understanding spatiality in the social network will enable public officials, policy makers, emergency management teams, and businesses to act and respond to criticism levied towards them in these highly volatile situations. An understanding of the communication in the social network will allow these officials to push information or allocate resources to the geographical regions in a response to the user’s critiques.

In the absence of traditional media outlets, whether the absence is due to the delay in the relay of information from the source to the network, strain on the system, or due to large volumes of caller activity, timely information to the public can be lost.
Perhaps due to structural collapse, such as an earthquake, terrorist attack, or civil unrest, the public loses its normal flow of information.

Around the world, many media outlets are under strict control of authoritarian governments, where censorship laws and a tightly controlled state-run media do not have permission to allow certain information to diffuse. In each of these cases, the tools developed in this research could be used to aid emergency response to a disaster or address public hostility towards the government, potentially saving property and lives.

While research suggests that local media and established emergency management agencies continue to be valued sources for information, as most messages in the social media originate with established news sources, the research also emphasizes the growing contribution individuals make to present timely spatial information in a crisis event (Starbird & Palen, 2010). It has yet to be determined if the event of a structural collapse would make social media users the main source of crisis information. However, the Arab spring and Boston Marathon bombings provide examples of social media being used as primary sources of information (Rogers, 2013), for better or worse.

Finally, in the wake of the revelations that the National Security Agency (NSA) collects contacts from people’s personal e-mail and instant messaging accounts (Gellman & Soltani, 2013), monitoring social media offers a far less intrusive means to monitor our digital activity. Social media privacy awareness has become ubiquitous today, as users are generally aware, and often times must acknowledge consent that they are posting comments for the entire public to disseminate.
1.3 Research approach

The emergence of social media, and the prior research on the content these networks can provide, will enable me to develop analytical tools. To address the above research question, I developed applications that leverage web programming API’s, GIS methods, and analytical cartography, to determine who is voicing negative opinion towards the government.

The key to understanding collective public sentiment is to collect and analyze the textual messages and spatial information from social media over an extended period of time. I make use of two case studies—one for the 2012 US Presidential Election and one for the 2012 Hurricane Sandy event—to demonstrate the utility of spatiotemporal analysis of social media.

I then determine if the messages discusses the government and whether or not they have a negative sentiment. The geographic information embedded in the messages metadata is used to relate the social media content to geographic census data (e.g., poverty, age, race, and population) using simple statistical measures, visual pattern matching, spatial overlay, and the assumption that most people use social media close to their residence or in a geographic area that has similar attributes as their home census block.

If they have negative sentiment towards the government, I examine whether or not there is a consistent pattern between normal social media behavior and the behavior that occurs during major political and natural events. Each of these events certainly will create a disturbance in the natural ebb and flow of the social network.
Therefore, it will be possible to determine how communities react towards their government in each of these scenarios (Figure 1).

1.4 Organization of thesis

The remainder of this thesis will present my initiative to enrich social media metadata by using the geographic content embedded in each message. Section 2 describes the methodology, divided into four subsections. First a survey of the social media platforms Twitter², Facebook³, and Foursquare⁴ is provided. Next it covers methods directly related to geographical theme extraction on social media, along with the most common issues that need to be resolved in order to pursue this line of research. This is followed by a section on the Twitter⁵ API and Gnip⁶, the exclusive provider of historical and unlimited Twitter activity. Next I discuss Hurricane Sandy and the 2012 Election, the case studies I decided to test. Then I discuss the data preparation and processing, and the means of temporal and spatial analysis. Section 3

---
² www.twitter.com
³ www.facebook.com
⁴ www.foursquare.com
⁵ www.dev.twitter.com
⁶ www.Gnip.com/twitter
presents the results of my tests and Section 4 concludes this thesis and presents future directions of research areas.
2. Methodology

2.1 Overview of methodology

2.1.1 The social media platforms: Facebook, Twitter, and Foursquare

Spatiotemporal research within this field is certainly quite new, given the recent growth of social media. The social media network Facebook was launched in 2004, while Twitter was released in 2006. But the popularity of these social networks has grown exponentially; Twitter grew 1460 percent between June 2008 and June 2009 to reach a global audience of an estimated 44.5 million visitors (Scanfeld, et al., 2010).

With more than 800 million users, Facebook would be the ideal network to examine. Here the relationships between the users must be reciprocal. In order to be "friends" with someone, each person must accept the friendship. Becoming friends allows users to read and comment on posted messages that their friends post on their personal Facebook page, and in turn your friend can do the same. There are very few limits as to what someone can post to Facebook. A user can post messages of unlimited length, photograph images, videos, or links to other internet pages.

The sum of the network is a "social graph?», generalized in Figure 2. The social graph refers to: 1. Community, our network of friends. 2. Conversation, the interactions users have with each other. 3. Identity refers to ourselves and how we are viewed within the community8. The Graph application programming interface (API) presents a simple, consistent view of the Facebook social graph, uniformly representing

7 www.developer.facebook.com
8 Social design described by Facebook
objects in the graph (e.g., people, photos, events, and locations) and the connections between them (Facebook). Facebook allows third party developers to access the Graph API using their own applications access to any information that Facebook users allow with the applications access request shown in Figure 3.

Authentication, or access permission, is not required of users that have public profiles. Unfortunately, for the purpose of studying Facebook relationships, the Graph does not allow the location field of public profiles to be viewed. In order to have access to profiles with their locations, developers need permission from the users, using a Facebook application. The Facebook Graph API therefore was not considered for this research, given the prohibitive amount of time necessary for developing an application and to acquire a representative sample of users accept the application in time to conduct the research.

Twitter, however, allows every message sent out on their network to be viewed by anyone using their API. It is almost an entirely public social network. It is possible to
create a private account and to block people from viewing this personal account, but any account can be accessed through Twitter’s development environment. This of course raises the issue of Internet privacy. For this research I do not seek out private accounts and therefore do not intend to explicitly address the issue of privacy.

Twitter and Facebook operate in very different ways. Both networks tend to be dominated by a smaller percentage of active users who make one or more postings a day. Many other users are considered passive users who may check the network daily, but post messages less frequently. This is referred to as degree centrality, where the importance of a node is determined by the number of nodes adjacent to it. High degree nodes naturally have more impact to reach a larger population, thus they are considered more important (Lei & Liu, 2010).

Degree centrality could be thought of as a hub and spoke distribution. Research on the two websites has not confirmed this relationship difference quantitatively, but it is widely assumed that Twitter has many more "hub" users than Facebook. That is because relationships in Twitter do not have to be reciprocal. A person can usually "follow" any user he or she wants to follow, without any acceptance or rejection available to the person being followed. Therefore, celebrities and other public figures may have many thousands, or even millions of followers, while many other users, if not most, may have less than fifty, thus giving them much less influence over the network.

Twitter is often referred to as a micro blogging service because it allows the users to post messages, known as tweets, using fewer than 140 characters. This is convenient because the messages aren't well formed comments or questions; rather
they are concise status updates that followers can understand (Sankaranaarayanan et al, 2009). Unfortunately this can lead to abbreviations and acronyms that may be harder for a search algorithm to interpret.

Users can specify a theme to their message using hash tags (#), which provide a valuable service to anyone attempting to gather contextual data, as these hash tags can assist in classification and filter the stream of messages. In addition, hash tags are also used by Twitter to calculate the trending topics of the day, which encourage the user to post in these communities (Li, et al., 2011).

Finally, the user can geo-tag their message in two different ways. The geo-tagging feature of Twitter assigns a physical location to the message metadata. A very small number of users, about 3–5 percent (Gnip, n.d.), actually opt in to this feature, which assigns longitude and latitude to each tweet using a cell phone's GPS signal, but the location data is very accurate.

Users may elect to add a location to their personal information that appears on the user’s landing page as well as the tweet’s metadata. However, many users lie about their location and often place a humorous message in the location field. Other times users may list their home as their location but they may tweet from work, school, or while travelling, making this data erroneous. Therefore, in order to use this field, researchers would have to verify the location of the message either by associating the user to his network of users to infer the location (Davis, et al., 2011) or by extracting text that refer to a geographical entity to geo-code the message.
Unlike the previously mentioned social media sites, Foursquare\(^9\) is a location-based social networking site developed specifically for mobile devices. This network is explicitly geographic, which would be valuable for this research. However, user expressions and relationships in Foursquare are much different than typical “real world” expressions and relationships. Foursquare is a game-theory designed social network that establishes user hierarchy based on how often users check-in to places. Individuals accrue points by checking in, the network also encourages its users to post relevant information about the locations they have checked into.

With its own API and developer terms of agreement, the data Foursquare provides could be valuable for analyzing location data and user movement. But for my research, I found that Foursquare did not contribute enough opinion data. Users have the ability to publish their Foursquare check-ins to Twitter and Facebook, and in preliminary testing of the networks, I found that the data I received from check-ins were largely generic, for example, "User checked-in to location". This type of message does not contribute knowledge and actually adds a great deal of noise to the database.

2.2 Selection of Social Media: Twitter

2.2.1 Reasons for Twitter

A comprehensive literature review of existing research and preliminary testing of each social media site’s API methods led to the conclusion that the Twitter API

\(^9\) www.foursquare.com
would be the best choice for mining social media data. Due to the large volume of tweets published on Twitter, topic modeling and content-based studies nearly exclusively utilize Twitter (Kireyev et al. 2009, Pozdnoukhov et al. 2011, Luo, et. al, 2011). My goal is to exploit Twitter's short message syntax and its easy-to-use API to extract negative opinion content.

Twitter's text limit of 140 characters and use of hash tag topic classification will simplify the task of government keyword identification and sentiment analysis. For instance, "#Obama2012" is clear sentiment that the author supports the President in the 2012 election, “#Sandy” or “#HurricaneSandy” immediately identifies the tweet as being on topic to the hurricane event.

The Twitter API, with its clear API documentation and limited restrictions on data access, provides the best resources for developers. The Twitter API is documented even further in the open-source community of developers that continuously post insights, code snippets, and best practices to their personal sites and blogs, and in even greater detail to code repositories such as Github\(^{10}\) and Stack Overflow\(^{11}\). Most importantly though, Twitter provides the best means of recording both location data and user sentiment, while exposing the greatest number of users, in the easiest to read format.

\(^{10}\) www.github.com
\(^{11}\) www.stackoverflow.com
2.2.2 Provision of Twitter Data

Twitter produces more than 200 million tweets a day and allows developers’ access to its freely available streaming API. The Streaming API is an approximately one percent random sample of the full streaming Twitter feed (Morstatter, et al, 2013). It is currently unknown as to how Twitter samples the data for public distribution. In their comparative analysis of the Streaming API vs. the Fire hose API, Morstatter and his colleagues found that the sampled dataset generally followed a one percent sample of the full stream, but surprisingly, when Twitter activity increased dramatically, the Streaming API actually decreased its coverage.

In order to yield a dataset that covers a particular natural and political event in space and time, it was determined that the full fire hose method would best serve my interests. Gnip is one of two exclusive Twitter business partners who offer this service for a moderate fee, depending on the purpose and volume of the request.

Purchasing the 100 percent full Twitter stream removes any ambiguity that is presented with Twitter's sampling methods. It also allows the ability to query terms post hoc, an important function because it is impossible to determine when and where a natural disaster will occur and what affect it will have on popular sentiment. Gnip allows the ability to review precise locations and historical time periods that span Twitter’s existence.
2.3 Discovering Context

A real challenge to my objective is the ability to filter latent noise. One can extract relevant topical information by identifying "seed users" such as newspapers, television stations, reporters, and bloggers (Sankaranarayanan, et al., 2009). According to one study, 10 percent of Twitter users are responsible for more than 90 percent of all tweets (Heil & Piskorski, 2009). To detect unusual geographic events (Lee & Sumiya, 2010; Li, et al., 2011), monitored crowd behavior on Twitter by establishing a "normal" baseline crowd behavior in a geographic region of interest. Sudden increases in twitter activity were taken to indicate crowd abnormality.

There is ample research suggesting that people use social media during a crisis to comment and share information about an abnormal event. Using trained data, such as predefined keywords like "shaking" or "earthquake" (Sakaki, et al., 2010), the geographic epicenter of earthquakes have been located with a high degree of accuracy.

Using a predefined lexicon method of qualitative theme extraction with search terms such as "antibiotic" or "flu", Scanfeld et al. (2010) determined that social media sites offer a means of health information sharing; furthermore, such tools could potentially be used to gather important real-time health data, combining health status updates with location-based information. To track outbreaks, for example, it would be relatively easy for a health organization to direct people to submit Twitter status updates with symptoms and location data using a predefined format so that the updates are machine readable and easily mapped (Scanfeld, et al., 2010).
The context of a message is crucial to the analysis of an event. Parsing a message, or any text string, is the process of analyzing the text syntactically to determine the grammatical structure. Rather than using a predefined lexicon, as described by Sakaki et al. (2010) or Liu (2010), it is possible to deconstruct text strings into elements of syntax to extract the semantics.

Researchers at Stanford developed an open source natural language processor\textsuperscript{12} (NLP) whose algorithm produces the correct classification of natural sentence at rate of 86 percent (Klein & Manning, 2003). This provides a framework to map phrases into "WHAT", "WHERE", "WHEN", the semantic reasoning that is a core component of theme discovery.

Mei, et al., (2006) proposed a model to discover spatiotemporal theme patterns of Weblogs\textsuperscript{13}. They determined themes using the count of each word in a blog, at a given time and location, and dividing it by the total word count to assign a proportion. The words in the weblogs can be classified into two groups, common English words such as "the", or "a", and words related to the global subtopic that they are interested in analyzing.

To extract sentiment from a constructed string, Hu and Liu (2004) identified opinion words that are used to express subjective sentiment. To determine consumer opinion from product reviews, they first identified features of the product about which customers expressed an opinion. For each feature, they identified a review sentence

\textsuperscript{12} http://nlp.stanford.edu/software/lex-parser.shtml provides an open source toolkit that allows the parsing of 5 natural languages: English, German, Chinese, Arabic, and French.

\textsuperscript{13} Often referred to as "blog", they can be written by any individual on a wide variety of topics.
that gave a positive or negative opinion, and then they produced a summary using the count of positive and negative opinions (Hu & Liu, 2004).

Hu and Liu used a lexical and corpus based approach over the course of many years to create a list of opinion words (See Appendix). They used their lexicon to sample consumer reviews of five items sold on two websites. A manual inspection of the sampled reviews determined that their method of sentiment analysis achieved a successful opinion classification with 84 percent of the reviews it analyzed.

A predefined lexicon of keywords approach (Sakaki, et al., (2010); Scanfeld, et al., (2010)) that define a sample of the influential government agencies and elected officials (i.e., “Obama”, “FEMA”, “NYPD”) is sufficient to identify my theme target. Furthermore, the brevity, unique short-hand text, and the acronyms used by Twitter users do not necessarily need to be analyzed by complex NLP programs. Neither is it an easy task to implement these types of machine-learning algorithms in the time needed to conduct the analysis.

Therefore, I chose to use the Hu & Liu (2004) opinion lexicon that targets Twitter status updates that mention certain high-profile, local, state, and federal government officials or agencies in the body of text. A key challenge in using this methodology is the selection of appropriate terms that describe the government officials and agencies.

I populated my government keyword lexicon with a list of all elected federal and state officials, as well as the mayors that represent the study area. A similar technique was used to select the government agencies. I chose to include all federal and state
emergency management agencies, along with agencies that were active or effected by the hurricane (Figure 4).

Furthermore, many terms chosen were excluded after analyzing a large sample of tweets and performing a frequency count of the terms. I removed status updates from the government related keyword lexicon that occurred infrequently.

2.4 Case Study: New York City, Hurricane Sandy, and the 2012 Election

New York City was carefully selected as a region of interest for two major reasons. First, having the benefit of hindsight, the election, and more importantly the hurricane, were highly publicized political and natural events, whose dates and impact geographies were well known.

Secondly, New York is the largest city in the United States. This presents a unique opportunity to understand how these types of events effect different social
classes. As its largest city, New York is one of the most socially and economically
diverse cities, not only in the U.S., but around the world. The chosen study contains
more than 6,500 census block tracks (U.S Census, 2010) that contain metrics about
age, race, and income. There are very few places in the world that offer such a large
slice of life to evaluate.

The particular bounding box of the Twitter data was selected for reasons both
geoGraphically and economically. The main objective was to cover as much of New York
City and its range of socio-economic data. But with the price of the fire hose method
provided by Gnip determined by the volume of records, a compromise was applied to
my search area. The bounding box was created using the center of New York City with
an approximately 15 mile wide (distance along latitudes) by 20 mile high (distance
along longitudes) rectangle.

For this research, all geo-tagged Twitter activity from October 6, 2012 to
November 26, 2012, in the New York City area, with a geographic bounding box of
(Longitude: -74.0424, Latitude: 40.5649) by (Longitude: -73.7622, Latitude: 40.8657)
ensuring that the election and hurricane events would be captured, and that the normal
baseline activity prior to and after the events would also be obtained. Using these
parameters and post hoc data analysis, the specific area of interest would be covered.

2.4.1 Hurricane Sandy and the Government’s Response

Hurricane Sandy was one of the deadliest and costliest storms in U.S. history.
An unusually late hurricane season storm for the northeastern portion of the United
States, Sandy’s path of destruction effected millions of lives in the United States and
the Caribbean Sea, killing approximately 286 people in seven countries, including 117 in the U.S. and causing approximately 68 billion dollars in damage (Petrecca, 2013).

Sandy began as a tropical depression that formed in the Caribbean Sea on October 22, 2012. Two days later it strengthened into a Category 1 hurricane and moved out to sea on October 27th. As the storm moved northeast, parallel with the Eastern Seaboard, it was intercepted by high-pressure cold front from the north, forcing it trajectory towards the northeastern United States (Figure 5).

By midafternoon, October 29th, Sandy’s diameter was around 1,000 miles wide, with the eye of the hurricane 300 miles off-shore (Drye, 2012) and headed towards the most populated region of the U.S. At 8 pm, Sandy’s eye came ashore across New Jersey, destroying Atlantic City’s iconic boardwalk and recreation area.

The storm surge crested with record high elevations in New York City, flooding Manhattan. Tunnels and subways were inundated and the flood caused a large explosion at the Consolidated Edison (ConEd) power plant, shutting off power for millions of New Yorkers.

The United States Government was active throughout the lifecycle of the storm using public outreach to create awareness for the storm and its potential for catastrophic damage (Federal Emergency Management Agency [FEMA], 2013). The President was in close contact with Governors Chris Christie of New Jersey and Mario Cuomo of New York. New York City Mayor Michael Bloomberg ordered evacuations of all low-lying neighborhoods and closed the public schools (CNN, 2013).
On Sunday, October 28th, President Barack Obama declared states of emergency for Connecticut, District of Columbia, Maryland, Massachusetts, New Jersey, and New York (FEMA). The following day Sandy passed through New York, with its backside winds dealing another blow to the region. President Obama declared major disasters for Connecticut, New Jersey, and New York, making disaster assistance available to the hardest hit regions affected by the storm (FEMA). The storm then weakened in strength and dissipated over Pennsylvania by October 31st.
2.4.2 The 2012 Election

Coincidently, the devastating hurricane event occurred near the climax of a hotly contested political campaign season, capped with the 2012 Presidential Election. At the national level, Democratic incumbent Barack Obama faced Republican challenger Mitt Romney.

In my geographic area of interest; New York City, one U.S. Senator was up for reelection and 12 U.S. Congressional Representatives were on the ballot. Major local politicians that were not up for reelection this cycle included New Jersey Governor Chris Christie, New York Governor Mario Cuomo, United States Senator Chuck Shummer, and New York City Mayor Michael Bloomberg.

As is the case in most presidential election cycles, the top of the ticket tends to dominate public discourse and politicians who share the ballot with an incumbent president are often directly affected by the president’s popularity, often referred to as the coattail effect (Enten, 2011).

Elections have a strong affect on public discourse and opinions toward government officials are amplified during the election cycle. Monitoring social media during the election season will reveal negative opinions of these officials with much greater frequency, and will elicit this information out of a greater percentage of the population.
2.5 Data Analysis

2.5.1 Database Schema

The database schema for this analysis was determined through early testing of the application development cycle with the Streaming API. The Tweets returned by the API are formatted using nested key-value pairs that serve as properties along with their associated values. For instance, each tweet returned by Gnip contains properties, or metadata, about the user. The tweet also contains metadata about the status update such the time, location, device, and language (Figure 6).

Currently there is no "one size fits all" database that accomplishes all the goals of the analysis. By utilizing the strengths and ignoring the shortcomings of different systems, one can build a cohesive database architecture that can store and represent the large volume of Gnip tweets spatially. I was able to store more than 10 million records daily using MySQL\textsuperscript{14}, and filter the body of text of the Tweet records, to determine the value of keyword filters, without running into any scalability issues.

To inspect the records visually, I used the easy-to-read, tabular interface of Microsoft Access\textsuperscript{15}. Access adds further functionality to my workflow with its ability to be imported to a spatial database with greater ease than most other platforms.

\textsuperscript{14} MySQL is an open-source relational database management system. www.mysql.com

\textsuperscript{15} www.office.microsoft.com
2.5.2 Filter by keyword

Following the (Sakaki, 2010, Hu & Liu, 2004) method, I used a list of keywords, in this case, well-known government officials and agencies to train a Python script that filters unneeded information in each tweet. Other data included in the database tables were the geographic location, the time of the status update, the device that created the update, and some metadata about the user for potential future use (Figure 7).
The device that generated the tweet attribute was used to confirm that only mobile devices or third-party location-based-services (LBS), such as Foursquare, created the geolocated tweets. Sampling of the Twitter feed revealed that a large number of Foursquare-generated tweets pollute the stream with unnecessary noise. Most Foursquare tweets are automatic status updates that only reveal that “the user has checked-in to this geographic place”, usually a business establishment. These updates, which constitute a large volume of data, are removed by querying which device created the update.

Sentiment analysis was performed using a similar filtering method. For each tweet in the database, the program scans through the body of text. If that body of text contains a term that exists in the negative keyword lexicon it will add a count of one to the sentiment score. If there are multiple negative words in the text, each negative word will add to the total sentiment score of the tweet (Figure 8). A tweet with a score of “-1” or less is considered a negative tweet for the purpose of this analysis. Cumulative scores generated were created for possible future testing.
2.5.3 Frequency analysis

A temporal window surrounding the dates of high negative sentiment is used to compare normal social media behavior with the periods with high frequency negative sentiment. A raw frequency histogram is needed to determine which events attract the
most social media attention. Abnormally high frequency dates will help to define the
dates to use and give a sense of overall social media patterns.

A normalized frequency (i.e., percentage of tweets that are negative) is needed to
determine which dates exhibit the most negative sentiment. A second frequency
analysis is performed on the data that is based on the geography of the social media
updates. Important dates such as the election attract a very high frequency of data,
which can be misleading to the study, as I am more interested in locating pockets of
negativity.

2.5.4 Socioeconomic overlay

During periods of high social media activity that coincide with major political
and natural disaster events, it would be ideal to understand what social and economic
conditions exist at the geographic location of the areas of high negative sentiment. An
overlay of geographic information can create a relationship between the social media
activity and the socio-economic conditions.

The metadata contained in the tweet reveals many things about the user and the
tweet itself, however, it does not reveal user attributes such as race, income, or age.
United States Census data from a geographic information system (GIS) contains all of
these attributes along with the geographic boundaries where these attributes exist
(Figure 9, 10, 11).

An overlay of the negative social media status updates, which are stored as
geographic points in a database, on the socio-economic layer will reveal what groups of
people might feel resentment towards the government. This approach of generalizing social media statuses from point locations to a polygon (whose overall socio-economic pattern is generalized by a census boundary) could reveal whether or not negative sentiment towards government is driven by socio-economic geographic conditions.

To identify sentiment within the census boundaries, social media points that exist within the boundary will be aggregated to the tract using a *count points in polygon* algorithm. Special consideration of the data is needed to ensure that census tracts with very low frequencies do not skew the results. Many census tracts may contain only one tweet, thus giving it a 100 percent opinion score.

To overcome this potential bias, the frequencies of socio-economic metrics are determined using classes of attribute indicators. The class breaks in the map and the frequency distribution will be based on the standard deviation from the mean. This is the appropriate method of classification because all of the metrics (i.e. race, income, age) follow normal or bimodal distributions.

No one census tract, irrespective of the negative sentiment frequency has greater influence than another census tract with similar socioeconomic attribute values in my analysis. Rather, the demographic attribute values of the tracts will be the controlling attribute for my analysis. The counts of negativity within these bins will simply provide a summary statistic on the socio-economic indicator.

Another potential bias in this analysis is that people are submitting social media updates from their home census block. This analysis does not have the ability to track user’s locations with respect to their homes. An assumption or generalization must be
made that most people are using social media within close proximity to their homes or that they the geographic range that they are travelling to exhibit similar socio-economic conditions to their home blocks.

Clockwise from top-left:

Figure 9: Median Income
Figure 10: Median Age
Figure 11: Race
3. Results and Discussion

The analysis revealed unique patterns of negative sentiment based on demographic geography when used in conjunction with periods of time when the percent of negative sentiment status updates peaks. The three metrics I was most interested in discovering, with respect to negative expressions towards the government, were income, age, and race, discussed in the following section.

3.1 Spatiotemporal Patterns of Negative Opinions

Once the entire database was filtered down to a usable set of records that contained the government related keywords and a resultant sentiment score, frequency plots of the time period were created (Figure 12).

As expected, the volume of social media activity within the collected area deviates from its baseline average of 810 geo-enabled tweets per day at pivotal points in the timeline. A small jump in total tweets occurs on October 11th that can be attributed to the Vice-Presidential debate. Much larger spikes in both total Twitter activity and negative counts occurs on October 16th and 22nd. These days represent the final two Presidential debates.
On Election Day, Twitter announced that it set a site record with 31 million election related tweets (which contained certain key terms and hash tags), with a peak one-second rate of 15,107 tweets per second that nearly tripled the Twitter’s average of 5,700 tweets per second (Rewashed, 2012). These astonishing numbers are reflected in the geo-located New York City dataset, as was expected with such a high profile event taking place.

A natural ebb and flow of negative sentiment is exhibited, with an average of 189 negative tweets per day occurring throughout the time period. Approximately 22 percent of all government related tweets have a negative sentiment. However, unlike the raw frequency of negative sentiment, the normalized represents the actual trend of negative sentiment (Figure 13).
In the time frame between the last debate and the election, the Twitter stream maintains a higher than normal state of traffic with a gradual increasing trend occurring as the election draws near. However, the second Presidential debate and the dates between October 29th and October 31st draw the most negative (as a percentage) tweets. Those days represent the day Hurricane Sandy hits New York and the two days that follow. The frequency of Twitter activity can be represented spatially as well. The normal frequency (Figure 14) exhibited on the off-peak dates is fairly sparse. But during the hurricane (Figure 15) the volume of data increases significantly. However, on Election Day (Figure 16), when Twitter set all-time records for data-usage, geo-enabled tweets are dense enough to cover all of Manhattan and multiply throughout the study area. Very large numbers of people receive, distribute, and are active participants in news gathering and diffusing. When significant events unfold, social media users take to their service to broadcast information from their own perspective.

Figure 13: Frequency of percent negative per day. The Y-axis is the percentage of negative social media statuses.
Figure 14: Spatial pattern of normal Twitter activity. Red dots represent negative sentiment tweets. White Dots represent neutral or positive sentiment tweets.

Figure 15: Spatial pattern of Twitter Activity during Hurricane Sandy. Red dots represent negative sentiment tweets. White Dots represent neutral or positive sentiment tweets.

Figure 16: Spatial pattern Twitter activity on Election Day. Red dots represent negative sentiment tweets. White Dots represent neutral or positive sentiment tweets.
3.2. Median Household Income

The median household income for New York City, within the study area, has a unimodal distribution, skewed to the right, with an average household income of about $57,000 (Figure 18A: Frequency). For the study area, pockets of significant wealth exist mostly along the Hudson River, Manhattan in New York, and the waterfront in New Jersey (Figure 17: Map).

A parallel pattern emerges (Figure 18A: Social media analysis) from election sentiment with hurricane sentiment as wealth increases. This pattern suggests that wealthy people have a higher negative attitude towards the government during each type of event. Because more wealthy folks are located along the sea wall, perhaps their higher negative attitude is caused by greater damage to their properties and way of life caused by Hurricane Sandy.

Both events create relatively low negative attitude based on income when compared to the age and income metrics. It is interesting that people that live below the average income seem to express little outrage towards the government during the hurricane and the lowest negativity of any class in the election. Nearly 1,200 tweets originated from census tracts that have a median income of $38,000. These tracts tend to have a much less negative perception of their government than those having a household income greater than $38,000.
Figure 17: Median Household Income
3.3 Median Age

Median age follows a uniform distribution (Figure 18B: Frequency). Spatially it has a fairly random distribution with large pockets of youth in Brooklyn and pockets of older population around Central Park in Manhattan as well as northeast Queens (Figure 19: Map).

The resultant negativity pattern created by social media at first seemed to be a fairly uniform response to a uniform distribution (Figure 18B: Social Media Analysis). However, polarizing negative sentiment with respect to each event occurs with folks with a median age younger than thirty.

Perhaps these differing trends are due to the nature of age. Census blocks where the median age is younger than twenty usually contain schools, both primary and secondary. Someone who is young, who does not own property, or has no children, may be less likely to be upset by hurricane events unfolding around them. However, homes that have a median beyond 27 could feel more impacted by the hurricane.

The election actually increased the negativity in age groups younger than 27, while older populations where less stimulated by this process. Perhaps people experienced in the nature of American politics are more accustomed to the negative culture that surrounds politics, whereas young folks, who tend to be more optimistic about the human condition, may feel alienated by the election process.
Figure 19: Median Age
3.4 Race

The most interesting and revealing results emerge within the racial study of negative sentiment covering these events. Census data reveals significant spatial clustering, or segregation of race in New York City (Figure 20: Map), despite it having a bimodal and diverse overall population (Figure 18C: Frequency). Segregation of race reveals deep divisions in society. The same disturbing divisions of geography appear to apply towards divisions in sentiment (Figure 18C: Social media analysis).

President Obama was the first black president elected in America. While great strides have been made since the time of slavery and the Civil Rights Act, racism and discrimination still thrive in the U.S. Therefore, the president’s race may contribute much of the critique and support that is expressed towards him.

The census tracts that contain less than 65 percent white population have a uniform negative sentiment of about 22 percent. But where white population is larger than 65 percent, its negativity towards government officials becomes quite significant.

Spikes in government related keywords occurred during significant presidential election events. And of course the keyword “Obama” dominated all other government related keywords, so most social media chatter in the region, and across the United States dealt specifically with the presidency. The negative sentiment towards government, and more specifically towards President Obama, seems to be directly correlated with race.
A second disturbing trend emerges as it pertains to race. Census tracts that are less white have higher negative sentiment towards the government during the hurricane than white communities.

The United States has a poor record with race relations and government response towards non-white communities during natural disasters. The negativity in social media during Hurricane Sandy could be a reflection of communities that feel neglected by their government and whiter communities who tend to receive preferential treatment from the government.
4. Conclusion and Future Direction

4.1 Conclusion

A social media data mining application was created that absorbed all geo-enabled tweets that contained one or more reserved keywords of government officials and agencies. A negativity score based on a negative keyword lexicon was assigned to all of the tweets that contained a government keyword. The output of the application was a database that contained information about the user, the body of text in the status update, the sentiment score, and the geographic location of the message.

Using publicly available U.S. Census demographic data I was able to plot these messages to my study area of New York City. Using a point-in-polygon algorithm, a frequency of negative messages was assigned to each census tract.

Many limitations occur throughout this analysis. A number of inferences are made about the nature of individuals with respect to the census blocks that the tweets originated from. It is assumed that users will make social media updates in and around their home or in neighborhoods with similar socio-economic attributes. It is also assumed that access to the Twitter service was unbroken during the Hurricane. There was a significant power outage during the hurricane that could have had a major effect on the number of negative status updates.

With those assumptions accepted, the result of this analysis revealed that when social media data has geographic information embedded, one can utilize geographic information systems to overlay the information on demographic data to generalize
which socio-economic groups are expressing negative reactions or opinions during natural disaster and political events. Analysis of the information overlay revealed that unique patterns emerge within the socio-economic classes with respect to the attribute being examined.

Patterns of social media frequency peaked on dates significant to the 2012 election and Hurricane Sandy. Spatially and temporally, social media volume increases when newsworthy events occur. People have always had a need to be involved in the spread of information, regardless of its accuracy. And during this time period, in New York City, conversation about the presidential election trumped conversation on Hurricane Sandy.

When examining age as a metric, the pattern of negativity is mostly uniform across all age groups. However, groups with a median age less than thirty exhibited much more variation in opinion that depended on the nature of the event taking place.

Median household income as an attribute exhibited lower overall negativity across all classes. Interestingly though, people with more economic wealth have a much greater negative sentiment towards the government than those with little wealth.

Race as the defining metric provided the most interesting results of the analysis. While race relations in America seem to have improved a great deal in the last 150 years, the results of this study suggest that we have a long way to go before race is not an issue. White people in this study exhibited much more hostility towards the election and the presidential candidate who headlined the contest. And conversely, less white groups exhibited hostility towards the government during Hurricane Sandy.
4.2 Future Direction

The development of Web 2.0 technologies has created a new paradigm in geographic information science. We now have the ability to mine and analyze social media behavior and public opinion with greater ease. This will create opportunities for spatial and social scientists to examine the populous who are contributing terabytes of volunteered opinion data every day.

Many interesting questions about this research came about as a result of the spatial analysis. The factual evidence is that there is segregation of sentiment in the three metrics I examined. However, the discussion that I provide about the results is speculative, rooted in the history of American culture. Further examination of the results by social scientists could lead to more accurate and prophetic knowledge of the conditions.

Using a simple lexical approach proved to be a sufficient mechanism to analyze opinion data using Twitter. A social media site that limits its users to 140 characters or less provides simplicity to both the users of the technology and the methods to examine it. As new social media outlets emerge and others increase access to their APIs, how can more sophisticated natural language processing algorithms improve the results of opinion mining?

Furthermore, the study area of New York and other very large metropolitan areas have a great deal of geographic and demographic data available. How well would my approach work using less populated, less diverse, and less data accessible sites?
As big data grows, and more users contribute their thoughts and opinions for the world to digest, more and more institutions will eagerly and actively consume their data for public service, consumer research, and possibly for more nefarious reasons. Therefore, further research by academia is needed to test the capabilities of social media data-mining, create innovative workflows, and question the motivations of those who seek out this wealth of information.
Works Cited


De Longueville, B., Smith, R., & Luraschi, G. (2009). "OMG! From here, I can see the flames!": A case for mining location based social networks to acquire spatio-temporal data on forest fires. ACM LBSN (pp.73-80). Seattle: ACM.


Lei, T., & Liu, H. (2010). *Graph Mining Applications to Social Network Analysis*. Springer US.


paltry, pan, pandemonium, pander, pandering, panders, panic, panick, panicked, panic, panicky, paradoxical, paradoxically, paralyze, paralyzed, paranoia, paranoid, parasite, pariah, parody, partiality, partisan, partisans, passe, passive, passiveness, pathetic, pathetically, patronize, paucity, pauper, paupers, payback, peculiar, peculiarly, pedantic, peeled, peeve, peevish, peevishly, penalize, penalty, perfidious, perfidy, perfunctory, peril, perilous, perilously, perish, pernicious, perplex, perplexed, perplexing, perplexity, persecute, persecution, pertinacious, pertinaciously, pertinacity, perturb, perturbed, pervasive, perverse, perversely, perversion, perversity, pervert, perverted, perverts, pessimism, pessimistic, pessimistically, pest, pestilent, petrified, petrify, pettifog, petty, phobia, phobic, phony, picket, picketed, picketing, pickets, picky, pig, pigs, pillage, pillory, pimple, pinch, pique, pitiable, pitiful, pitifully, pitiless, pitilessly, pittance, pity, plagiarize, plague, plasticky, plaything, plea, pleas, plebeian, plight, plot, plotters, ploy, plunder, plunderer, pointless, poison, poisonous, poisonously, pokey, poky, polarisation, polemize, pollute, polluter, polluters, pollution, pompous, poor, poorer, poorest, poorly, posturing, port, poverty, powerless, prate, pratfall, prattle, precarious, precariously, precipitate, precipitous, predatory, predication, prejudge, prejudice, prejudices, prejudicial, premeditated, preoccupy, preposterous, preposterously, presumptuous, presumptuously, pretense, pretend, pretense, pretentious, pretentiously, prevaricate, pricey, pricier, prick, prickly, prickles, prideful, prik, primitive, prison, prisoner, problem, problematic, problems, procrastinate, procrastinates, procrastination, profane, profanity, prohibit, prohibitive, prohibitively, propaganda, propagandize, proprietary, prosecute, protest, protested, protesting, protests, protracted, provocation, provocative, provoke, pry, pugnacious, pugnaciously, pugnacity, punch, punchy, punish, punishable, punitive, punk, puny, puppet, puppets, puzzled, puzzlement, puzzling, quack, qualm, qualms, quandary, quarrel, quarrelless, quarrellously, quarrels, quarrelsome, quash, queer, questionable, quibble, quibbles, quitter, rabid, racism, racist, racists, racy, radical, radicalization, radically, radicals, rage, ragged, raging, rail, raked, rampage, rampant, ramshackle, rancor, randomly, rankle, rant, ranted, ranting, rantingly, rants, rape, raped, raping, rascal, rascals, rash, rattlin', rattling, rattles, ravage, raving, reactionary, rebellious, rebuff, rebuke, recalcitrant, recant, recession, recessionary, reactionary, recessionary, reckless, recklessly, recklessness, recoil, recourses, redundancy, redundant, refusal, refuse, refused, refuses, refusing, refutation, refute, refuted, refutes, refuting, regress, regression, regressive, regret, regretted, regretful, regretfully, regrets, regrettable, regrettably, regretted, reject, rejected, rejecting, rejection, rejects, relapse, relentless, relentlessly, relentlessness, reluctance, reluctant, reluctantly,