COMPARATIVE GEOSPATIAL ANALYSIS OF TWITTER SENTIMENT DATA

DURING THE 2008 AND 2012 U.S. PRESIDENTIAL ELECTIONS

by

JOSEF GORDON

A THESIS

Presented to the Department of Geography
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Master of Science

September 2013

THESIS APPROVAL PAGE

Student: Josef Gordon

Title: Comparative Geospatial Analysis of Twitter Sentiment Data during the 2008 and 2012 U.S. Presidential Elections

This thesis has been accepted and approved in partial fulfillment of the requirements for the Master of Science degree in the Department of Geography by:

Amy Lobben	Chairperson
Xiaobo Su	Member

and

Kimberly Andrews Espy	Vice President for Research and Innovation; Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded September 2013

THESIS ABSTRACT

Josef Gordon

Master of Science

Department of Geography

September 2013

Title: Comparative Geospatial Analysis of Twitter Sentiment Data during the 2008 and 2012 U.S. Presidential Elections

The goal of this thesis is to assess and characterize the representativeness of sampled data that is voluntarily submitted through social media. The case study vehicle used is Twitter data associated with the 2012 Presidential election, which were in turn compared to similarly collected 2008 Presidential election Twitter data in order to ascertain the representative statewide changes in the pro-Democrat bias of sentiment-derived Twitter data mentioning either of the Republican or Democrat Presidential candidates.

The results of the comparative analysis show that the mean absolute error lessened by nearly half – from 13.1% in 2008 to 7.23% in 2012 – which would initially suggest a less biased sample. However, the increase in the strength of the positive correlation between tweets per county and population density actually suggests a much more geographically biased sample.

CURRICULUM VITAE

NAME OF AUTHOR: Josef Gordon

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene

DEGREES AWARDED:

Master of Science, Geography, 2013, University of Oregon
Bachelor of Science, Geography, 2010, University of Oregon
Bachelor of Science, Environmental Studies, 2004, University of Oregon

AREAS OF SPECIAL INTEREST:

Geographic Information Systems
Social Media and Volunteered Geographic Information
Big Data

PROFESSIONAL EXPERIENCE:

Graduate Research Assistant, Spatial and Map Cognition Research Lab, University of Oregon, September 2010 – September 2013

GIS Intern / Private GIS Contractor, Lane Council of Governments, January 2011 – June 2012

GIS Assistant, InfoGraphics Lab, University of Oregon, January 2010 – September 2010

GIS Analyst, Ecosystem Workforce Program, University of Oregon, September 2009 – September 2010

ACKNOWLEDGMENTS

I would like to thank Amy Lobben and Xiaobo Su for their guidance and advice during the course of this research project, as well as for their editorial input during the writing of this thesis. Additionally, I would like to thank my parents, Steven and Susan Gordon, for their lifelong support of my desire to explore the world and learn for myself.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

The overarching goal of this thesis is to assess and characterize the representativeness of sampled data that is voluntarily submitted through social media. The case study vehicle used is Twitter data associated with the 2012 Presidential election. More specifically, those data were compared to similarly collected 2008 Presidential election Twitter data in order to ascertain the representative statewide changes from 2008 to 2012 in the pro-Democrat bias of sentiment-derived Twitter data mentioning either of the Republican or Democrat Presidential candidates. Additionally, representativeness is inferred from a change between 2008 and 2012 in the correlative strength of overall Twitter use and population density per county.

2012 data were collected for the comparative analysis during the nine weeks leading up to Election Day, November 6th, 2012, and the 2008 data are referenced from the results of a separate study conducted by Gayo-Avello (2011). The states of interest for the 2012 comparative analysis are California, Florida, Indiana, Missouri, North Carolina, Ohio, and Texas, as originally chosen by Gayo-Avello for his 2008 study. California and Texas were chosen as being representative of Democrat- and Republican-leaning states, respectively, and the rest were chosen due to their characterization as swing states. In 2008, Gayo-Avello’s findings suggested a strong pro-Democrat sample bias in Twitter sentiment relative to candidate mentions, which he suggests is due in large part to the self-selection of urban and young voters favoring Obama.

The measure of sample bias for Gayo-Avello's study (and repeated again for the 2012 comparative analysis) is derived from the attribution of "votes" to Twitter users, which are then compared with election results. Twitter "votes" are attributed by the sentiment of individual tweets mentioning Obama or Romney (or McCain in 2008), and then aggregated to individual users. Sentiment is measured by counting the number of positive or negative words appearing in conjunction with a candidate's name. For both this thesis and the comparative analysis by Gayo-Avello (2011), the positive and negative words used in the sentiment analysis are sourced from a previously published subjectivity lexicon (Wilson 2005).

Based on the 2008 Presidential election Twitter data analysis, Gayo-Avello (2011) concludes that Twitter data results in a biased sampling that cannot reliably predict Presidential election outcomes. These results provide a foundation for comparative analysis. By conducting a similar study, this thesis can document the extent to which the pro-Democrat sample bias lessened or increased relative to the 2008 study, as evidenced by a change in the mean absolute error (MAE) between statewide actual election results and Twitter sentiment-derived results, as well as a change in the correlative strength of overall Twitter use and population density per county. This thesis is driven by two central research questions.

1. What is the extent to which geographic information volunteered through social media has diffused from 2008 through 2012?

2. What are the relative urban and rural patterns of geographic information volunteered through social media in 2012?

Although the predictivity of social media relative to mass events is an important component of the comparative analysis, the central conceptual issue at hand is the extent of participation, or rather, the diffusion of use of the technological means of data production (in this case, Twitter) from 2008 to 2012, whereby the Presidential elections simply represent convenient and much-discussed occurrences, rather than events of primary interest.

The results of the comparative analysis show that the MAE lessened by nearly half – from 13.1% in 2008 to 7.23% in 2012 – which would initially suggest a less biased sample. However, the increase in the strength of the positive correlation between tweets per county and population density actually suggests a much more geographically biased sample. In all likelihood, the lessening of the MAE is due to increased electoral contestation, although the effect of the diffusion of use into pro-Republican demographic sectors cannot be ruled out.

Increased geographic bias, trending toward a concentration of use in high-density urban areas is also documented and not surprising in light of the sheer magnitude of overall increased use of Twitter – from 20,000 users collected over 6 months in 2008 to 490,000 users collected over 2 months in 2012 – as well as in light of the relative newness of the technology in 2008.

CHAPTER II

BACKGROUND

GIS has benefited in the last few decades from the advent of internet and mobile technologies, which have in turn helped facilitate the growth of participatory GIS in the 1990s, volunteered geographic information (VGI) alongside Web 2.0 at the turn of the century, and currently the popular field of Big Data research. This very general developmental trend has been paralleled by a critical interest in the democratization of GIS and the process of data production.

Three fundamental questions have arisen throughout the development of geospatial and data production technologies, namely (1) who takes part, (2) how does this participation occur, and (3) what are the motivating factors for the participants in terms of data production?

Participatory GIS and Citizen Science

Participatory GIS has its roots in citizen science, and throughout much of its development has been acted out in the sphere of academic and public agency research, where the use of GIS has been facilitated or fostered by traditionally expert knowledge-holders. The focused and well-planned recruitment of participants on the part of researchers has often been central to participatory GIS projects, and has been motivated in part by the desire to both democratize the process of data production and promote civic engagement. Through the democratization of the traditionally top-down expert-oriented process of data production, the fulfillment of public interest via participatory GIS may be strengthened by the inclusion of local citizen knowledge.

In the case of Weiner's study (1995) in South Africa, 100 village elders were recruited to provide localized geographic and cultural information for a GIS-based land reform project, which was largely an attempt to allow alternative voices to contribute to the traditionally top-down development planning process. Guided by the primary stakeholders, Weiner was able to identify important data, goals, and decisions that may not have been illuminated without community participation. Similarly, Jordan's study (2002) recruited locals in Nepal to help inform assessments of forestry-related issues regarding conflict resolution, access to resources, and biodiversity preservation.

In both instances, although the empowerment of local community members alongside the democratization of development processes via data production were explicit goals of the participatory GIS projects, the recruitment of individuals was also predicated upon the idea that the locals simply had a better idea of what was actually happening on the ground, and thus the data would be of higher quality and subsequent policy decisions would be more informed.

Historically, as geospatial tools developed, the methods of recruitment and means of data production (e.g., online surveys, smartphone apps) lent themselves toward greater flexibility of the application of participatory GIS projects, ranging from public online surveys of attendants of U.S. National Parks (Brown 2011), to citizen-based noise studies of cityscapes with mobile phones (Kanjo 2009 and Maisonneuve 2009) to smartphone transportation analyses based on bicycle commuter routes (Kessler 2011).

Web 2.0 and VGI

The same development of geospatial and internet technologies that diversified the spectrum of research in participatory GIS also brought about the proliferation of newer, less solicited types of data production, which have generally been referred to as volunteered geographic information (VGI). The lines between participatory GIS and VGI blend, and sometimes distinctions of levels and types of participation/solicitation become meaningless, but essentially, the general mode of data production in VGI centers around more open-ended and unsolicited avenues of voluntary participation, as compared with the more focused or guided studies that have comprised much of participatory GIS.

The advent of Web 2.0 has created an active rather than a passive Internet society, where users produce rather than simply consume data. The bulk of citizen-based data production has in recent years shifted away from targeted participatory studies, which is not to say that participatory GIS has waned, but rather that the means (e.g., smartphone) and ends (e.g., FourSquare check-ins) of unsolicited VGI production have grown in popularity in day-to-day lives. Parallel to the recent rise of social media, the magnitude and variety of unsolicited types of data production have rapidly increased (e.g., Facebook, Yelp, TripAdvisor, Flickr, and Twitter). Over the past decade, it has become normalized for individuals to catalog and publish daily activities, emotions, and opinions, resulting in an immeasurable amount of publicly available, freely volunteered (and geographic) Big Data.

Big Data

Understandably, the increased magnitude of citizen-based data production has garnered interest in a wide range of Big Data and social media research, including studies on demographics (Mislove 2011), online information diffusion (Van Liere 2010), internet geographies (Takhteyev 2012), sentiment analysis (Thelwall 2011), ethics and privacy (Henderson 2012), emergency response (Starbird 2011), health and epidemiology (Signorini 2011), citizen journalism (Murthy 2011), media discussions (Bruns 2012), urban and transportation analysis (Ueno 2012), financial and market trends (Bollen 2011), social activism (Vallina-Rodriguez 2012), and mass events ranging from movie premieres (Asur 2010) to elections (Gayo-Avello 2011). Nonetheless, despite the topical and technological novelty, the excitement around Big Data is mitigated by the intellectual need to address the questions of who participates, how, and why they participate. Considerations of the nature of participation implicitly reference the idea that large internet-based digital repositories simply reflect population distributions, and thus show us pictures of only urban life while underrepresenting rural or unconnected areas. Due to the necessary technological facilitation of participation, the issue of access has been inextricably intertwined with citizen-based data production since its beginnings in participatory GIS and VGI, where despite the fact that access has generally "expanded for the most advantaged users, at the bottom of the digital divide relatively little has changed" (Elwood 2006: 694).

The advent of cheaper and more advanced technologies may serve to eventually lessen the digital divide by providing a wider range of users access to the means of data production, as has been argued with the potential of cheap smartphones to connect the developing world (Boyera 2007, Jensen 2012). In opposition, there are claims that smartphone ubiquity is a myth, as in the case of the United States, where nearly 16% are left out of the data production process due to poverty (Pernisco 2011). Furthermore, as the growing interests of government, academia and industry focus upon the potential of citizen-based data to provide meaningful insights into societal trends or environmental conditions, Big Data research should be approached with the critical understanding that new technologies may inadvertently widen the "'digital divide' between those adopting/having the technology and those avoiding/lacking it" (Newman 2012: 301). Perhaps the most meaningful part of Newman's statement is that it's not always about access, but also about choice, which is central to the issue of the types of participation that occur in the very general realm of Big Data. In other words, not every person with a smartphone chooses to put their thoughts, opinions, or ideas up for public consumption.

Likewise, with regards to any responsible analysis of Twitter data, access to the means of data production is certainly an issue, as evidenced by Twitter's reflectance of overall patterns of internet use globally (i.e., the disproportionate presence of U.S. and Europe) and nationally (i.e., the disproportionate presence of highly populated areas). However, access alone is not the only reason people use Twitter; participation is characterized by age and culture demographics. While some demographics may currently be the primary contributors to publically generated data via social media, the demographic composition will likely change over time, as has been evidenced by the continued dissemination of Facebook out of the initial college-aged users. For the time being, however, the majority of Twitter users are under 30, and they continue to be the fastest growing demographic (Smith 2012).

Despite the uneven representation of sample populations, the attractiveness of citizen-based Big Data is understandable, due in part to its unprecedented size, as well as to its potential reactivity to real-world events. One of the obvious low-hanging fruits in Big Data research, and in Twitter research particularly, is political analysis. The predictive capacity of Twitter data has been a popular topic in recent years, both internationally and in the United States, and so far there has been mixed reviews regarding accurate Twitter-based election predictions.

Twitter Elections

Initial studies often presented optimistic results regarding the predictive capacity of Twitter data relative to election results. Some researchers had found that the volume of candidate or party mentions alone reflected election results. In the case of Spanish elections, Borondo (2012) found that votes and tweets correlated closely and predicted a Zapatero victory in 2004 and a Partido Popular victory in 2011. In the 2010 Brazilian Presidential election, Dilma was the predicted winner, by virtue of single tweet mentions, over Serra and Marina respectively, with a mean absolute error of 4.07% (Trumper 2011). Perhaps most famous, the 2009 German federal elections were found to have incredibly high correlations between party tweet mentions and election results, with a mean absolute error of 1.65% among six parties (Tumasjan 2010 and 2011).

Other studies have found that although volume of tweets alone are predictive, these measures are supplemented by the sentiment of tweets affiliated with mentions

(Bermingham 2011, Cummings 2010). Additionally, another vein of election-oriented Twitter research has focused solely upon the efficacy of sentiment analysis (Ceron 2012, Lampos 2012, and Sang 2012).

In an attempt to draw out some of the complexities of Twitter data, further research has shed light on the effect that temporal and user sampling has on the predictivity of tweets. One study presents varying degrees of predictivity according to aggregation of types of users (i.e., high or low frequency of use), as differentiated by magnitude of tweets (Chen 2012). Another explores temporally segmented samples prior to election (Bermingham 2011).

Despite the perceived general optimism, contrarians are skeptical about the veracity of studies that promote the predictive capacity of Twitter data. Some of this skepticism relates to the issue of transparency, simplicity, and replicability of methods, especially of those studies whose findings are based upon word counts (Gayo-Avello 2012 and Jungherr 2011).

Combined with skepticism regarding methodology, other studies have found weak correlations between election-oriented Twitter data and real-world election results, and thus serve as empirical refutations of positive results. O’Connor’s (2008) study found that, although sentiment correlated with Obama’s job approval ratings, it did not show meaningful positive correlations with electoral predictions, as derived from the polarity lexicon OpinionFinder (O’Connor 2010). Similar results were found in Gayo-Avello’s 2008 study of the U.S. Presidential election, and were attributed in part to the issue of the ineffectiveness of simplistic sentiment analyses, in which tweets were mislabeled as positive or negative due to subtleties or linguistic complexities (Gayo-Avello 2011).

Another 2008 study found that sentiment analysis of time-series data too volatile to predict future opinions in the U.S. Presidential election (Bravo-Marquez 2012). In the 2010 U.S. Senate special election in Massachusetts, Chung (2011) found that neither volume of tweet mentions nor sentiment analysis accurately predicted the winner.

Additional skepticism surrounds the inherent issue of uncorrected sample bias (Gayo-Avello 2012, Metaxas 2011 and 2012). Gayo-Avello flatly states that "until Social Media becomes regularly used by the vast majority of people, its users cannot be considered a representative sample and forecasts from such data will be of questionable value at best" (2011: 14). Some studies have recently begun to model results using age-based corrections, though some assumptions are required due to uncertainty regarding the age of the subset of Twitter users who take part in election-oriented discussions (Choy 2011, Choy 2012).

Thesis Data Baseline

The findings and conclusions from Gayo-Avello's 2008 research (2011) serve as the baseline for the 2012 comparative analysis. At the time of publication, this study was one of the more critical refutations of the optimism originally associated with Twitter data's predictive capacity. He noted several fundamental flaws in the previous research. One of his criticisms was that simple word counts of candidate mentions in the tweeted body of text alone did not provide evidence of the actual meaning of the tweet. In turn, deriving meaning or opinion from the tweet by way of simple sentiment analysis – such as counting words associated with other words (i.e., positive or negative words associated with Obama or McCain) – was also inadequate for Twitter data analysis due to the shortness of the tweets and the inability to parse linguistic subtleties.

After testing four types of sentiment analysis, and comparing the results of these against a subset of tweets for which he knew the user’s true voting intent – as published on TwitVote – Gayo-Avello found that all of the analyses performed poorly. Nonetheless and without an alternative, Gayo-Avello chose to proceed with the least poor form of sentiment analysis, the polarity lexicon associated with OpinionFinder (Wilson 2005), with the knowledge in hand that tweets themselves are often too short and, for many infrequent users, too few, to provide meaningful aggregate user sentiment. Again, this is no slight against the polarity lexicon itself, but rather an issue with the structure and content of the tweets.

Overall, the study’s results found that Twitter sentiment largely over-predicted Obama wins in all seven states of interest. He suggests the young age and urban geographic bias skewed the apparently pro-Democrat sample.

CHAPTER III

METHODOLOGY

In order to make a comparative analysis between the available 2008 data from Gayo-Avello (2011) and the 2012 data gathered in this current study, similar methods were applied. Though, these methods were extended for this study to investigate spatial patterns associated with the 2012 Presidential election tweets. The following section documents methods associated with the 2012 data collection and analysis.

Twitter Streaming API, Amazon Web Services, and Hive

Twitter data were collected for nine weeks prior to the 2012 U.S. Presidential Election – from 12:00 a.m. on Tuesday, September 4th until 11:59 p.m. on Tuesday, November 6th. The method of collection entailed the use of Twitter's Streaming API, which was accessed using a cURL command with Mac OS X Terminal. Irrespective of location, Twitter data was collected by tracking the keywords "Obama" or "Romney".

The nine weeks of tweets mentioning "Obama" or "Romney" were uploaded to Amazon Web Services (AWS) S3 storage servers using the Import/Export option. The Import/Export option entailed delivering a hard drive to AWS, whereupon the data were directly uploaded to pre-existing user-specified data buckets. Although more expensive than using an internet connection, the Import/Export option was quicker.

The subsequent computation was run with an interactive Hive session, which parsed the Twitter data located on S3 servers. The interactive Hive session was set up using Elastic Compute Cloud (EC2) and Elastic MapReduce (EMR) tools.

In order to parse the Twitter data, one tweet at a time – with each tweet stored as a single line of text – a JSON Serde was needed. A JSON Serde essentially functions as a means of identifying and then accessing different types of data within nested JSON files, where some of the data is embedded within parent data formats (e.g., array, map, struct). These nested data types populated the fields that would comprise the tabular/spatial GIS database in the subsequent steps. The particular Serde that worked best for the nested JSON files was downloaded from GitHub (github.com/mattbornski/Hive-Demo/tree/master/exercises/lib). The two previous versions of this Serde (e.g., 0.1 and 0.2) did not entirely work with the complicated nested structure of Twitter data, and would sometimes return NULL values even when the data entries were known to exist. The third version of the Serde (e.g., 0.3) was able to extract the necessary fields.

The extracted fields included the tweet text, unique tweet ID (i.e., numeric and string format), longitude, latitude, timestamp, default profile place settings (i.e., city, state, and/or country), user ID (i.e., numeric and string format), and self-identified user location, which was typed in by the user and varied in levels of specificity. In order to facilitate the comma-separated importation of text data into tabular format, commas were stripped from three fields – including the tweet, the place full name (i.e., city, state, and/or country), and user location.

Due to the compartmentalized storage of files on S3 servers, single chunks of data (e.g. original Twitter JSON files) are not allowed to exceed a specified file size. This compartmentalization of data was mirrored in the final product of the Elastic MapReduce job, where each chunk of JSON data resulted in an individual comma-separated text file, resulting in hundreds of small output files that were downloaded and joined for the subsequent sentiment and geospatial analysis.

Parsing Locations from Tweets

Location data were extracted from four fields in the raw JSON files. Two of the fields were geotagged and contained longitude and latitude, respectively, while the other two fields contained semantic user locations, which included reference to individual cities and states. Most (approximately 90%) of the geographically-referenced tweets were identified based on the user location field, which was typed into the profile by the user.

The geotagged tweets (i.e., containing lat/long information) were identified and extracted with Hive and the JSON Serde by detecting the presence of longitude and latitude coordinates. These tweets were then brought into ArcGIS by importing a CSV file into ArcMap as an XY Event Layer, a process which references the appropriate coordinate fields. These tweets were then cross-referenced against U.S. Census data in order to attribute them with county names.

Other locations were extracted by identifying tweets that contained a location reference – either in the user location field or the place full name field – to the seven states of interest. Subsequent analysis at the county level was done by matching mentions of census places (e.g., cities, townships, villages) within either the user location or place full name field.

The process of matching mentions of census places within the user location or place full name fields of the tweet involved developing a Python script that would match place names within substrings of tweet location descriptions. The need to identify substrings was due in part to the messy nature of the user location field, which could hypothetically include any variation of individually customizable location references (i.e., *Upper East Side Indianapolis*, or *+254 Fort Wayne In*). The script required two CSV inputs – a file containing the tweets themselves, and a file containing a list of Census places.

In order for the location parsing process to work, the list of Census place names needed to have their categorical descriptors stripped from within the name field (e.g., city, village, and township). The Python script matched the Census place name with any mention of it in the location description field of the tweet, and then extracted the matching place name from within the tweet's location fields and wrote the place name to a new field. The final result of the location parsing process was an individual CSV file for each state, which was then joined back into a GIS database containing the Census place name field, and analyzed in ArcMap as points (i.e., centroids of Census places).

Sentiment Analysis

The polarity lexicon published by Wilson (2005) is comprised of positive and negative words, and the sentiment analysis for both 2008 and 2012 simply measured the ratio of positive and negative words in a single tweet that appeared in conjunction with the mention of a Republican or Democrat candidate's name. A vote was determined for each tweet according to its sentiment, or ratio of positive and negative words, and these individual "tweet votes" were aggregated to the user IDs in order to determine a user's overall sentiment toward the candidates, thus inferring a user's individual vote. This method mitigated the effect of varying degrees of user participation, where 1% of users contributed approximately 25% of tweets.

The sentiment analysis involved using a Python script similar to the one used in the aforementioned location analysis process. The script required three CSV inputs – one containing the Twitter data and one list each of positive and negative words from the polarity lexicon. The Python script detected, extracted, and counted any matching words from the polarity lexicon corpus with any matching words in the body of the tweets themselves.

Comparison with Election Data

Once all of the tweets were georeferenced, the analysis of correlations between Twitter vote sentiment, Census, and election data was done in ArcMap and Excel. Twitter points were aggregated to Census county polygons. 2012 Presidential election results were sourced from CSV files downloaded from the individual state-level Huffington Post election websites (http://elections.huffingtonpost.com/2012/results).

The CSV files containing election results were brought into ArcMap and joined with the feature layer that contained both the Twitter results and Census data. From there, statewide and county level tabular correlations were calculated for total amounts of Twitter use, Twitter sentiment results, Census population statistics, and actual 2012 Presidential election results.

CHAPTER IV

RESULTS

Following a report of the overall sentiment analysis, this section organizes the results by each of the two research questions.

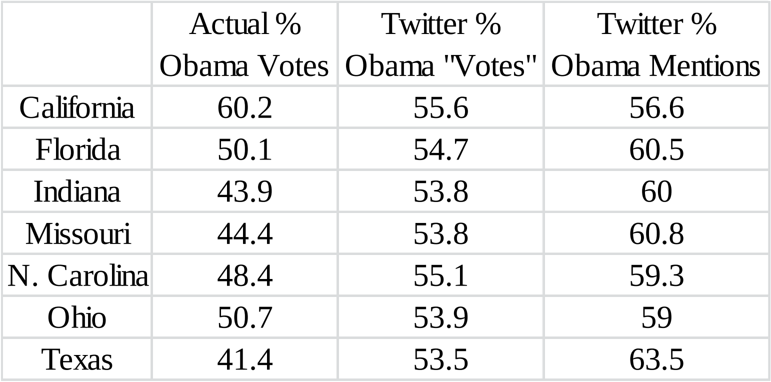
Sentiment Analysis

Approximately 1/3 of the 3.6 million tweets that contained both location information and candidate mentions were discarded due a neutral categorization by the sentiment analysis. Neutral categorization is not caused solely by a lack of sentiment in the tweets, but also because of the limitations of the polarity lexicon, which, for instance, does not contain the word *voted*, and thus the term – *I voted Romney because I'm Republican* – is categorized as neutral, when it is an obvious endorsement. Likewise, the phrase - *once again the media treated Obama with kid gloves during WH news conference* – despite being implicitly critical of Obama, or at least critical of an uncritical media, is considered neutral due to the lack of any explicitly negative words. Also, tweets that contain the same amount of positive and negative words are understandably considered neutral, as in the case of the phrase - *best thing about driving from Dallas is reading the ridiculous anti-Obama billboards* – where *best* and *ridiculous* cancel each other out.

Although revelations surrounding the limitations of simple sentiment analysis are far from novel, full disclosure is warranted nonetheless – which Gayo-Avello (2011) points out, as well – that without sufficient regard to linguistic complexities and implied meanings, the method of counting positive and negative words is a significant source of undetermined error and omission.

By measuring candidate mentions alone, the topical bias is more pronounced than sentiment-derived bias, as shown in Table 1. This difference is likely due in part to the twofold effect of incumbency, where Obama is referenced in non-election tweets, simply as a side-effect of being the acting President, and also because the overall rhetoric of the election is often situated around the performance of the defending incumbent.

Table 1: Possibly an effect of incumbency, the statewide percentages of Twitter mentions of Obama outweigh the percentages of sentiment-derived Twitter “votes” for Obama.



	Actual % Obama Votes	Twitter % Obama "Votes"	Twitter % Obama Mentions
California	60.2	55.6	56.6
Florida	50.1	54.7	60.5
Indiana	43.9	53.8	60
Missouri	44.4	53.8	60.8
N. Carolina	48.4	55.1	59.3
Ohio	50.7	53.9	59
Texas	41.4	53.5	63.5

Research Question One: What Is the Extent to Which Geographic Information Volunteered through Social Media Has Diffused from 2008 through 2012?

The predictive capacity of the 2012 Twitter-derived sentiment data showed a fair improvement over the 2008 results of Gayo-Avello’s study, with a Mean Absolute Error (MAE) of 7.23% in 2012, as compared with 13.1% in 2008. Table 2 compares the results by state and shows a lessening in 2012 of Twitter error in every state except California.

One possible inference from the lessened MAE is that the pro-Democrat bias evidenced in 2008 lessened in 2012. The effect of a lessened bias, however, cannot be entirely differentiated from the conflating effect of Twitter users simply changing opinions, or in other words, the difference in MAE may simply suggest a more contested election.

Table 2: Comparison of 2008 and 2012 U.S. Presidential Elections using sentiment derived from tweets which mentioned either Obama or Romney. The pro-Democrat bias from 2008 has lessened according to the MAE.

State	2008 % Actual Obama Votes	2008 % Twitter Votes	2008 Twitter Error	2012 % Actual Obama Votes	2012 % Twitter Votes	2012 Twitter Error
California	62.28	62.70	0.42	60.24	55.62	-4.62
Florida	51.42	66.20	14.78	50.10	54.69	4.59
Indiana	50.50	64.70	14.20	43.93	53.84	9.91
Missouri	50.07	68.10	18.03	44.38	53.76	9.38
N. Carolina	50.16	66.60	16.44	48.35	55.07	6.72
Ohio	52.31	59.80	7.49	50.67	53.94	3.27
Texas	44.06	64.40	20.34	41.38	53.53	12.15
		MAE	13.10		MAE	7.23

The influence of greater electoral contestation, as opposed to the possibility of a lessening of pro-Democrat bias, is supported by the lessening of percentages of actual Obama votes for each state. Additionally, as shown in Table 3, the general lessening of the strength of the correlation between the Democrat vote percentages and population density by county suggests a more contested election in the traditionally urban strongholds that supported Obama in 2008. On the other hand, as the blatant exception to the trend, the increase of the positive correlation in Missouri between Democrat vote percentages and population density conversely suggests an increased geographic polarization of political opinion within that state.

Table 3: A general lessening between 2008 and 2012 of the positive correlation (Pearson's *r*) for Democratic vote percentage and population density.

	2008 Democrat Vote % vs Population Density	2012 Democrat Vote % vs Population Density
California	0.4069	0.407
Florida	0.474	-0.4858
Indiana	0.5452	0.4894
Missouri	0.5239	0.7104
N. Carolina	0.3968	0.2288
Ohio	0.5676	0.5245
Texas	0.4789	0.2755

In terms of pro-Democrat Twitter bias, the most telling correlation is between Twitter use and population density by county, as shown in Table 4. With the exception of California, the positive correlation between Twitter use and population density increased substantially in the 2012 election, which shows that the issue of geographic sample bias is currently more pronounced than it was in 2008. Perhaps unsurprisingly, Twitter use is growing along existing patterns of population distribution.

Table 4: A general increase between 2008 and 2012 of the positive correlation (Pearson's *r*) for number of Twitter users and population density.

	2008 Twitter Users vs Population Density	2012 Twitter Users vs Population Density
California	0.9452	0.6435
Florida	0.1768	0.6465
Indiana	0.2956	0.9239
Missouri	-0.0079	0.7998
N. Carolina	0.5425	0.8366
Ohio	0.6343	0.9099
Texas	-0.0535	0.8699

Research Question Two: What Are the Relative Urban and Rural Patterns of Geographic Information Volunteered through Social Media in 2012?

Based on the available data from the Gayo-Avello (2011) results from the 2008 election, only gross geographic patterns can be compared. But, the data gathered for the current study, provide a more detailed look at current urban and rural trends in the use of social media to share volunteered geographic information.

The geographic bias toward areas of increased urban density is highlighted when we look at only the counties that have statistically representative Twitter samples, as seen in Table 5. The vast majority of tweets come from a subsection of counties that contain high popoulation density and overall high proportion of state populations. For example, in Texas 97.2% of the overall Twitter sample comes from approximately 1/3 of the state's counties, which in turn contain 88.1% of the state's total population. In another example, 94.7% of the overall Twitter sample in North Carolina comes from 2/5 of the state's counties, which in turn contain 77.6% of the state's total population.

The fairly low threshhold of counties with statistically representative Twitter samples, as compared with the high percent of population living in those counties, points out the unequal spatial distribution of Twitter, and the general undersampling of less populated areas.

Essentially, except for California and Florida, the other five states in question show an oversampling of heavily populated counties, as evidenced by the difference between the percentage of Twitter users from these counties and the percentage of the state's total population that these counties comprise. The fact that Twitter use mirrors population distribution, as shown in Table 4, is not surprising, but the oversampling of heavily populated areas is an important consideration in election analysis, where distinct correlations also exist between voting trends and population distributions.

Table 5: Where the percentage of tweets outweighs the percentage of population (from counties with statistically representative Twitter samples), high population counties are being disproportionately weighted.

	% of Twitter Users from Counties with Statistically Representative Twitter Samples	% of Counties with Statistically Representative Twitter Samples	% of Population from Counties with Statistically Representative Twitter Samples
California	99.5	66.7	98
Florida	99.1	53.7	94.5
Indiana	91.3	28.7	69.7
Missouri	87.8	13	60.1
N. Carolina	94.7	40	77.6
Ohio	96.8	53.4	86.3
Texas	96.8	21.7	84.9
Total	97.2	33.2	88.1

Another way to understand the effect that oversampling heavily populated areas has on potential election-oriented Twitter sentiment, as shown in Table 6, is to show that the counties with statistically representative Twitter samples also tend to have higher Democrat vote percentages in the 2012 election.

In other words, most of the Twitter data are being sourced from counties with increased average percentages of Obama voters. Although the difference in statewide percentages in Table 6 are not enormous, the increased average Democrat voting results in heavily populated counties, in combination with the oversampling of Twitter data from those counties, serves as a possible means of inflation of pro-Democrat Twitter sentiment.

Table 6: The counties with statistically representative Twitter samples also have higher Democrat vote percentages.

	% Actual Obama Vote for All Counties (Mean)	% Actual Obama Vote for Counties with Statistically Representative Twitter Samples (Mean)
California	52.6	56.1
Florida	39.9	45
Indiana	38	43.2
Missouri	33.7	42.7
N. Carolina	44.5	44.7
Ohio	43.1	46.7
Texas	28.3	34.4
Total	36.8	44.3

Twitter’s pro-Democrat geographic bias toward heavily populated counties is readily apparent when displayed cartographically, as seen in Figure 1. High concentrations of tweeters visibly stand out in the heavily populated counties that voted for Obama in the 2012 election. The pro-Democrat geographic bias is especially visible in California and Florida, where all of the counties with high total populations (shown in black) are also counties that contained high amounts of Twitter users, which in turn voted for Obama in the 2012 Presidential election (shown in dark blue), with the exception of Orange County, California (shown in dark red). Again, in Indiana, the same pro-Democrat geographic bias is apparent, where the only county with high population levels, Marion County (shown in gray), is also simultaneously a place with high amounts of Twitter users and a pro-Democrat election result (shown in dark blue).

The pro-Democrat geographic bias is also evident in North Carolina and Ohio, as seen in Figure 2, where all of the counties that have high populations (shown in gray or black) also have high amounts of Twitter users and pro-Democrat election results (shown

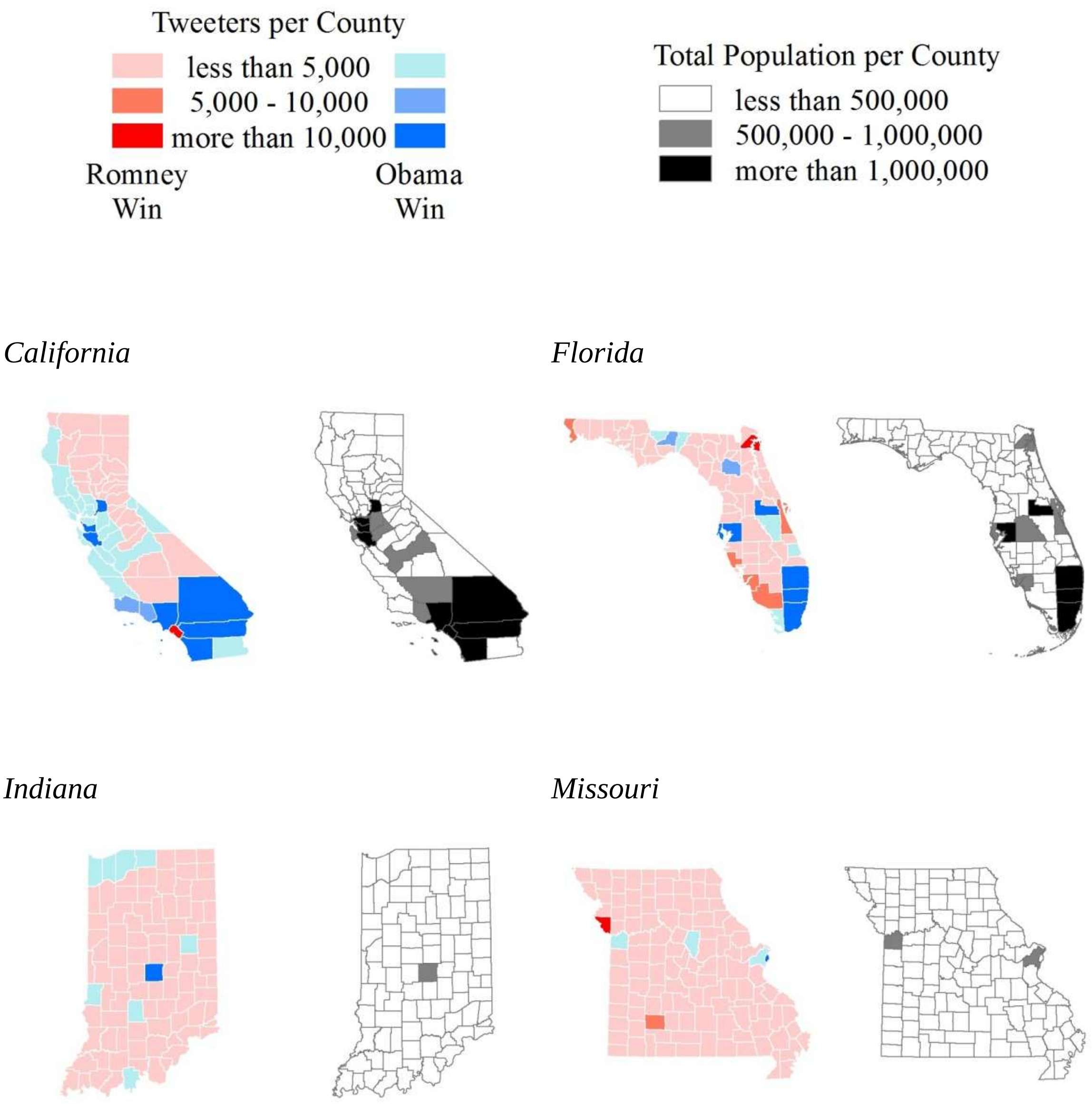


Figure 1: For California, Florida, and Indiana, counties with high amounts of Twitter users (shown in dark blue or red) also tend to have high total populations (shown in grayscale), which in turn tend to vote Democrat (shown in dark blue).

in dark blue). Texas is similar in its display of Twitter's pro-Democrat geographic bias, although there are some exceptions in counties around Dallas. These exceptions to the trend, as also evidenced in Missouri around Kansas City (seen in Figure 1), show that geography is only a part of the story. Despite having high amounts of Twitter users in highly populated counties that voted for Romney (shown in dark red), the demographic composition of Twitter use in these pro-Romney counties in Texas and Missouri is likely to be young, urban, and pro-Democrat.

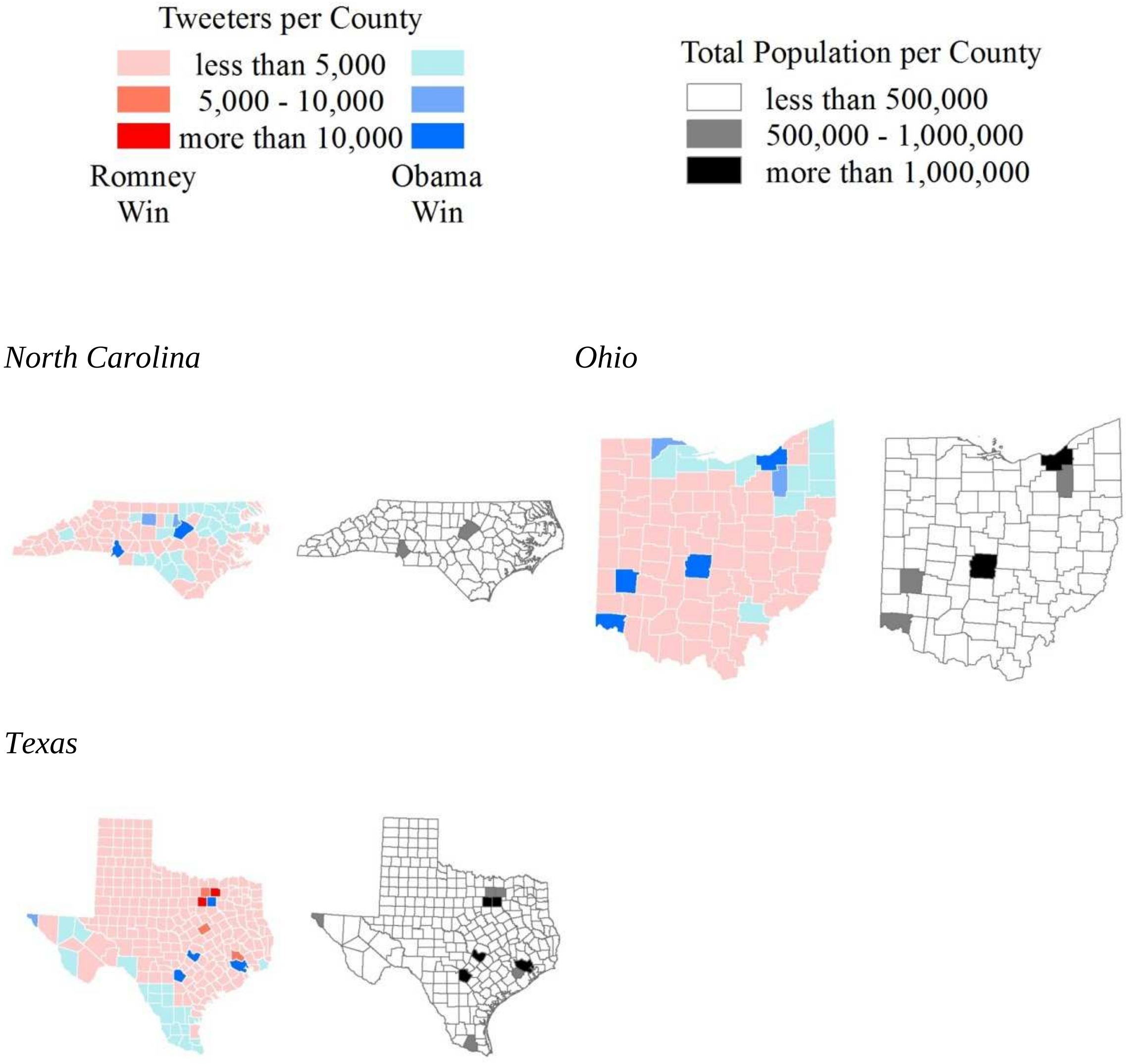


Figure 2: For North Carolina, Ohio and Texas, counties with high amounts of Twitter users (shown in dark blue or red) also tend to have high total populations (shown in grayscale), which in turn tend to vote Democrat (shown in dark blue).

CHAPTER V

CONCLUSION

The lessening of the MAE of the sentiment-derived Twitter election results from 13.1% in 2008 to 7.23% in 2012 would initially suggest that (1) Twitter is more predictive in 2012 than it was in 2008, because (2) Twitter's pro-Democrat bias had lessened due to a diffusion of use over the previous four years. Though, in actuality, the lessening of the MAE is likely due in large part to a more contested election in 2012, as evidenced primarily by closer statewide election results, and also by the lessening of the correlation between actual Democrat votes and population density.

The likelihood that the lessened MAE does not actually provide evidence of a greater diffusion of Twitter use is also statistically corroborated by the exponential increase in magnitude of use – from 20,000 users over six months in 2008 to 490,000 users over two months in 2012 – alongside the strong increase in positive correlation between Twitter use and population density. The increased correlation between Twitter use and population density shows that Twitter use in 2012 is more geographically biased, which would suggest that Twitter use in 2008 was at an early stage of demographic maximization, and that the concentrated utilization of Twitter in urban areas had yet to fully materialize.

Because of Twitter's tendency to closely follow lines of population distribution, and in combination with the current tendency in national politics to show a positive correlation between population density and counties with pro-Democrat election results, there is also a natural correlation between high levels of Twitter use and counties with pro-Democrat election results. Perhaps more interestingly, however, Twitter data tend to slightly oversample highly-populated counties, which in turn tend to further weight the importance of counties with pro-Democrat election results.

Nonetheless, despite strong evidence of geographic bias in Twitter use, and a prevalence of Twitter to slightly statistically overweight areas of greater population with pro-Democrat voting results, geography alone does not explain the pro-Democrat bias, as evidenced in part by the anomalous counties in Texas and Missouri that show high amounts of Twitter use in highly populated counties that voted for Romney in 2012.

One very prevalent non-spatial consideration, of course, is that Twitter use in 2012 has steadily increased among young people, as evidenced by the Pew Research Institute (Smith 2012). Another perhaps more elusive and non-spatial reason for the pro-Democrat bias, in the specific case study of Twitter, is related to cultural factors of participation, where urban areas have a greater lifestyle predilection toward internet and social media use than rural areas. This might also be due, in part, to simple issues of infrastructural capacity and access to the means of participation (e.g., internet, cellphone networks).

Sample bias in social media along lines of population distribution are not surprising, and pro-Democrat bias in Twitter in particular is also understandable, as evidenced by the aforementioned results and conclusions, but it should also be understood that, in a manner of speaking, these types of Big Data studies can make mountains out of molehills. The “votes” in this study are often based upon individual users’ few casual observations of candidates over a period of months, the “sentiment” of which is being derived by arbitrarily counting positive and negative words. The link is tenuous at best, and the method runs the risk of over-valuing otherwise offhanded and reactive moments in time. The primary difference between analyzing unsolicited Twitter data (as it pertains to political sentiment) and analyzing political surveys, is that the respondents of traditional surveys are (hopefully) thoughtfully engaging in explicit questions referencing the extent of their approval of the candidates. A tweet, on the other hand, hypothetically, may simply be a negative response to a poorly designed jingle or montage in a candidate's ad, and the polarity lexicon might very well mistake the meaning anyway.

Nonetheless, Big Data research will likely continue to grow in popularity as more data and means of analysis become available, and the arguments toward infrastructural proliferation of the means of data production will also probably bear fruit in the long run. The less urban areas will likely become more represented over time. For now, it is safe to say that Big Data, or at least big Twitter data, as Gayo-Avello originally found in his 2008 study (2011), is too biased a sample of urban and young people to be used for general purposes of electoral prediction, and more importantly, Twitter is more geographically biased now than it was in 2008.

APPENDIX

CODE ASSOCIATED WITH THESIS

The cURL command used to call on Twitter's Streaming API

-
  ```
  curl -u username:password
  "https://stream.twitter.com/1/statuses/filter.json?track=Romney,Obama" >
  $(date +%Y_%m_%d_%H_%M_%S).json
  ```

The HiveQL statements used to parse the data on AWS

-
  ```
  ADD JAR s3://gordonjar/hive-json-serde-0.3.jar ;
  ```

-
  ```
  CREATE EXTERNAL TABLE tweets (
  text string, id_str string, id bigint, coordinates_type string,
  coordinates_long float,
  coordinates_lat float, created_at string, country_code string, country
  string, full_name string, name string, user_id_str string, user_id bigint,
  user_location string
  )
  ROW FORMAT SERDE
  'org.apache.hadoop.hive.contrib.serde2.JsonSerde'
  WITH SERDEPROPERTIES (
  "text" = "$.text",
  "id_str" = "$.id_str",
  "id" = "$.id",
  "coordinates_type" = "$.coordinates.type",
  ```

```
"coordinates_long" = "$.coordinates.coordinates[0]",
"coordinates_lat" = "$.coordinates.coordinates[1]",
"created_at" = "$.created_at",
"country_code" = "$.place.country_code",
"country" = "$.place.country",
"full_name" = "$.place.full_name",
"name" = "$.place.name",
"user_id_str" = "$.user.id_str",
"user_id" = "$.user.id",
"user_location" = "$.user.location"
)
LOCATION 's3://gordontweets/twitter/' ;
```

-
  ```
  CREATE EXTERNAL TABLE tweetsexport (
  text string, id_str string, id bigint, coordinates_type string,
  coordinates_long float, coordinates_lat float, created_at string,
  country_code string, country string, full_name string, name string,
  user_id_str string, user_id bigint, user_location string
  )
  row format delimited fields terminated by ','
  STORED AS TEXTFILE LOCATION 's3://gordontweets/twitter/' ;
  ```

-
  ```
  FROM tweets
  INSERT OVERWRITE TABLE tweetsexport select regexp_replace(text,
  ",", "") as text, id_str, id, coordinates_type, coordinates_long,
  ```

```
coordinates_lat, created_at, country_code, country,
regexp_replace(full_name, ",", "") as full_name, name, user_id_str,
user_id, regexp_replace(user_location, ",", "") as user_location ;
```

The Python script used for the sentiment analysis

-
  ```
  # somewhere to hold the hit words in memory
  scoreDictionary = {}
  # read the dictionary file and store all entries with the given value
  def initDictionary(filename, value):
      file = open(filename,"r")
      for line in file:
          # drop the new line from the end
          word = line.strip().lower()
          # store the word in the dictionary with given value
          scoreDictionary[word] = value
      file.close()
  def parseLine(line):
      plusScore = 0
      plusWord = ""
      minusScore = 0
      minusWord = ""
      words = line.strip().split(",")[0].split(" ")
      for word in words:
          score = observeWord(word.lower())
  ```

```
        if score > 0:
            plusWord+=";"+word
            plusScore+=score
        elif score < 0:
            minusWord+=";"+word
            minusScore+=score
    return
plusWord+","+str(plusScore)+","+minusWord+","+str(minusScore)+","+s
tr(plusScore+minusScore)
def observeWord(word):
    if scoreDictionary.has_key(word):
        return scoreDictionary[word]
    return 0
def processTextFile(filename):
    # input data
    file = open(filename,"r")
    # where the output will go
    output = open(filename+"-output.csv","w")
    # write the header line to the output file
output.write(file.next().rstrip()+",plusWords,plusCount,minusWords,minu
sCount,totalScore\n")
for line in file:
    output.write(line.rstrip()+","+parseLine(line)+"\n")
output.close()
initDictionary("polarity_positive.csv",1)
initDictionary("polarity_negative.csv",-1)
processTextFile("Points.csv")
```

The Python script used to parse Census place names from Twitter location fields

```
• # somewhere to hold the hit words in memory
scoreDictionary = {}
# read the dictionary file and store all entries with the given value
def initDictionary(filename, value):
    file = open(filename,"r")
    for line in file:
        # drop the new line from the end
        word = line.strip()
        # store the word in the dictionary with given value
        scoreDictionary[word] = value
    file.close()
def parseLine(line):
    placeScore = 0
    placeWord = ""
    for word in scoreDictionary:
        if word in line:
            # do real stuff
            placeWord+=word
            placeScore+=scoreDictionary[word]
            #print word
    return placeWord+","+str(placeScore)
def processTextFile(filename):
    # input data
    file = open(filename,"r")
    # where the output will go
    output = open(filename+"-output_places_indi.csv","w")
    # write the header line to the output file
    output.write(file.next().rstrip()+",placeWords,placeCount\n")
    for line in file:
        output.write(line.rstrip()+","+parseLine(line)+"\n")
    output.close()
initDictionary("indiana_places.csv",1)
processTextFile("indiana_geo.csv")
```

REFERENCES CITED

Asur, Sitaram, and Bernardo A. Huberman. "Predicting the Future with Social Media." *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference On*. Vol. 1. (2010): 492–499.

Bermingham, Adam, and Alan F. Smeaton. "On Using Twitter to Monitor Political Sentiment and Predict Election Results." *Retrieved from http://doras.dcu.ie/16670/* (2011).

Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter Mood Predicts the Stock Market." *Journal of Computational Science* 2.1 (2011): 1–8.

Borondo, J. et al. "Characterizing and Modeling an Electoral Campaign in the Context of Twitter: 2011 Spanish Presidential Election as a Case Study." *Chaos: An Interdisciplinary Journal of Nonlinear Science* 22.2 (2012): 023138.

Boyera, Stéphane. "The Mobile Web to Bridge the Digital Divide." *IST-Africa Conference*. 2007. 9–11.

Bravo-Marquez, Felipe et al. "Opinion Dynamics of Elections in Twitter." IEEE, 2012. 32–39.

Brown, Greg, and Delene Weber. "Public Participation GIS: A New Method for National Park Planning." *Landscape and Urban Planning* 102.1 (2011): 1–15.

Bruns, Axel, and Jean Burgess. "Researching News Discussion on Twitter: New Methodologies." *Journalism Studies* 13.5-6 (2012): 801–814.

Ceron, Andrea, Luigi Curini, and M. Stefano. "Tweet Your Vote: How Content Analysis of Social Networks Can Improve Our Knowledge of Citizens' Policy Preferences. An Application to Italy and France." *Retrieved from http://www.sisp.it/files/papers/2012/andrea-ceron-luigi-curini-e-stefano-iacus-1414.pdf* (2012).

Chen, Lu, Wenbo Wang, and Amit P. Sheth. "Are Twitter Users Equal in Predicting Elections? a Study of User Groups in Predicting 2012 US Republican Presidential Primaries." *Social Informatics*. Springer, 2012. 379–392.

Choy, Murphy et al. "A Sentiment Analysis of Singapore Presidential Election 2011 Using Twitter Data with Census Correction." *arXiv preprint arXiv:1108.5520* (2011).

---. “US Presidential Election 2012 Prediction Using Census Corrected Twitter Model.” *arXiv preprint arXiv:1211.0938* (2012).

Chung, Jessica, and Eni Mustafaraj. “Can Collective Sentiment Expressed on Twitter Predict Political Elections.” *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence. San Francisco, CA, USA. Retrieved from http://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/viewPDFInterstitial/3549/4126* (2011).

Cummings, David, Haruki Oh, and Ningxuan Wang. “Who Needs Polls? Gauging Public Opinion from Twitter Data.” *Unpublished manuscript. Retrieved from http://nlp.stanford.edu/courses/cs224n/2011/reports/nwang6-davidjc-harukioh.pdf* (2010).

Elwood, Sarah. “Critical Issues in Participatory GIS: Deconstructions, Reconstructions, and New Research Directions.” *Transactions in GIS* 10.5 (2006): 693–708.

Gayo-Avello, Daniel. “Don’t Turn Social Media into Another ‘Literary Digest’ Poll.” *Communications of the ACM* 54.10 (2011): 121.

---. “I Wanted to Predict Elections with Twitter and All I Got Was This Lousy Paper a Balanced Survey on Election Prediction Using Twitter Data.” *arXiv preprint arXiv:1204.6441* (2012).

Henderson, Tristan et al. “Ethics and Online Social Network Research–developing Best Practices.” *Retrieved from http://www.cs.st-andrews.ac.uk/~tristan/pubs/bcsethics2012.pdf* (2012).

Jensen, Kasper Løvborg. “Sensible Smartphones for Southern Africa.” *interactions* 19.4 (2012): 66–69.

Jordan, Gavin. “GIS for Community Forestry User Groups in Nepal: Putting People before the Technology.” *Community participation and geographic information systems* (2002): 232–245.

Jungherr, A., P. Jurgens, and H. Schoen. “Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. ‘Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment’.” *Social Science Computer Review* 30.2 (2011): 229–234.

Kanjo, Eiman. “NoiseSPY: A Real-Time Mobile Phone Platform for Urban Noise Monitoring and Mapping.” *Mobile Networks and Applications* 15.4 (2009): 562–574.

Kessler, Fritz. “Volunteered Geographic Information: A Bicycling Enthusiast Perspective.” *Cartography and Geographic Information Science* 38.3 (2011): 258–268.

Lampos, Vasileios. “On Voting Intentions Inference from Twitter Content: a Case Study on UK 2010 General Election.” *Retrieved from http://eprints.pascal-network.org/archive/00009559/* (2012).

Maisonneuve, Nicolas et al. “NoiseTube: Measuring and Mapping Noise Pollution with Mobile Phones.” *Information Technologies in Environmental Engineering*. Ed. Ioannis N. Athanasiadis et al. Berlin, Heidelberg: Springer Berlin Heidelberg, (2009): 215–228.

Metaxas, Panagiotis T., and Eni Mustafaraj. “Social Media and the Elections.” *Science* 338.6106 (2012): 472–473.

Metaxas, Panagiotis T., Eni Mustafaraj, and Daniel Gayo-Avello. “How (not) to Predict Elections.” *Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom).* (2011): 165–171.

Mislove, Alan et al. “Understanding the Demographics of Twitter Users.” *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM’11), Barcelona, Spain. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2816/3234* (2011).

Murthy, D. “Twitter: Microphone for the Masses?” *Media, Culture & Society* 33.5 (2011): 779–789.

Newman, Greg et al. “The Future of Citizen Science: Emerging Technologies and Shifting Paradigms.” *Frontiers in Ecology and the Environment* 10.6 (2012): 298–304.

O’Connor, Brendan et al. “From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series.” *Proceedings of the International AAAI Conference on Weblogs and Social Media* (2010): 122–129.

Pernisco, Nick. “The Ubiquity Myth.” *Retrieved from http://www.understandmedia.com/journals-a-publications/44-scholarly-articles/141-the-ubiquity-myth* (2011).

Roche, Stéphane et al. “WikiGIS Basic Concepts: Web 2.0 for Geospatial Collaboration.” *Future Internet* 4.4 (2012): 265–284.

Sang, E. T. K., and Bos, J. “Predicting the 2011 Dutch Senate Election Results with Twitter.” *Proceedings of the Workshop on Semantic Analysis in Social Media* (2012): 53-60.

Signorini, Alessio, Alberto Maria Segre, and Philip M. Polgreen. “The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. During the Influenza A H1N1 Pandemic.” Ed. Alison P. Galvani. *PLoS ONE* 6.5 (2011): e19467.

Smith, Aaron, and Joanna Brenner. “Twitter Use 2012.” *Pew Internet & American Life Project*. *Retrieved from http://alexa.pewinternet.com/~/media/Files/Reports/2012/PIP_Twitter_Use_2012.pdf* (2012).

Starbird, Kate, and Leysia Palen. “Voluntweeters: Self-organizing by Digital Volunteers in Times of Crisis.” *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*. (2011): 1071–1080.

Sui, Daniel Z. “The Wikification of GIS and Its Consequences: Or Angelina Jolie’s New Tattoo and the Future of GIS.” *Computers, Environment and Urban Systems* 32.1 (2008): 1–5.

Takhteyev, Yuri, Anatoliy Gruzd, and Barry Wellman. “Geography of Twitter Networks.” *Social Networks* 34.1 (2012): 73–81.

Thelwall, Mike, Kevan Buckley, and Georgios Paltoglou. “Sentiment in Twitter Events.” *Journal of the American Society for Information Science and Technology* 62.2 (2011): 406–418.

Trumper, Diego Saez, Wagner Meira, and Virgilio Almeida. “From Total Hits to Unique Visitors Model for Election’s Forecasting.” *Retrieved from http://journal.webscience.org/473/* (2011).

Tumasjan, A. et al. “Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape.” *Social Science Computer Review* 29.4 (2010): 402–418.

---. “Where There Is a Sea There Are Pirates: Response to Jungherr, Jurgens, and Schoen.” *Social Science Computer Review* 30.2 (2011): 235–239.

Ueno, Kenta Sasaki Shinichi Nagano Koji, and Kenta Cho. “Feasibility Study on Detection of Transportation Information Exploiting Twitter as a Sensor.” *Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/download/4752/5105* (2012).

Vallina-Rodriguez, Narseo et al. “Los Twindignados: The Rise of the Indignados Movement on Twitter.” *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*. (2012): 496–501.

Van Liere, Diederik. “How Far Does a Tweet Travel?: Information Brokers in the Twitterverse.” *Proceedings of the International Workshop on Modeling Social Media*. (2010): 6.

Vieweg, Sarah. “The Ethics of Twitter Research.” *Revisiting Research Ethics in the Facebook Era: Challenges in Emerging CSCW Research, Savannah, GA. Retrieved from http://www. cc. gatech. edu/∼ yardi/ethicscscw2010_files/AcceptedPapers. htm* (2010).

Weiner, Daniel et al. “Apartheid Representations in a Digital Landscape: GIS, Remote Sensing and Local Knowledge in Kiepersol, South Africa.” *Cartography and Geographic Information Science* 22.1 (1995): 30–44.

Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. “Recognizing Contextual Polarity in Phrase-level Sentiment Analysis.” *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. (2005): 347–354.