SENTIMENT MAPPING: POINT PATTERN ANALYSIS OF SENTIMENT CLASSIFIED TWITTER DATA

By

Kenneth Camacho

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Geography – Master of Science

2020

ABSTRACT

SENTIMENT MAPPING: POINT PATTERN ANALYSIS OF SENTIMENT CLASSIFIED TWITTER DATA

By

Kenneth Camacho

Varieties of sentiment analysis and point pattern analysis are being applied to social media data to address a broad range of questions, but they are rarely used in tandem. This study outlines a methodology that combines these two approaches to analyze the spatial distribution of sentiment classified opinions from social media data. Twitter postings on natural gas were downloaded and classified using a variety of sentiment analysis methods into positive, negative, and neutral categories. The classifications were then converted into spatial points using the location data associated with the tweets, whereby point pattern analysis techniques were applied to the points to examine the patterns of positive and negative tweet locations with respect to a background rate of neutral tweets across the contiguous Unites States. Basic temporal visualizations were also constructed to explore the variations in sentiment over time. Considerations are discussed on the accuracy limitations of sentiment analysis and the potential for a variety of applications using these techniques. With careful implementation, this methodology can open the door to a range of spatiotemporal analyses of social media sentiment.

ACKNOWLEDGEMENTS

I am indebted to all who supported me in the writing of this Thesis. This includes my advisor, Dr. Raechel Portelli, and my other committee members, Dr. Ashton Shortridge and Dr. Bruno Takahashi, whose guidance and curiosity were essential in shaping this document. I also extend my gratitude to Dr. Nathan Moore and Dr. Shiyuan Zhong for their mentoring and friendly optimism throughout my stay in the department.

I thank Pietro Sciusco, who gave his time to serve as the secondary coder of the manually classified tweets. I would also like to express thanks to him and the other Lunch Buddies for their camaraderie as we navigated our way through graduate school together.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

From Volunteered Geographic Information to Geospatial Twitter Mining

Ever-expanding in their response to new technologies, the contributions made by volunteered geographic information (VGI) have transformed geography. Data collection has become informalized to the point where citizens are now the primary data collectors for many studies, in fields as disparate as ornithology, epidemiology, and forest pathology (Connors, Lei, & Kelly, 2012). There is a new appreciation for the collective capacity of ordinary people to provide data for researchers to take in, analyze, and share as results.

In a review of volunteered geographic information, Goodchild (2007) describes the people contributing this information as sensors. This mindset is useful when considering volunteered information as it implies that researchers should observe the normal considerations that are used for other geographic sensors such as seismographs and weather stations; these considerations include how the sensors are distributed in space, their precision and accuracy, how often they report readings, how to calibrate them, and so on. At the time, Goodchild was referring primarily to the uses of Wikimapia, Flikr, Google Earth, and OpenStreetMaps, but today the literature has expanded to include the immense depths of data available from social media platforms such as Twitter, a microblogging platform with over 300 million active users (The Statistics Portal, 2018).

Few papers take the "citizens as sensors" view more literally than *#Earthquake*, an article which demonstrated that posts about an earthquake on Twitter could make a beneficial supplement to other spatial VGI sources such as the USGS 'Did You Feel It?' website. The paper analyzed the reaction time of Twitter users to an earthquake in order to understand where

and when the shaking was felt. The paper also stands testament to the scale of Twitter. Despite having access to only 1% of tweets, the authors were able to identify over 100 geolocated earthquake tweets posted within two minutes of the event and nearly 1,000 within five minutes (Crooks, Croitoru, Stefanidis, & Radzikowski, 2013).

If Twitter users are data collecting sensors, what are they capable of sensing? Evidently, their capacities go well beyond those of a seismograph. Users are sensing the experiences of other people and sharing their own lived experiences which, rather than being limited to the recording of events, emphasize sentiments that reflect their engagement with the world. The real value of Twitter users as sensors is not likely to be found in their ability to detect readily observable environmental phenomena, which other technologies and platforms are better designed for, but instead in their ability to sense and report on the human social environment.

Twitter Mining Human Sentiment

There are a growing number of disparate researchers exploring new ways to analyze social media posts and interpret them in ways that reveal something meaningful about the nondigital, human world. Broadly speaking, their papers examine different expressions of human *sentiment*; that is, the opinions or feelings expressed by people in general or with regard to a specific topic. The study of sentiment can be limited by means of geography, time period, demographic group, and topic to produce answers to different research questions. Opinion polling is a popular example of measuring broad human sentiment that has been hugely impactful on American politics and other areas of life (Erikson & Tedin, 2001; Herbst, 1995). In a less direct way, Twitter sentiment can address similar topics to opinion polling as well as a host of other issues. A small sample of this work will be explored in this section.

Twitter provides an imperfect window into human psychology that, with enough data mining, has the potential to offer insights on human social dynamics that other sources of data cannot easily provide. Early in its lifespan, Twitter was established as a platform that could be used to monitor the emotional states of its user base, with researchers tracking the changing moods of Twitter users over time (Roberts, Roach, Johnson, Guthrie, & Harabagiu, 2012). Emotions on the website have been found to track real-world social, cultural, and political events (Bollen, Mao, & Pepe, 2011). Others have used this mood data to pursue a direction focused on trends in mental health (Larsen et al., 2015). That the sentiment on Twitter has been demonstrated to respond to time-sensitive real-world phenomena suggests an abundance of other research opportunities as well.

In addition to tracking mood, Twitter mining was also quickly adapted for use as a means of studying opinions such as political sentiment. Following the spread of Occupy Wall Street protests, Twitter data were employed to examine the spread of anticapitalistic ideas and learn more about the messaging backing them, as well as how members of the movement organized themselves (Conover et al., 2013). A number of studies also came out in response to the Brexit referendum. Some attempted to use Twitter to predict the outcome of the voting while others sought to explore links between political statements on twitter and general opinion on political issues (Freitas et al., 2016). While the latter exploration of opinions seems to be an appropriate use of Twitter data, election predictions may fall outside of what is currently feasible. Gayo-Avello (2012) provided a cautionary tale concerning the use of Twitter posts to make predictions about political outcomes. At the start of the decade, it was common to see publications and conference speakers claiming an ability to predict election results using Twitter data. These results were short-lived, however, with few researchers able to make reliable predictions about

future elections. Gayo-Avello explains how the promise shown by these early studies was a consequence of faulty assumptions, biased data, and the file-drawer effect, which is the tendency of researchers to abstain from reporting negative results (Fanelli, 2010).

An important lesson can be learned from this and other cautionary tales found in the literature. When generalizing data on public sentiment, compare apples to apples only – Twitter does not necessarily represent the true average of the public's feelings on a given topic. There is a selection bias inherent in using Twitter data in that Twitter posts are made by people who find it prudent to share their opinion on potentially contentious topics publicly. Thus, the demographics of Twitter users do not align neatly with US demographics. By margins of 10% or greater, users are higher income, younger, and more highly educated than the general population. Politically, 60% of users lean Democratic (compared to 52% in the US population) and 35% lean Republican (compared to 43% in the US population). Further, 10% of US tweeters are responsible for 80% of US tweets, skewing the results further. These top 10% of users are 65% female and nearly four times more likely to post political content than other users (Wojcik & Hughes, 2019). It can make sense to compare sentiments of Twitter users in one geographic area to those in another area, but to make assertions about the general public from these demographics without due diligence is careless. That said, there are still researchers who try. For example, by analyzing public mood through tweets, Bollen & Mao (2011) attempted to predict changes in the stock market. Like many others that tried to use Twitter as a true mirror of the nondigital world, this research did not provide long-term success.

While many of these studies show promise in analyzing the opinions and moods expressed on Twitter, they do not take advantage of all the dimensions of the data. Twitter users

are sentiment sensors of the population that are located somewhere in space. The distribution of these users is also open to analysis and may be just as significant as the sentiment they are expressing.

Geospatial Twitter Mining

A multitude of studies demonstrate that new insights can be obtained by applying a geographic lens to Twitter data. These studies range in topic but are generally congregated in the social sciences. Lessons learned from the areas of public health, environmental analysis, and sentiment-related approaches will be covered in this section.

Public Health

From its beginnings, one of the prominent uses of geospatial Twitter data has been the analysis of the spread of infectious diseases across time and space. Lessons can be drawn from this research that are worth repeating here. Early work using Twitter to identify influenza outbreaks began around 2010, initially focusing on the methods of classifying tweets and constructing accurate models that could match with CDC reports of influenza outbreaks (Achrekar, Gandhe, Lazarus, Yu, & Liu, 2011; Aramaki, Maskawa, & Morita, 2011; Culotta, 2010). The flu was an exciting spatiotemporal issue to explore at the time using social media trends, in part fueled by the early success of Google's Flu Trends service, a web-based tool that used Google search queries as a method of early disease detection (Carneiro & Mylonakis, 2009).

It is prudent to highlight here that this success did not last. Google Flu consistently overpredicted flu outbreaks, resulting in a service that did little to improve on existing tools for tracking and predicting the spread of the flu. The story has become something of a parable for

the limitations of big data, primarily when used in fields where there are already well-established methods for fulfilling similar tasks. Big data should prove to be most useful in areas where it can touch on novel topics rather than areas that have already been thoroughly developed (Lazer, Kennedy, King, & Vespignani, 2014).

Since the shutdown of Google Flu trends in 2015, the social media big data space surrounding disease appears noticeably different. Some have changed course to focus on related but perhaps more appropriate topics for social media investigation, such as the public's awareness of influenza outbreaks (Smith, Broniatowski, Paul, & Dredze, 2015). Other researchers have chosen to focus on alternative diseases, such as heart disease (Eichstaedt et al., 2015) or Zika (McGough, Brownstein, Hawkins, & Santillana, 2017), while others have given their attention to the social side of disease outbreak and public health more broadly (Daughton, Paul, & Chunara, 2018; Karami, Dahl, Turner-McGrievy, Kharrazi, & Shaw, 2018; Lee et al., 2016).

Environmental Issues

Twitter data is often used during and after environmental calamities to foster situational awareness and understand the human response to these events. Already mentioned is the use of *#Earthquake* (Crooks et al., 2013), which used humans as sensors of a natural disaster. Other researchers have used Twitter to examine human responses to earthquake events (Vo & Collier, 2013) as well as wildfires (Z. Wang & Ye, 2016), focusing on the emotional reaction and content of the discourse surrounding these events, similar in spirit to some of the sentiment-related topics discussed above. Other natural disasters have been analyzed as well. Using a combination of tweet content and analysis of the movement patterns before, during, and after hurricane events

(as gleaned from the same tweets), researchers can learn information about behavioral responses to natural disasters (Stowe, Anderson, Palmer, Palen, & Anderson, 2018). This information can be used for a variety of purposes, from the behavioral sciences to the disaster response planning of emergency services.

Some work has been done in analyzing Twitter data to learn about environment-related perspectives across space, but this research area is largely unexplored. One example is a national study on fracking perspectives in the US. The authors analyzed tweets from a variety of stakeholders to determine how perspectives were distributed spatially and which users were more likely to have their ideas diffused across the platform. The authors of this study also made an essential point regarding environmental perspectives: a user will probably be less likely to share their viewpoint if they feel that their perspective is the dominant one in their area, or if the issue no longer pertains to their geographic location. For example, the authors found far less discussion on fracking in New York, where the practice has been banned, than in California, where active policy debate on the subject occurred during the study period (Sharag-eldin, Ye, & Spitzberg, 2018).

Sentiment-Related Topics

Less studied geospatially are the subjects that are unique to humans and are expressed primarily through language: emotions, personal preferences, and opinions. Being a platform that reflects the collective sentiment of individuals, Twitter has the potential to provide a plethora of insights on these topics so long as appropriate analyses are run. In terms of the effort and financial inputs needed to run the research, it is an inexpensive means of comparing geographic trends over broad areas. The added value can be made clear by briefly returning to research on

Brexit sentiment. A recent study used tweets to examine the global distribution of "stay" and "leave" sentiments during the time of the voting period, allowing for comparisons across countries inside and outside the European Union (Agarwal, Singh, & Toshniwal, 2018), and demonstrating the value that can be added when geospatial components are incorporated into data analysis. An opinion poll covering the same countries would have been an expensive exercise.

A range of studies exist that can shed light on the breadth of possibility in geospatial analysis on more human topics, but they also highlight the disconnectedness of the research. Many of these studies stand alone, with no clear predecessor or follow-up. One study examined the connectivity of users online with comparison to their physical location in the world. Perhaps unsurprisingly, distinct geographic and cultural groups have developed on Twitter, a platform that is aspatial (Kulshrestha, Kooti, Nikravesh, & Gummadi, 2012). Some authors have examined the spatial distribution of music tastes, developing clusters of similar artists and analyzing local trends based on network interactions and the spatial distribution of listening trends (Hauger & Schedl, 2014). Some focus on extracting twitter data over time from a single location. The instrument *EmoTwitter* detects and visualizes Twitter discussions taking place near a given location over time (Kobayashi, Mozgovoy, & Munezero, 2016). A more prominent example tracked the spatial spread of ideas across nearly six million tweets, finding that comparatively small clusters of tweets played a vital role in the diffusion of discussion (Ardon et al., 2011). However, even this more prominent study did not have an obvious follow-up. It is not entirely clear why using geospatial twitter data for these purposes tends to have minimal impact on the literature. Perhaps the studies are too novel and disconnected, or there is a lack of interest or trust in sourcing human data from social media platforms.

Geospatial Sentiment Analysis

The aforementioned studies rely primarily on simple, straightforward analyses such as the presence of certain hashtags to determine the sentiment of tweets. More complex methods for determining the sentiment of a piece of text are often accomplished using techniques from *sentiment analysis*, a field of natural language processing specializing in extracting sentiment from a body of text. The most common methods in sentiment analysis either compare the words in the text to a lexicon of words with known sentiment or use machine learning with a body of manually classified training data to determine the sentiment (Liu, 2012). Using sentiment analysis to analyze tweets rather than more simplistic methods introduces additional uncertainty into the analysis – the highest accuracy of these methods tends to lie in the mid 80% range (Saif, He, Fernandez, & Alani, 2016). However, utilizing sentiment analysis to classify tweets provides many new opportunities for deeper analysis, particularly when it is used to inform geographic analyses.

The literature combining sentiment analysis of tweets with geographical analysis is extremely sparse, but some examples do exist. A 2013 study using geotagged tweets analyzed the spatial distribution of Twitter sentiment in New York City to determine areas with positive and negative sentiment (Bertrand, Bialik, Virdee, Gros, & Bar-Yam, 2013). The tweets were classified with a machine learning approach and the resulting maps offer an exploratory means of visualizing mood at the census block level. The study excelled in producing high resolution maps of the mood across New York City over the data collection period and tying these sentiments to real-world features, though the accuracy of the sentiment analysis is questionable, and the geographic analysis was very limited. Another example of visualizing emotions through sentiment analysis is the We Feel project by Larsen et al. (2015). The sentiment analysis in this

case was performed with a lexicon-based method on tweets globally, aggregating the data at the national scale to compare the results to mental health indices of different nations. Temporal analyses were also conducted to examine the rates of different emotions over time.

A tool entitled GeoSentiment was developed to track the sentiment response to events on Twitter interactively across different locations (Pino, Kavasidis, & Spampinato, 2016). Although the concept is promising, the methodology is unclear on the sentiment analysis techniques and only geotagged tweets were used, which excludes the vast majority of Twitter posts. One interesting component is the use of kernel density maps to display the density of tweets across the area of study. A final paper worth mentioning used twitter data to analyze the political sentiment in different areas of East Java, Indonesia using a machine learning approach to the sentiment analysis (Fahrur Rozi et al., 2018). Again, the concept is exciting, but the data set was quite small, there are significant issues with the sentiment analysis, and the geographic component of the data was only used for visualization purposes.

Common threads tie these studies together. They often succeed in grounding their findings with reference to real-world phenomena, which is key for building credibility in Twitter data. Although they use primarily geotagged tweets, which are accurate but limit the sample size, the geographic component of the data is used only for aggregation and/or visualization rather than any substantive analysis. In addition, the sentiment analysis is consistently underdeveloped. Where the methods are clear enough to interpret, only one method of classification was applied to the data, and normally toward the classification of moods. Although geographic mood data is useful, there is potential for far more expansive applications of sentiment analysis.

These studies point in the direction of a powerful new application of Twitter data without fully realizing its extent. With a more thorough exploration of sentiment analysis, truer and more

meaningful sentiment can be extracted from the data. With a more developed geographic analysis, patterns and trends can be explored in a more substantive and quantifiable manner. For the latter, there is potential to draw from *point pattern analysis*, a geographic toolkit for analyzing the distribution of points that is often applied in the fields of epidemiology and ecology (Gatrell, Bailey, Diggle, & Rowlingson, 1996; Velázquez, Martínez, Getzin, Moloney, & Wiegand, 2016). Like diseases or living things, the distribution of tweets of varying sentiment can be quantified and explored spatially to learn meaningful things about their spread and location in space. What remains is to implement these techniques.

Developing a Methodology for Sentiment Mapping

At this time, the research surrounding Twitter mining is unconnected and without any standardized methodology. Despite its already broad applications, there is considerable room for the use of Twitter data to be expanded into new fields and analyzed spatially with point pattern analysis to answer new questions or revisit old inquiries, especially toward the study of human systems such as healthcare, economics, politics, psychology, and human-environment issues. Like any area of big data, the results of Twitter mining should be interpreted with caution; Twitter users are not ambassadors for humanity as a whole. Many exciting projects have started well, only to fail later (Gayo-Avello, 2012; Lazer, Kennedy, King, & Vespignani, 2014), highlighting the need for prudence. However, Twitter mining shows promise as a useful tool for understanding mass human perceptions, behaviors, and experiences – a side of research that has been underexplored. With methods to analyze this sentiment used in tandem with point pattern analysis techniques, a new space can be created for the spatiotemporal interpretation of social media sentiment. In this paper, potential methodology is outlined for these exploratory techniques, the result of which is dubbed sentiment mapping.

2. METHODS

A summary of the methods outlined in this Methods section is visualized in Figure 2.1. Like the bolded headings in the figure, the section is divided into four main parts: data collection, location analysis, sentiment analysis, and spatial analysis. The methodology is presented as a linear sequence of steps but experimentation across all stages of the methods should happen concurrently. For example, experimentation with sentiment analysis should begin as soon as the first sets of tweets are downloaded. To do otherwise is to risk months of data collection on a subject that may be difficult or impossible to classify with the sentiment analysis techniques discussed below. In the case of this research, the topic that was chosen as a demonstration of the spatial analysis of Twitter sentiment is natural gas. This topic has been explored previously, but only at the state level, and without the use of point pattern analysis or algorithm-assisted sentiment analysis (Sharag-eldin et al., 2018).

A theme of these methods is the cycle of automated work with human validation and modification. The majority of the process is handled through scripting, but every step requires careful investigation of the results to avoid errors, find patterns, rework assumptions, choose the best algorithms, and so on. Sentiment mapping is therefore a subjective process, making it paramount that methods be as clear as possible about intermediate results and the decisions that led to the final maps. Further, given that the process involves interdisciplinary methods and that the results of the research will be more applicable to social scientists than computer sciences, it seems sensible to work with approaches, as much as they are available, that do not require a deep understanding of machine learning algorithms or spatial statistics.

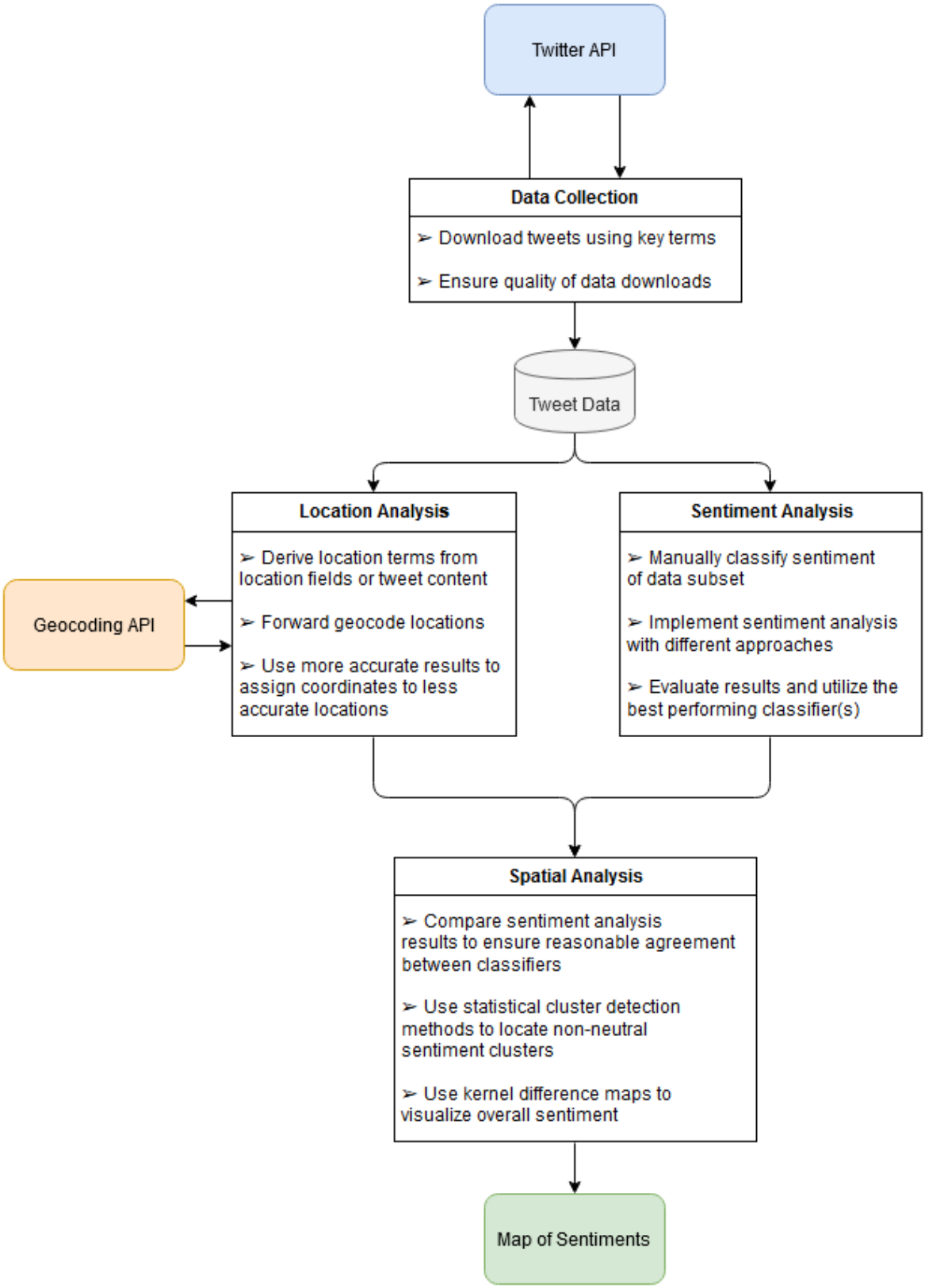

Figure 2.1. Flowchart of sentiment mapping methodology.

Data Collection

Tweets were downloaded and stored over an eight-month period from June 2019 to January 2020. The data were retrieved via requests to Twitter's Standard API, available to all Twitter developer accounts, which provides access to a sample of 1% of tweets from the previous 6-9 days (Twitter, 2020). Over the data collection period, tweets were downloaded in batches once per week with the programming language R (R Core Team, 2020) and with the assistance of the "rtweet" package (Kearney, 2019). The convenient package formats the returned data as a data table, which is a format that is easy to manipulate and export from R. Data were stored locally and backed up with OneDrive's cloud storage service.

Requests to Twitter's Standard API require the name of the account, a consumer key, and an access token. With this, a set of key terms can be passed through as a request along with a number of other parameters, such as a limit on the number of tweets retrieved, the language of the text, and whether or not to include retweets or media items. For this analysis, tweets were limited to the English language and retweets were not retrieved. Retweets were ignored to avoid exceeding the maximum number of allowable requests and because retweets are not necessarily signals of agreement – with some additional text accompanying a retweet, the retweet can either support or oppose the sentiment of the original tweet, creating problems for the sentiment analysis phase of the methodology. In short, a retweet is not equivalent to "I agree."

The key terms passed to the API were prominent nonrenewable electricity production methods including "petroleum", "natural gas", "coal", and "nuclear energy". These terms were intentionally broad in order to include the broadest range of the conversation on the national scale. Studies on a more local scale or of a more specific phenomenon would benefit from more specific search terms. By examining the returned tweets across these different topics, it was

decided that the focus of this study is tweets containing the key words “natural gas”. Out of the range of topics examined, these tweets were the most focused, with discussions revolving almost exclusively around natural gas energy production. They were also the most balanced, with a similar amount of positive and negative sentiment on display toward natural gas. In total, 366,353 tweets containing the phrase “natural gas” were collected over the eight-month period.

For each request, the API returns 88 fields for every tweet available that matches the key terms given. Although all fields were saved, the analyses performed were completed using only seven dimensions: 1) *created_at*, a datetime field that contains the timestamp of each tweet, necessary for temporal analysis 2) *screen_name*, the field that contains the username of the poster, useful for filtering out bots and finding other data quality issues, 3) *text*, containing the text content of each tweet, 4) *location*, a user-entered field from the profile wherein users describe their location, 5) *geo_coords*, the latitude and longitude coordinates from where the tweet was posted, available only if the tweet was geo-tagged, 6) *place_full_name*, which has the full name of the Twitter Place of the tweet, which similarly is only available if provided with the tweet, and 7) *lang*, a language field used to filter for tweets in the English language. Fields 4-6 were used to ascribe a location to each tweet, as will be described in more detail in the next section.

Following a filter for tweets in the English language only, a total of 283,297 tweets remained for further analysis, totaling approximately 111 MB when stored in a single a comma-delimited file. Of these, 217,746 had information in one or more of the *location*, *geo_coords*, or *place_full_name* fields, allowing for the opportunity of assigning geographic coordinates to these tweets. From there, a random subset of 5,000 of these tweets was set aside for preliminary location analysis and the training and parameterization of the different sentiment analysis

techniques. This 5,000-tweet sample will hereafter be referred to simply as the *sample*. In addition to its foundational necessity for location and sentiment analysis, the sample also serves a key role in evaluating the final spatial analysis and map of sentiment. Although the sample should not be expected to align neatly with the final analysis, it can be used as a sanity check. The relative ratio of positive to negative to neutral sentiment, clustering of the data, and geographic trends can be expected to reflect some of the qualities of the sample.

Location Analysis

The majority of tweets are not geotagged, and therefore do not provide direct coordinate information. One study found that only 1.5% of tweets have a geotagged location associated with them (Zandbergen & Barbeau, 2011). More analytic methods must then be applied to the data to infer the location from which a user is posting, using both the *location* and *place_full_name* fields. However, it cannot be assumed that these fields are always accurate. A 2011 study estimated that only 64% of users input accurate location information in their profile, though accuracy level (city versus state level, for example) can vary (Hecht, Hong, Suh, & Chi, 2011). See Ikawa, Vukovic, Rogstadius, & Murakami (2013) for an example of an evaluation of these different location accuracy levels and their effects on the spatial precision of tweets. The takeaway message is that assigning coordinates to tweets can be an inaccurate process if care is not taken to evaluate geocoding returns carefully and remove erroneous results.

The process of assigning coordinates to tweets was carried out in R software with assistance from the "rgeocodio" package, designed to assist with API connections to the forward geocoding service Geocodio (Rudis & Thompson, 2018). The initial geocoding runs were performed on the 5,000-tweet sample in order to evaluate the returns from Geocodio. In response to text cleaned and submitted from the *location* field and *place_full_name*, the service returned

not only the estimated latitude and longitude for each location but also an approximated accuracy score ranging from zero to one, with one being the most accurate; the accuracy level of a location, be it rooftop level, street level, a place (such as a city or zip code), or state; and the source of the data used to find the coordinates. The most common data source is the TIGER/Line dataset from the US Census Bureau, indicating the majority of returned coordinates are centroids of US Census tracts, an accuracy level that is more than sufficient when analyzing spatial trends at the national scale.

Several lessons were learned from the initial geocoding runs performed on the sample. First, only 3508 out of the 5000 tweets were successfully geocoded – and these were tweets that had already been filtered to ensure there was text present in one of the location fields. This indicates that much of the information users input into the location fields in their user profiles is inscrutable as location data.

Second, out of the records with coordinates returned, 2668 had an accuracy score of one, indicating the highest possible accuracy return from Geocodio. After going through the returns with an accuracy below one, it was established that only tweets with the highest accuracy score would be utilized in the final geographic analyses. This is was decided upon after seeing the high frequency of errors present in lower accuracy levels. With the filtering out of unsuccessfully geocoded locations and accuracies below one, approximately half of the tweets with information in a location field were retained for geographic analysis.

Third, after evaluating each record by comparing the user's location to the coordinates returned by Geocodio, it was discovered that even some of the returns with accuracy scores of one were spatially inaccurate. For example, all instances of the location "Earth" were assigned to the small town of Earth, Texas with high confidence despite it being unlikely that any given user

was actually from that location. These falsely accurate errors were commonly associated with a relatively small list of locations, many of which are located outside of the United States. For the final geocoding of the full dataset, places containing the following names removed: China, Ontario, England, Mexico, Ireland, Laguna, Russia, Turkey, Ottawa, Alberta, Columbia, Earth, Norway, Delhi, Arab, Edmonton, and Nottingham. Locations referencing Silicon Valley also created some inaccurate coordinates, so for the full dataset, instances of "Silicon Valley" were replaced with "San Jose" before geocoding.

With this information gleaned from the subset, the process for assigning coordinates to the full dataset was developed. First, the *location* field was replaced by the *place_full_name* field where it was available, because unlike *location* it is selected from an established list Twitter locations and is thus easier to geocode. Second, the listed names above were altered or removed from the location because of the accuracy issues they presented. Third, the *location* text of the tweets was uploaded to the Geocodio servers as a CSV file, with the coordinates and accompanying information returned after approximately one hour. Fourth, the results were examined to find and correct any common errors in accuracy scores. Fifth, the coordinates were overwritten for tweets with coordinates in the *geo_coords* field, their accuracy score was updated to be one, and their accuracy level was reassigned to "coordinates". Sixth, tweets without any coordinate information were removed from the dataset.

This process resulted in 160,768 tweets with assigned coordinates. According to the Geocodio return, dozens of sources were used to geocode these coordinates, with the most common being the TIGER/Line dataset from the US Census Bureau, which was used to geolocate 135,766 of the tweets. The rest of the sources were mainly geocoded using state or city-level datasets. As can be seen in Table 2.1, the overwhelming majority of the points are at

the accuracy level of *place*, which would not be precise enough for any local analysis but will suffice for analyzing national trends. Aside from tweets at the state accuracy level, the remaining categories are more than precise enough for analysis at a broad scale. The state-level data are too coarse even for the national scale and must either be removed or processed further before they can be utilized in the geographic analysis.

Table 2.1. Accuracy level of the tweets following the completion of location analysis. An accuracy score of 1 indicates that the geocoder had the highest possible confidence in the coordinates returned.

Accuracy level	**Number of tweets**	**Tweets with accuracy score of 1**
Place (cities, zip codes, etc)	133,233	94,029
State	21,329	21,329
Coordinates	3710	3710
Street center	1224	887
Rooftop	1067	1060
Intersection	193	0
Range interpolation	12	11

In order to retain the state-level data, it must be localized somehow, which is possible only though an imprecise process. If left as state centroids, these tweets can create artifacts in the spatial analysis phase that do not correspond with any real phenomena (see Figure 2.2). The process for localizing the state-level tweets involves knowing the sentiment of each tweet, so it will be explored further in the spatial analysis section.

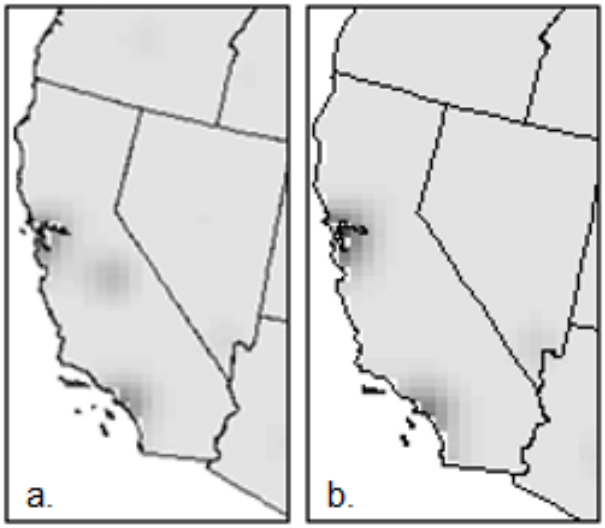

Figure 2.2. A kernel density map of California tweets (a) before and (b) after localizing state-level tweets. An apparent clustering of tweets appears at the centroid of the state in (a) as an artifact of state-level geocode returns being assigned to state centroids.

Sentiment Analysis

There are a variety of methods to choose from when classifying the sentiment of text. For this analysis, one lexicon-based technique and three machine learning techniques were applied to the data. Analyses were carried out using the python programming language and two essential libraries: 1) the "Natural Language Toolkit" (NLTK) library (Loper & Bird, 2002), which provides an array of corpuses, preprocessing tools, sentiment classifiers, and other modules for simplifying the sentiment analysis process, and 2) the "scikit-learn" library (Pedregosa, Weiss, & Brucher, 2011), which is a broad machine learning library used for this research to build and evaluate the different machine learning classifiers used for the sentiment analysis.

Before performing any classifications using computer algorithms, a set of manually classified tweets is necessary as a benchmark for assessing the accuracy of the classifiers and training the machine learning algorithms. The 5000-tweet subset was manually classified into three sentiment categories: positive, negative, and neutral, based on each tweet's position toward

natural gas. In order to determine the sentiment of each tweet, the question asked was, roughly, "does this text show a preference for the continued use or expansion of natural gas as an energy source?" and the classification was recorded. See Table 2.2 for exemplar classifications. This was a subjective process, prone to misinterpretation and personal bias, so a second reviewer evaluated 200 tweets from the subset. Between the original reviewer and the second reviewer, there was an 88% agreement overall in the classification of tweets.

A classification challenge for "natural gas" tweets is their strong bias toward neutral sentiment (see Figure 2.3). In the manually classified subset, only 6.4% of tweets were scored with a positive sentiment and 7.9% of tweets were given a negative sentiment, with the remaining 85.7% of tweets having a neutral sentiment. Non-neutral tweets will have less representation in the training data and will thus be more challenging to classify. This creates an issue because it is these same non-neutral tweets that are key to visualizing patterns in "natural gas" sentiment spatially. With these biased proportions in the sentiment categories, the assessment of text classification scores will need to pay special attention to the performance of

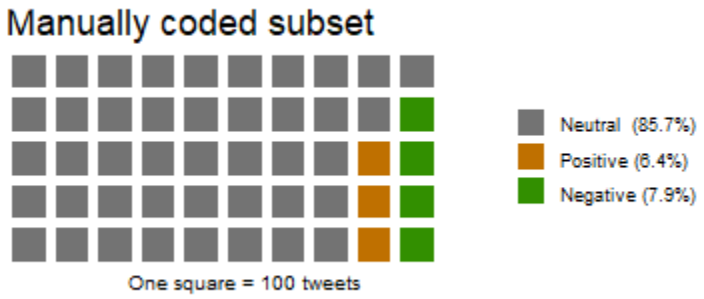

Figure 2.3. Classification results for the manually coded subset of 5,000 tweets.

different classifiers with respect to the positive and negative categories, not merely the overall accuracy scores.

With a set of tweets manually classified, the text in the tweets could then be run through an automated sentiment classification process. Prior to any classification, the text in tweets requires several preprocessing steps that improve sentiment analysis performance. For all analyses, Twitter handles and URLs were removed due to their irrelevance in text classification.

Table 2.2. Example tweets with their manually coded sentiment scores.

Text	**Sentiment**
In the US, renewable natural gas used as a vehicle fuel has displaced over seven million tons of carbon dioxide equivalent (CO2e) over the past five years. Learn more about the benefits of #RNG:	positive
Go green with natural gas and keep some green in your wallet! Triple the efficiency, savings and benefit to the environment.	positive
Fracked natural gas has resulted in the US decreasing its CO2 emissions levels drastically since 2007. Any sane energy policy starts with not outlawing the practice.	positive
A cursory review of the chilling documentary 'Gas Land' tends to prove that Natural Gas extraction in the United States on a large scale is simply UNTENABLE.	negative
Tell USA Governors: Stop Touting Gas as a "Bridge Fuel" and Reject All New Natural Gas Infrastructure	negative
Flaring of natural gas increased by 11% to peak globally this year. A wasteful and harmful practice. Did we consider this increase in our #ClimateChange Models #Environment #ClimateChangeIsNow #ClimateAction #Economy	negative
This Northstar Village residence features a king bed, kitchenette, washer/dryer and a natural stone gas fireplace.	neutral
Asian LNG spot prices remain steady	neutral
Energy officials within the Trump administration referred to natural gas exported by U.S. energy companies as "freedom gas" and "molecules of U.S. freedom" in official statements	neutral

Words shorter than three letters were also removed, as they generally do not provide useful sentiment-related information. For the machine learning analyses, all non-letter characters

including numbers and punctuation were removed, and the remaining words were converted to lower case and split into individual strings. By the end of the preprocessing phase, each tweet was reduced to a list of its individual words and hashtags, the presence or absence of which were used to train the machine learning classifiers along with the sentiment scores assigned from the manual sentiment analysis.

The first sentiment analysis method applied to the data was VADER, an unsupervised lexicon-based technique designed to perform well on social media data such as tweets divided into positive, negative, and neutral categories (Hutto & Gilbert, 2014). Lexicon-based methods, such as VADER, are the most common approach to sentiment analysis (Liu, 2012). The VADER algorithm utilizes a dictionary of words and phrases (with their positivity/negativity attached) along with a set of rules to evaluate the words in each tweet along with emoticons and contextual clues such as negating words and punctuation to return scores for each tweet. Individual scores are given for the positive, negative, and neutral categories for each tweet as well as a compound score ranging from -1 to 1, with lower values relating to negative sentiment and higher values relating to positive sentiment. With different cutoff values in the compound score, the tweets can be classified into one of the three categories. The VADER sentiment analysis was performed using the “vader” module included in the NLTK python library (Loper & Bird, 2002).

Unfortunately, despite VADER seeming well-positioned to perform well with the “natural gas” tweets, given that they are social media data, its accuracy scores were very low across the positive, negative, and neutral categories. Depending on the threshold set for the cutoff between the categories, accuracy scores would range from approximately 10% to 40% for the positive and negative categories and 30% to 80% for the neutral category, with improvement in the former coming at the expense of the latter. With a specific interest in classifying positive

and negative tweets, these accuracy levels were deemed unacceptable, and the VADER results were left unused. Lexicon-based methods designed for general use, even when targeted explicitly at the type of short form text under analysis, are likely better at detecting the mood of a piece of text than they are at more nuanced tasks such as assessing the sentiment of political views. For a lexicon-based method to successfully classify any single political topic, a new dictionary of lexical features would most likely need to be created that incorporates the language, slogans, irony, and other parts of speech relevant to the topic.

With an unsuccessful attempt at using a lexicon-based sentiment classifier, the focus shifted to supervised machine learning approaches. Three different machine learning techniques were applied to the tweets: naïve Bayes, Support Vector Machine (SVM), and logistic regression (also called maximum entropy or MaxEnt), as shown in Table 2.3. Naïve Bayes and SVM are two of the most commonly used classification algorithms in sentiment analysis, with different researchers finding one or the other to be more accurate when applied to Twitter data (Kolchyna, Souza, Treleaven, & Aste, 2015; Pak & Paroubek, 2010). Logistic regression is used less

Table 2.3. The seven machine learning algorithms applied to the tweet text for classification into positive, negative, and neutral categories.

Approach	**Algorithm**
Naive Bayes	Gaussian (GNB)
	Multinomial (MNB)
	Complement (CNB)
Support Vector Machine	C-Support (C-SVM)
	Linear (L-SVM)
	Nu-Support (N-SVM)
Logistic Regression	Maximum Entropy (MaxEnt)

frequently for the classification of Twitter data, though it is not uncommon (Gautam & Yadav, 2014; González-Ibáñez, Muresan, & Wacholder, 2011). The variants of naïve Bayes classifiers

used took the forms of Gaussian (GNB), Multinomial (MNB), and Complement (CNB) algorithms, of which the latter two are best suited for text classification. Three Support Vector Machine algorithms were also used, with these being C-Support (C-SVM), Linear (L-SVM), and Nu-Support (N-SVM) classifiers. All machine learning classifiers were implemented using the scikit-learn library (Pedregosa et al., 2011).

Hyperparameters for each classifier were optimized on the data subset using a grid search method. Every combination of a range of values for each parameter was entered into the classifiers to determine which combination returned the highest accuracy scores across a set of train/test data splits. The highest performing parameters were retained. See Table 2.4 for an overview of the resulting performance of each classifier, which includes the overall accuracy as well as the *recall* and *precision* for each sentiment category. It is crucial when analyzing the performance of classifiers to look beyond overall accuracy because accuracy alone obfuscates critical information about classifier performance. Recall indicates the percentage of tweets in a particular category that were correctly classified, ignoring false positives. For example, if a dataset has 500 true negative tweets, and a classifier correctly assigns a negative score to 250 of these, its recall score would be 50%. Precision indicates the percentage of classified tweets of a specific category that are correct. To continue the previous example, if 300 total tweets were assigned a negative score by the classifier, and the same 250 were correct, the precision score would be 83%. With the primary goal of the sentiment analysis being the confident categorization of positive and negative tweets, close attention was paid to the recall and precision of non-neutral tweets when optimizing the hyperparameters.

Table 2.4. Mean performance results for all classifiers from 50 random samples of training and testing data using a 90/10 train/test split. The ± symbol denotes the size of two standard deviations around each mean value.

		Positive		Negative		Neutral	
	Accuracy	Recall	Precision	Recall	Precision	Recall	Precision
GNB	**23.09** ± 0.41	**87.79** ± 2.72	**8.27** ± 0.14	**89.78** ± 2.96	**18.34** ± 1.01	**15.62** ± 0.44	**98.19** ± 0.23
CNB	**83.13** ± 0.39	**50.24** ± 1.26	**26.53** ± 1.77	**62.88** ± 2.38	**41.75** ± 1.87	**92.61** ± 0.56	**88.91** ± 0.22
MNB	**87.08** ± 0.12	**41.56** ± 1.59	**69.23** ± 3.85	**50.68** ± 1.69	**69.77** ± 3.26	**98.85** ± 0.15	**87.73** ± 0.08
C-SVM	**87.45** ± 0.10	**41.61** ± 1.28	**78.23** ± 3.97	**52.30** ± 2.16	**79.32** ± 3.06	**99.25** ± 0.11	**87.87** ± 0.11
L-SVM	**87.41** ± 0.12	**40.99** ± 1.39	**70.99** ± 4.11	**53.12** ± 2.12	**73.70** ± 2.66	**98.97** ± 0.13	**87.97** ± 0.09
N-SVM	**87.47** ± 0.11	**41.23** ± 1.35	**82.63** ± 3.72	**51.55** ± 2.22	**80.37** ± 2.48	**99.36** ± 0.10	**87.80** ± 0.11
Maxent	**87.48** ± 0.07	**31.41** ± 1.25	**90.04** ± 2.84	**40.93** ± 1.60	**86.54** ± 1.59	**99.74** ± 0.04	**87.43** ± 0.07
Ensemble	**87.47** ± 0.11	**40.10** ± 1.37	**78.65** ± 3.71	**51.72** ± 1.83	**83.12** ± 2.81	**99.20** ± 0.10	**87.84** ± 0.09

This paper will now run through the performance and some of the hyperparameters of the classifiers evaluated to find the highest performing option. The numbers reported are the cross-validation results of 50 random sets of train/test splits with 90% of tweets used as training. The GNB classifier performed the poorest of any classifier used. It classified far too many tweets as positive or negative, resulting in low accuracy scores overall and particularly low precision for the positive and negative categories. Adjusting the parameters did little to redeem the overall performance of this classifier, which had an accuracy score of 23% on the subset of tweets. The results of this classifier were retained because they may provide an interesting counterfactual, showing the spatial effects of over-assigning polarized sentiments to tweets.

The CNB classifier was expected to perform highest among the naïve Bayes classifiers because it was written specifically to deal with skewed data classes such as those present in the "natural gas" tweets. Additionally, this classifier was designed to excel in classifying text data (Rennie, Shih, Teevan, & Karger, 2003). With an alpha of 0.5 and using two normalizations, the CNB classifier obtained an overall accuracy of 83%, lower than most other classifiers.

Interestingly, the recall for positive and negative tweets was relatively good, but at the cost of lower precision.

Best performing among the naïve Bayes classifiers was the MNB classifier. Like CNB, it is commonly used for text classification tasks (Frank & Bouckaert, 2006). With an alpha of 0.6, its overall accuracy was quite good (87%) and its precision for positive and negative tweets was about 70%, far better than the other naïve Bayes algorithms used. At this point, it appears the MNB classifier may be best suited to represent the naïve Bayes family of machine learning algorithms in the spatial analysis portion of the research.

The range of scores among the SVM classifiers used is far smaller than what was observed between the different naïve Bayes classifiers. The overall accuracy, for example, varied by a fraction of a percent, with all returning a mean accuracy score of about 87.5%. The higher performance was unexpected as previous research indicates that naïve Bayes algorithms may be superior at classifying short texts (S. Wang & Manning, 2012). The implementation of these classifiers is based on the LIBSVM implementation of SVM classification tasks (Chang & Lin, 2011). The C-SVM classifier was implemented with C set to 10,000, an RBF kernel, and balanced class weights. L-SVM was implemented with C set to 1, an L2 penalization norm, and a squared hinge loss function. The N-SVM classifier was applied with Nu set to 0.1, a linear kernel, and equal class weights. The N-SVM classifier performed best, showing very similar accuracy and recall when compared to the C-SVM, but with slightly improved precision. With precision around 80% for positive and negative tweets, relatively high confidence can be placed in the N-SVM classification of non-neutral tweets, and it will likely be used to represent the SVM genre of classifiers going forward.

The logistic regression classifier (MaxEnt) returned some of the most favorable scores in terms of its overall accuracy (87.5%) and precision (90% for positive tweets and 86.5% for negative tweets). MaxEnt is a discriminative classifier using a logistic function that can be used to classify text based on its word features. It is commonly used for text classification, including sentiment analysis (Jurafsky & Martin, 2019). The classifier was run with C set to 1, an L-BFGS solver, and an L2 penalization norm. Overall, the MaxEnt and SVM classifiers provided the best classification performance for the data subset.

To avoid pitfalls that may be present in any of the classifiers individually, one additional classifier was developed as an ensemble of the three different types of machine learning algorithms applied. Many methods exist for combining the output of different classifiers into an ensemble classification, with voting being the easiest to understand and most straightforward to implement (Kotsiantis, Zaharakis, & Pintelas, 2006). For this research, the three learning types, naïve Bayes, SVM, and logistic regression, were given one vote each, with the ensemble classifier being assigned whatever the majority vote was, with three-way ties being assigned a neutral score. The score for each learning type was simply a majority vote of the classifiers within each type. Figure 2.4 visualizes this voting process.

For the data subset, the ensemble model performed well, roughly similar to the N-SVM model. Its real use, however, will come into play when the ensemble classifier is applied to the full dataset of tweets rather than the subset, where there is a higher likelihood of unforeseen flaws emerging as the classifiers extrapolate their learning into new territory. The associations between each classifier were quantified by calculating a Pearson's correlation coefficient on the classification results, ranked -1 for negative, 0 for neutral, and 1 for positive, between all pairs of

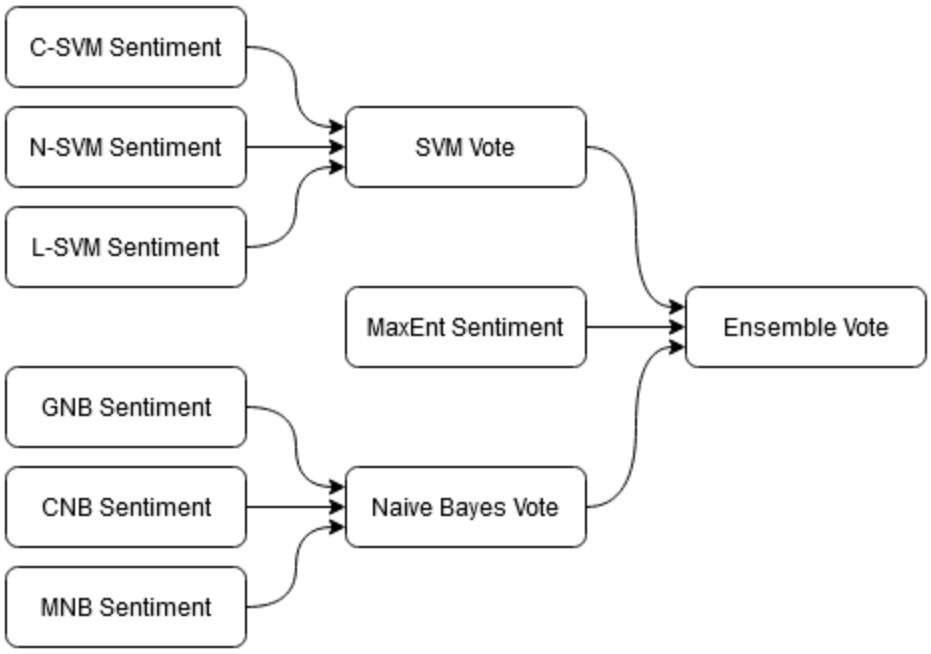

Figure 2.4. The voting process for creating the ensemble classifier. Each arrow represents a vote, where the majority decides the next "vote" classifier.

classifiers with a Bonferroni correction. By examining the correlations between each pair of classifiers, it is possible to approximate how representative each classifier is of the other classifiers. Unsurprisingly, the Ensemble classifier showed the highest average correlations, indicating that it is successfully representing the full range of classifiers. These findings will be discussed further in the results section.

Although these sentiment analysis methods focused on machine learning approaches, it is important to remember that other approaches exist. A lexicon-based approach could potentially perform well if a new dictionary of lexical features was created to match the sentiment of words used in the discourse surrounding natural gas. While this would be a laborious process, manually classifying the training subset of 5000 tweets was also a time-consuming process. Regardless of

the method applied, sentiment analysis is a hands-on task that is likely to consume a plurality of the working hours involved in preparing data for sentiment mapping.

Spatial Analysis

With sentiment analysis and location analysis complete, the tweets are now prepared for the final steps of processing. The spatial analysis presented here relies on point pattern analysis techniques and was carried out in R with the use of several spatial packages: "rgdal" (Roger Bivand et al., 2015), "spatstat" (Baddeley & Turner, 2005), "splancs" (R. Bivand et al., 2017) "sp" (Pebesma & Bivand, 2013), and "raster" (Hijmans et al., 2015). All spatial processing was performed in the USA Contiguous Albers Equal Area Conic projection, using only the tweets with an accuracy score returned from the geocoding equal to 1. In most of the analyses, the neutral tweets were used to represent a background rate of "natural gas" tweets. As was alluded to at the end of the location analysis section, this section will begin with a discussion of updating the coordinates of state-level tweets.

Localizing state-level tweets is a multi-step process, with the idea being to use tweets with more local accuracy levels to create a probabilistic raster to reassign the coordinates of state-level points. First, a kernel density surface was created for all tweets assigned with a neutral sentiment in order to create a surface of the background rate of "natural gas" tweets. Then, using a US shapefile, this surface was clipped into 49 rasters corresponding with the 48 contiguous states plus Washington, DC. Finally, using the name of the state returned with each state-level tweet, the coordinates were randomly reassigned to the center of a grid cell of the corresponding state raster probabilistically based on the value in each cell, plus or minus half the cell width and height in the x and y directions, respectively (see Figure 2.5). An assumption behind this process is that negative and positive tweets cluster in the same locations as neutral tweets. If the data

show strong in-state variation in sentiment, or if this assumption is otherwise invalid, the state-level tweets should be removed from the point-level spatial analysis.

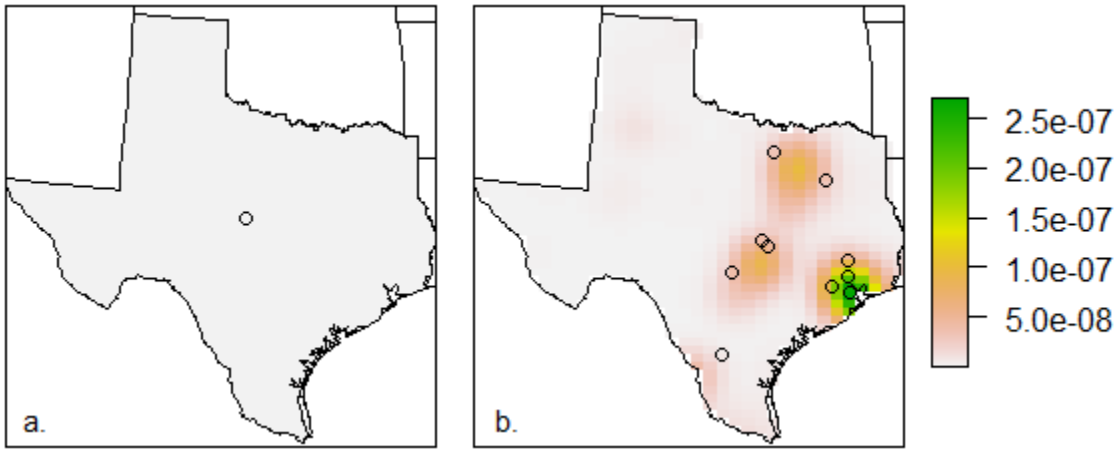

Figure 2.5. An example of reassigning coordinates from state-level accuracy to more local coordinates based on the background rate of neutral tweets. Here, 10 state-level points returned from Geocodio located at (a) the centroid of Texas are moved to (b) locations placed probabilistically on a kernel density surface derived from neutral tweets.

This process has the potential to introduce considerable error into the process at a sub-state regional level that affects the overall interpretation in the final results. With these "natural gas" data, the state-level tweets comprise approximately 1/6th of the total tweets and removing them entirely from the dataset did not cause significant changes to the results. It was determined that, ultimately, removing the state-level tweets may be the safest choice, especially considering that some of the spatial statistics described in this section are sensitive to the precise locations of points in a point pattern. That said, when working with a dataset with less observations, or if working in a different spatial context (such as a global study or a region of the US with smaller states), it is worthwhile to consider this or similar processes for localizing state-level tweets.

With the decision to remove the state-level data resolved, the data were prepared for basic point pattern analysis. First, a variety of global cluster detection methods were applied to

the subset and full dataset to characterize the overall spatial structure of the data. Ripley's K and L statistics (Ripley, 1979) were applied to the positive, negative, and neutral point patterns to evaluate their second-order structure individually. Although the point patterns were clustered, which was not surprising, confidence intervals would be necessary to ascribe significance to any differences between the patterns. Simulated confidence envelopes were utilized to determine if any single sentiment category was more spatially clustered than another (19 simulations). The tests were also completed on the full dataset with and without state tweets to determine if there was any effect by the relocated state tweets on the second-order spatial properties of the point pattern.

The next test, which is the first of the "sentiment mapping" visualizations, is a binomial spatial scan statistic. This was used to identify clustering of positive and negative tweets relative to the background rate of neutral tweets. The spatial scan computes the likelihood ratio of observing the clustering of one point pattern in a given radius to the clustering of the background point pattern, returning statistically significant locations where more clustering is present than what is expected (Kulldorff, 1997). A radius of 150 km was selected for these data as a representation of the regional scale. With smaller radii, too many local clusters are generated to allow for meaningful interpretation of the resulting map, whereas larger radii simply mirrored the results at the 150 km scale. A total of 99 simulations were run to obtain a confidence level of $p < .01$. The spatial scan statistic was repeated with the sentiment results of different classifiers, and over varying time slices, in order to glean a well-rounded picture of the true clustering underlying the point patterns.

Another check performed to develop confidence in the final results was to plot the spatial distribution of misclassifications resulting from the different sentiment analysis techniques.

Errors were saved from the cross-validation runs performed in the sentiment analysis of the data subset as points. These points were then run as a point pattern into the spatial scan statistic described previously. The concept behind this step is that the word choice and issues discussed around the terms “natural gas” may vary regionally, causing some areas to be more prone to sentiment misclassification. Areas that the scan statistic highlights as containing more clustering of error than expected should be interpreted with lowered confidence. As seen in Figure 2.6, the primary area of concern is in Central and Eastern Pennsylvania, the region where three different classes of machine learning classifiers had the highest proportion of error. Also registering on each plot were regions of California and New England, though less so than Pennsylvania. Results in these areas – particularly non-definitive results – will need to be interpreted with more caution.

Finally, in order to visualize the positive and negative sentiment spatially, kernel density difference maps were created. This was done by first creating kernel density maps for the positive, negative, and neutral point patterns using a radius of 50 km. The negative kernel density raster was then subtracted from the positive kernel density raster to create the finished kernel density difference maps. Like the spatial scan statistic, these steps were repeated over monthly time slices to create a visualization of the changes in the polarized sentiment over time. In a similar vein, one final temporal analysis performed was the aspatial density of positive, negative, and neutral tweets over time.

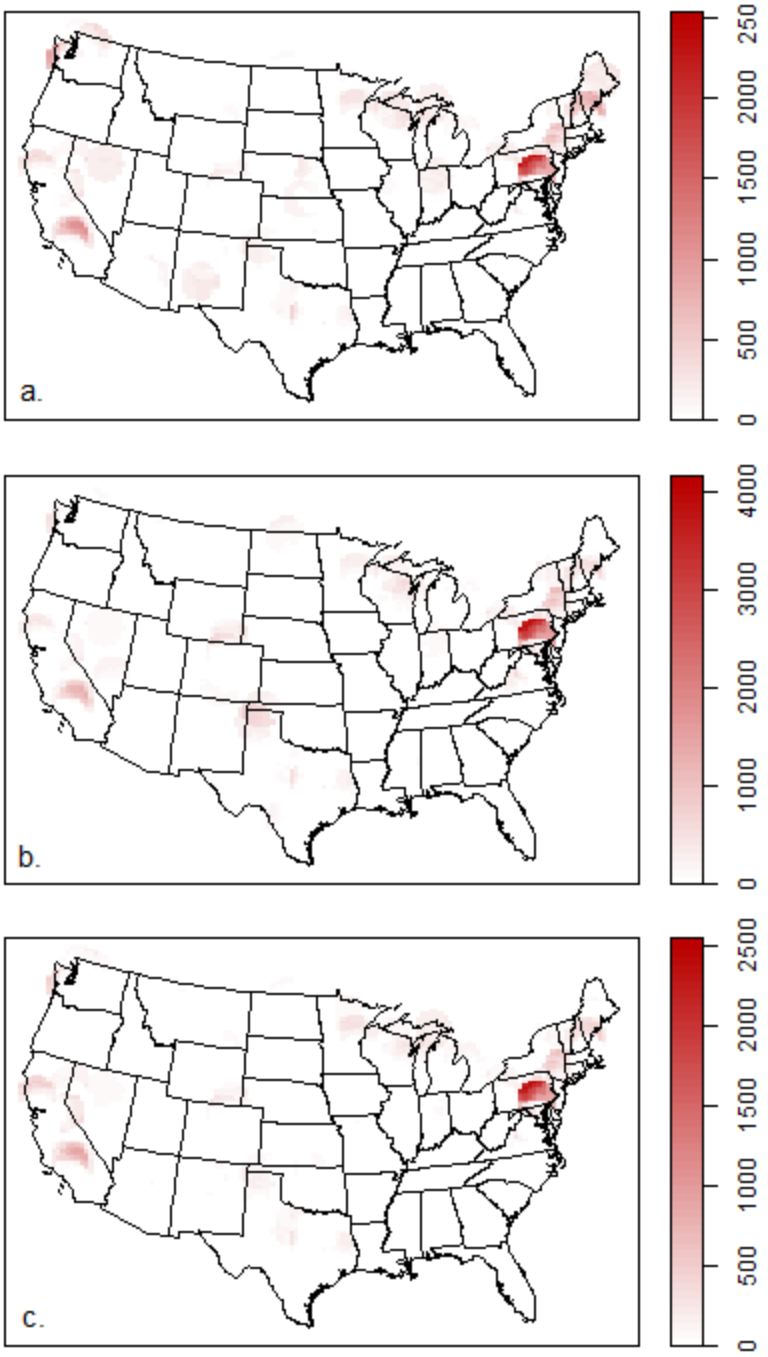

Figure 2.6. Spatial scan statistics comparing the clustering of errors from different classifiers to the background rate of neutral tweets. The errors pictured are from three machine learning classifiers, (a) MaxEnt, (b) MNB, and (c) N-SVM. Errors were obtained from 50 cross-validation runs of the classifiers using different sets of training and testing data. Plots were generated from 99 simulations of the spatial scan statistic with a radius of 150 km, $p < .01$.

3. RESULTS

Tweets in Space

Following the data processing, a total of 121,026 “natural gas” tweets across the contiguous United States were prepared for sentiment mapping. The text in each tweet had been used to assign each to a positive, negative, or neutral sentiment category, and only the tweets with the most accurate locations geocoded from their location information were retained. The results presented here attempt to visualize these data using a plurality of different techniques, a necessity for providing a complete picture of the results and also for establishing their reliability. Decisions made throughout the methodology introduced many opportunities for subjectivity, making a clear and straightforward presentation of the full range of results an essential priority.

Which areas of the US are responsible for the “natural gas” conversation on Twitter? As might be expected, the tweets cluster in urban population centers and follow the general east-to-west population gradient present across the States. Figure 3.1 below depicts the locations of tweets as a kernel density map, with accompanying kernel density plots of the latitude and longitude of each tweet. In the Western US, Seattle, Portland, the San Francisco Bay Area, Los Angeles, and Phoenix are visible. Across the Central US, Denver, Milwaukee, Detroit, and Indianapolis stand out, as do several cities in Texas, including Dallas and Fort Worth, Houston, and Austin. In the Eastern US, the conversation is concentrated primarily in the northeast. Boston, Pittsburgh, a corridor from New Jersey through New York City, Washington, D.C., and Atlanta stand out.

Two of these locations stand out as disproportionately dense. Although its presence is not as unexpected, the density of “natural gas” tweets in Pittsburgh, Pennsylvania seems

overabundant given the size of the city. Further, the highest density of "natural gas" tweets is located in Washington, D.C., and although the area is highly populated, this density still overrepresents the population. There is a good explanation for both of these high densities of tweets. In the first case, Pennsylvania is the second-largest producer of natural gas in the US, just behind Texas, the largest producer. Pittsburgh is located in the western side of the state on the Marcellus Shale, the largest natural gas field in the country (U.S. Energy Information Adminstration, 2019b). It seems reasonable that such a city would be responsible for an

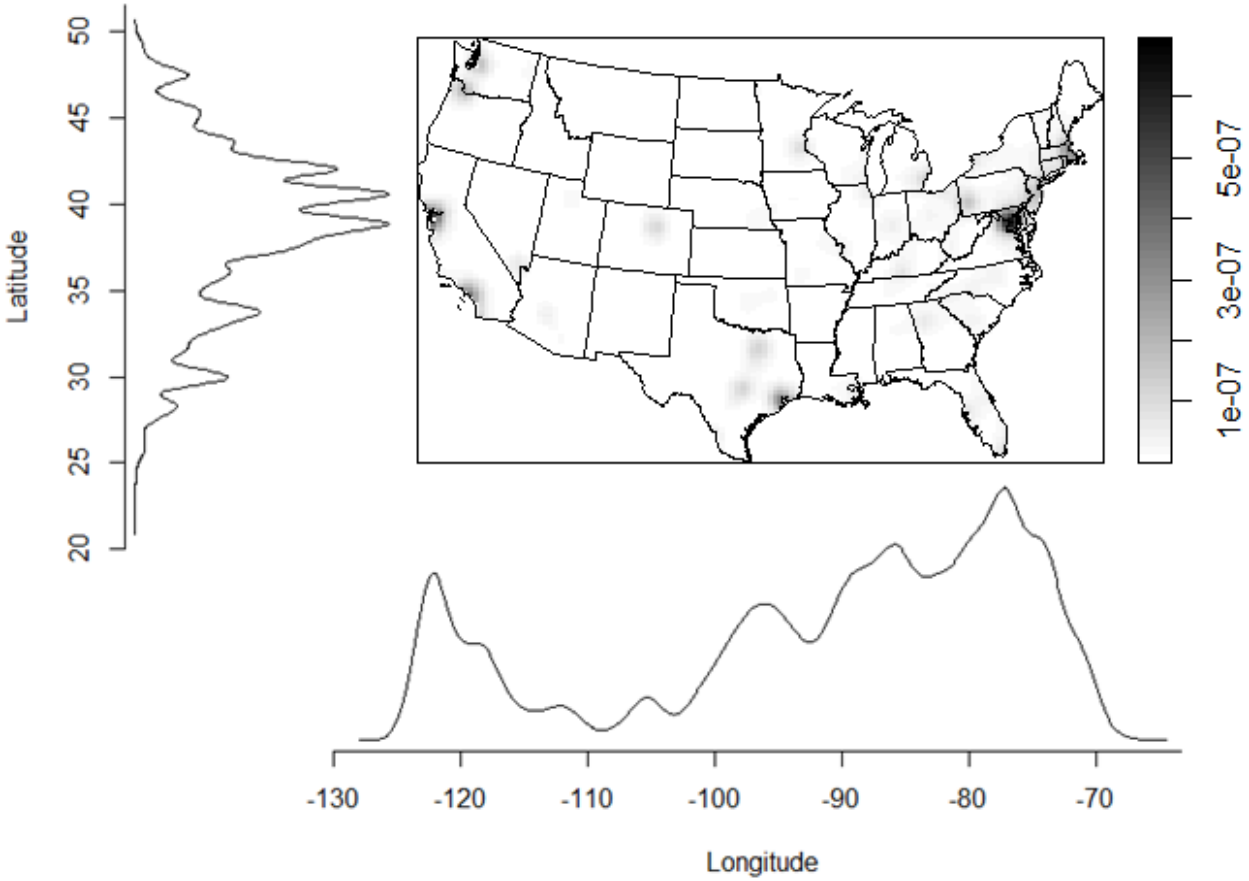

Figure 3.1. Kernel density plots for latitude and longitude alongside a kernel density map of tweet locations across the contiguous United States (CONUS) with an accuracy score equal to 1 (n = 121,026). A 50km bandwidth was used for the kernel density map.

abundance of natural gas posts on Twitter. Washington, D.C, on the other hand, makes sense as a source of postings because of its function as the political center of the US. Natural gas and its extraction are political topics of interest to lobbying groups, policy makers, and the politically minded citizenry of the capitol. Beyond natural gas, it seems likely that any topic of national political interest would see a high density of tweets in D.C.

Classification Comparisons

Before analyzing the sentiment of these tweets toward natural gas spatially, some time will be devoted to a comparison of the results of the different sentiment classifiers. As there is no way to know the actual accuracy of the classifiers once they have been extrapolated to the full dataset, comparison is the only means of assessing their performance. Based on evaluating the results of Table 2.4 in the Methods, the SVM classifiers performed best overall, and the ensemble classifier appeared to be a good middle ground in the performance of all the classifiers. The best naïve Bayes classifier working on the subset appeared to be MNB, while the best SVM classifier was the N-SVM. However, these classifiers were not guaranteed to be the highest performers on the full data set.

In examining Figure 3.2, it appears that MNB and N-SVM had some issues with over-fitting when compared to other classifiers of their algorithm type. While this may have granted some advantage with performance scores in the data subset, positive and negative tweets were under-classified in the full dataset (compared to the 6.4% positive and 7.9% negative tweets classified in the subset). When working with 121,026 tweets, there is enough data that over-fitting is unlikely to result in missing any large spatial clustering of positive or negative sentiment. However, even if some smaller clusters are lost, overfitting is preferable to the

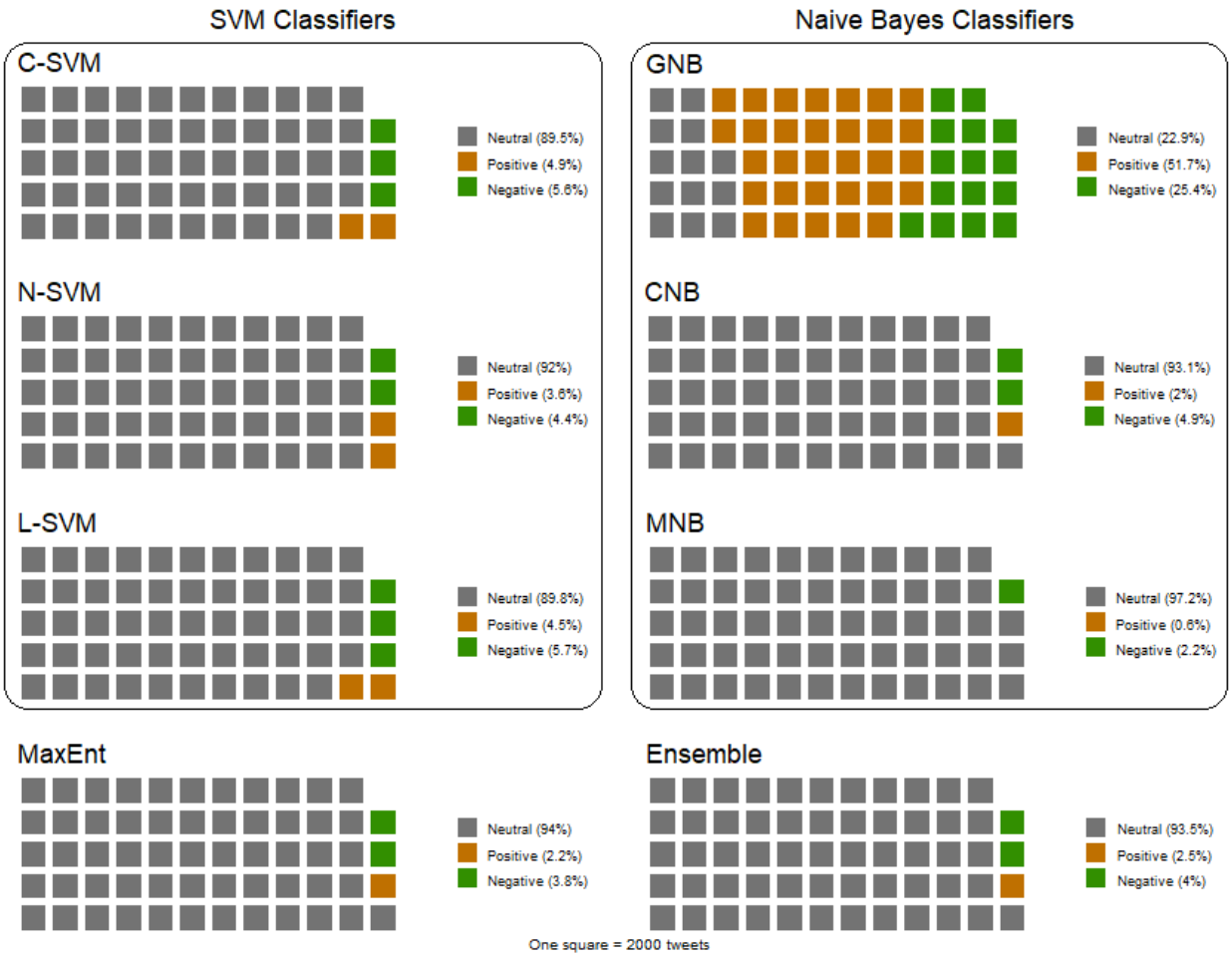

Figure 3.2. Total classification results for all machine learning algorithms by category. Each square represents 2000 tweets.

alternative where classifiers attribute too much sentiment to neutral tweets and generate false sentiment clusters. Overall, the C-SVM classifier attributed non-neutral scores to the highest number of tweets of its class (4.9% positive and 5.6% negative) and MNB the lowest (0.6% positive and 2.2% negative). Again, the ensemble classifier's results appear to land in the middle of the other classifiers, which is a good omen and indicates that the different families of classifiers agree on the sentiment of a large proportion of the tweets.

Aside from overfitting, another issue present with most classifiers is the exaggeration of the ratio of positive to negative tweets. Whereas in the subset there were about 1.2x as many negative tweets as there are positive tweets, the naïve Bayes results have 3.7x, 2.5x, and 0.5x as many negative tweets as positive. There are also 1.7x as many negatives for MaxEnt and 1.5x as many for the ensemble classifier. If the data subset is reflective of the full dataset, the spatial results from these classifiers exaggerate the magnitude of negative sentiment. Again, these results support the idea that the SVM classifiers are the best performers on these data as their ratios were very near 1.2 negative tweets for every positive one. The GNB results continue to be wildly inaccurate (also the only results with more positive tweets than negative tweets), and only continue to be retained as a study of the downstream consequences of overly polarized sentiment classification.

The takeaway from Figure 3.2 is increased confidence in the overall results. Although individual classifiers such as MNB appear to have missed a significant portion of non-neutral tweets, the general profiles of the classifiers look similar to one another, aside from the expected exception of GNB. Overall, the SVM and MaxEnt classifiers appear to have classified a reasonable number of tweets into each category given the expectations set by the recall and precision of Table 2.4. In contrast, the naïve Bayes classifiers may have been overfit to the subset and thus underperformed. The ensemble classifier continues to represent a middle ground.

The associations between the results of the different classifiers is summarized in Table 3.1 below using Pearson's correlation coefficient between each pair of classifiers. Here it is demonstrated that the Ensemble classifier is indeed the most representative of all classifiers with an average correlation of .648. It matched very closely with the MaxEnt classifier in particular, mirroring results seen in previous figures. Among the SVM family of classifiers, C-SVM

correlates most strongly with the others, indicating that it would be a good choice for representing its family. As expected, the naïve Bayes classifiers have weakest associations with one another and with all other classifiers, suggesting that they are categorizing the tweets substantially differently. Considering that this family also had the lowest accuracy scores, the categorizations of the naïve Bayes classifiers are undoubtedly the least reliable, though they will be retained throughout the results as a demonstration of the range of outcomes possible with the use of multiple classifiers.

Table 3.1. Pearson's correlation coefficients between classifier pairs. All correlations were significant at the $p < .05$ level with a Bonferroni correction.

	GNB	**CNB**	**MNB**	**C-SVM**	**N-SVM**	**L-SVM**	**MaxEnt**	**Ensemble**
GNB		0.253	0.199	0.269	0.246	0.262	0.223	0.242
CNB	0.253		0.615	0.411	0.413	0.432	0.431	0.539
MNB	0.199	0.615		0.395	0.428	0.397	0.428	0.504
C-SVM	0.269	0.411	0.395		0.802	0.835	0.699	0.771
N-SVM	0.246	0.413	0.428	0.802		0.781	0.772	0.807
L-SVM	0.262	0.432	0.397	0.835	0.781		0.733	0.783
MaxEnt	0.223	0.431	0.428	0.699	0.772	0.733		0.890
Ensemble	0.242	0.539	0.504	0.771	0.807	0.783	0.890	

Sentiment Cluster Maps

Further evaluation of the classifiers can be performed spatially by comparing the spatial scan statistics for the positive and negative sentiment categories. Figure 3.3, displaying "sentiment cluster maps," groups the machine learning classifiers into the same three groupings as the previous figure. The clusters of positive (left) and negative (right) sentiment are plotted with respect to the background rate of neutral tweets. The MaxEnt and ensemble classifier results are consistent with the previous figure in that they represent a middle ground between the other classifiers. Here, they identify many of the same clusters as the other classifiers but without any

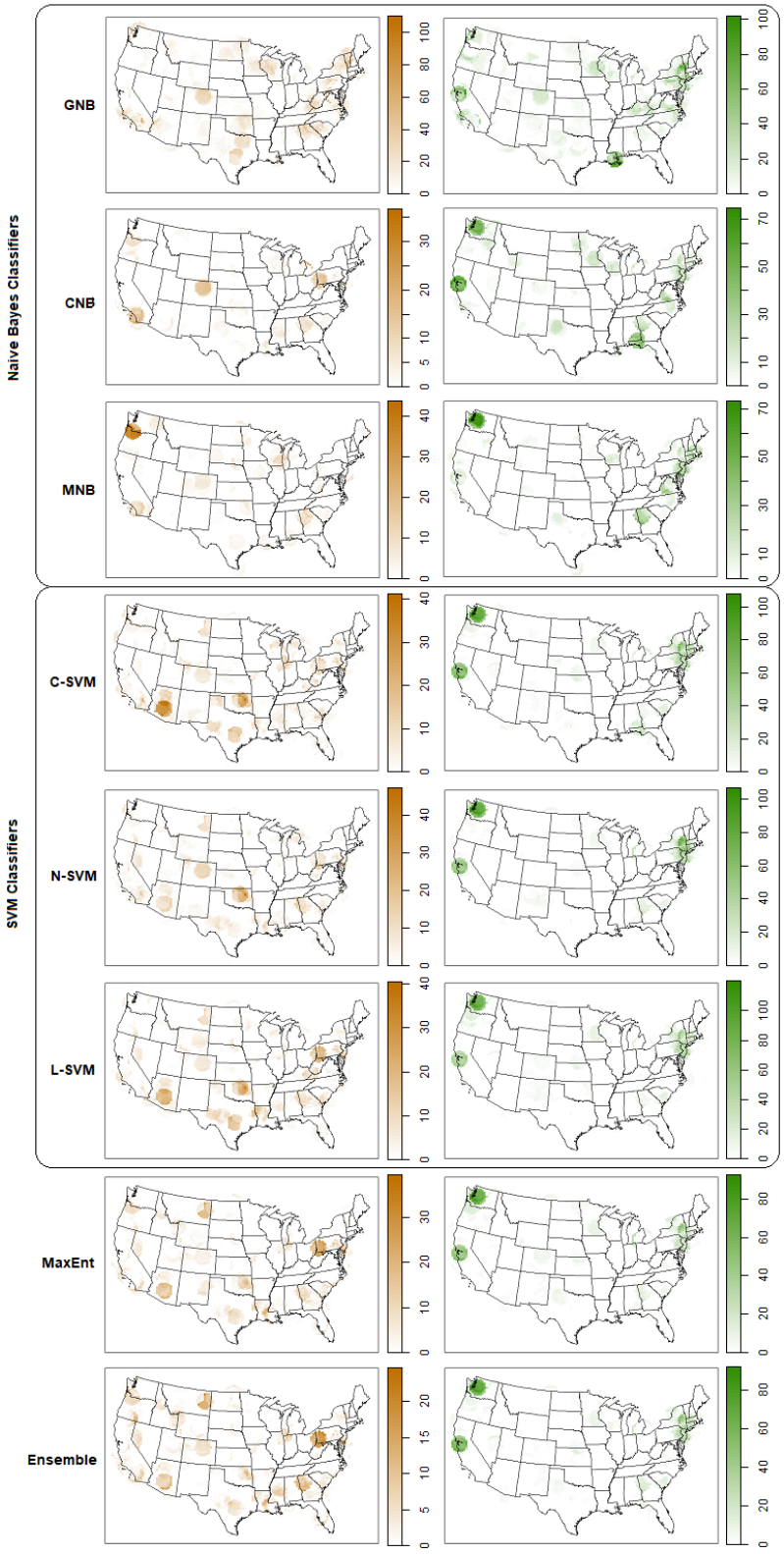

Figure 3.3. "Sentiment cluster maps." Spatial scan statistic results for positive (left) and negative (right) tweets with respect to a background rate of neutral tweets for all classifiers. The spatial scan was completed with a 150 km radius and 99 simulations, $p < .01$.

strong emphasis on individual cluster in the case of positive sentiment, and more distinct clustering in the case of negative sentiment.

Because it was created through a voting process, focusing on the spatial scan results for the ensemble classifier is a way of summarizing the trends across all the classifiers. There are clusters of positive sentiment in Reno, Phoenix, Pittsburgh, and Atlanta, as well as other nonspecific regions of eastern Montana, North Dakota, Louisiana, Alabama, and Oklahoma. Positive sentiment appears to cluster in a broad range of locations across the interior of the US. Examining the maps above the ensemble classifier reveals that these positive cluster locations can vary considerably depending on the classifier used, indicating that the individual locations are less meaningful than this overall trend of broad clustering. Conversely, negative sentiment is consistent and highly clustered in three distinct coastal areas: Seattle, the San Francisco Bay Area, and the New York City area. Note that although the negative sentiment is clustered in highly populated areas that Figure 3.1 revealed are responsible for a large portion of the "natural gas" tweets, these results account for the underlying population because the background rate of neutral tweets is utilized with the spatial scan statistic. This means that these clusters of negative sentiment are clustering to an extent that they are significant even in light of the very high rates of neutral tweeting in these areas.

When comparing the classifiers to one another and previous figures, many of the same trends emerge that were seen in prior comparisons. The SVM family maps similarly for both positive and negative sentiment. For the positive sentiment, the cluster most contested is centered on Phoenix, where C-SVM finds it to be one of the most robust clusters but N-SVM finds it to be one of the weakest. The other differences are mostly minute. The negative sentiment maps are nearly identical, both internally within the SVM family and when compared to the MaxEnt and

ensemble classifiers. Overall, the results of the SVM classifications continue to follow the same pattern of consistency with each additional visualization. Thus far, there have been no indications that the SVM family is producing unreliable results.

The naïve Bayes classifiers disagree with one another most strongly, with GNB once again setting itself apart as the outlier. This is most apparent in the spatial scan of negative sentiment, where the Seattle cluster is nearly absent from the map though it appears in every other negative scan. It also underemphasizes the southern portion of the Bay Area cluster while managing to find a Louisiana cluster than no other classifier identified. The MNB classifier, which Figure 3.2 revealed to be stringent in assigning non-neutral scores, also performed very poorly. Consistent with its underrepresentation of polarized sentiment, the Bay Area cluster is nearly missing entirely from the negative sentiment plot. The positive sentiment scan fares even worse for MNB, bearing little resemblance to the other maps of positive sentiment and also being the only to find a strong positive sentiment cluster around Portland. Of the naïve Bayes classifiers, CNB is the only to compare with the non-NB classifiers relatively well. However, it found weak sentiment clusters on both plots in regions where the other classifiers found nothing. This is somewhat unsurprising, as Table 2.4 reveals that CNB correctly identified many of the non-neutral tweets (recall), but it was also quick to mislabel the non-neutral tweets (precision). Overall, the flaws in the naïve Bayes family of classifiers appear to have had seriously detrimental effects when the classifications are plotted spatially. The results disagree internally and also externally compared to the other classifier families. The spatial scan statistics resulted in significantly decreased confidence in the naïve Bayes results.

The Khat statistic is another means of characterizing the spatial distribution of tweets. Taken individually, the Khat values for a given point pattern reveal the degree to which the

points are clustered at different spatial scales, which only reveals that both the positive and negative sentiment tweets are clustered. It is in taking the difference between the Khat values for positive and negative tweets that more revealing information is extracted. Figure 3.4 depicts the results of subtracting the Khat values of positive tweets from the Khat values of negative tweets.

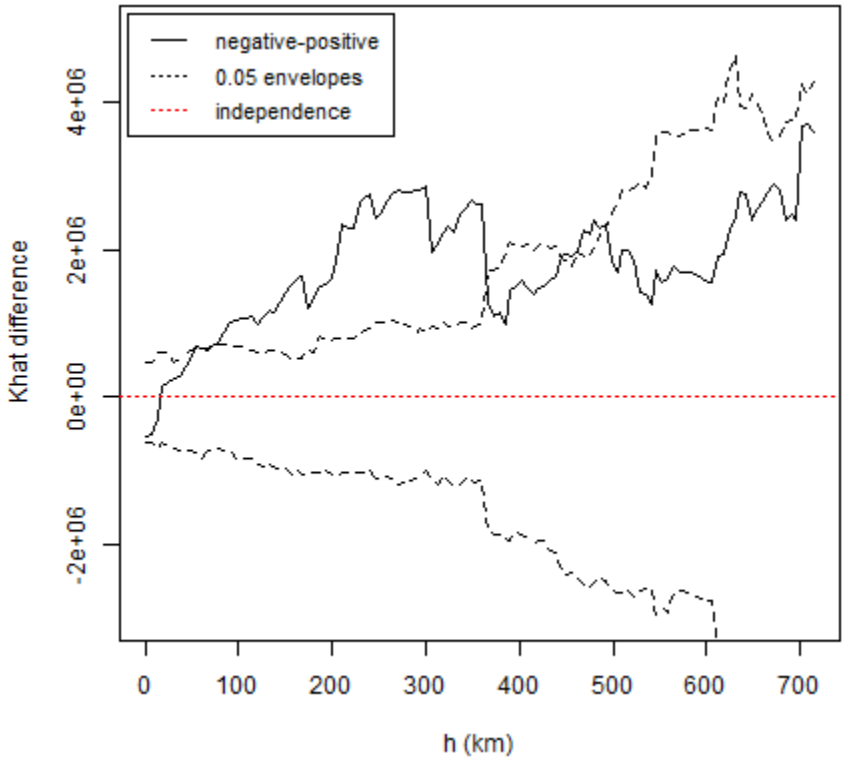

Figure 3.4. Khat difference between positive and negative tweets classified by the ensemble classifier. Envelopes represent p = .05 confidence intervals from 19 simulations. Khat difference values above 0 indicate greater clustering of negative tweets at the distances shown.

Up to a radius of approximately 375 km, the negative tweets are significantly more clustered than the positive tweets, rising well above the p = .05 confidence envelopes. This matches the expectation set by the nature of the clustering in Figure 3.3., where negative tweets were highly clustered in a few major population centers while positive sentiment clusters were more spatially diffuse. The Khat difference plot confirmed the visual interpretation that the negative tweets

were the more clustered pattern using a statistical test rather than merely a visual cue. Further, the test indicates that this more intense clustering holds true out to a radius of about 375 km, beyond which neither category of tweets is significantly more clustered than the other.

Raw Sentiment Maps

With an understanding of the performance of different classifiers and the overall spatial patterns of the positive and negative sentiment categories established, the results will now move on to the final sentiment maps. Depicted in Figure 3.5, the "raw sentiment map" is the difference between the density of positive and negative sentiment across the contiguous Unites States. Areas with a higher raw density of positive tweets appear in orange while locations of higher negative density appear in green. The eight classifiers were reduced to four in this plot to highlight only the classifiers that showed the highest performance of their family throughout the

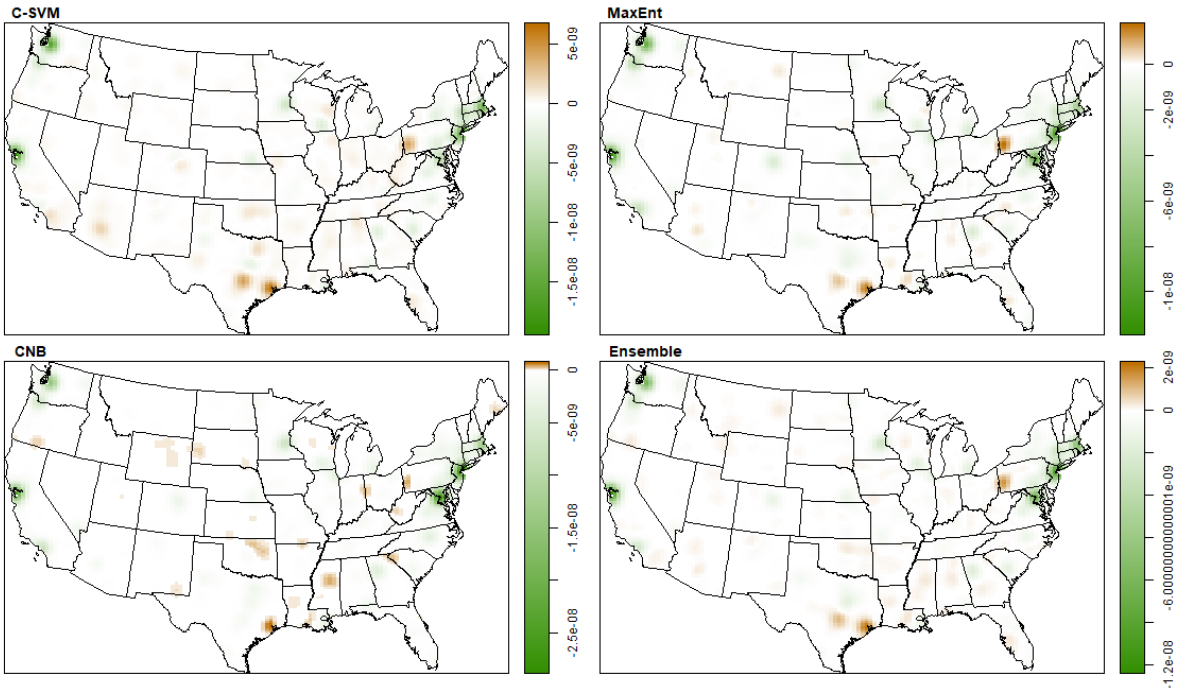

Figure 3.5. "Raw sentiment maps." Kernel density difference maps from four different classifiers computed by subtracting the kernel density map of negative sentiment (green) from that of positive sentiment (orange). Values above zero indicate a higher density of positive natural gas tweets.

different visualizations. The four classifiers included in the plot tell similar but somewhat varying stories, indicating that the kernel density difference map is less sensitive to different classifiers than the spatial scan was. The map that is most dissimilar compared to the others is the CNB plot. This is unsurprising given that the naïve Bayes classifiers consistently performed the most poorly, and with results most different from the other classifiers.

Once again, all classifiers agree on the hubs of negative sentiment that were identified with the spatial scan statistic. Seattle, the Bay Area, and the northeast from New Jersey to Boston have a high density of tweets classified as negative toward natural gas. Added to this list are Washington, D.C., the Twin Cities, Portland, Detroit, and Los Angeles, which appear on the density difference maps for most classifiers. While it may seem unusual that new locations are highlighted with this visualization, recall that these visualizations are independent of the background rate of neutral tweets. In an area such as Washington, D.C. with a high density of neutral tweets, there is a very high threshold to cross before the density of negative tweets becomes statistically significant. However, despite the clustering of negative tweets being insignificant when compared to the background rate, most classifiers agreed that there was more negative sentiment toward natural gas in the D.C. area than positive sentiment.

Areas that register with more considerable positive sentiment toward natural gas are again more dispersed and more variable than the negative sentiment hotspots, similar to what was observed in Figure 3.3. Houston, Austin, Oklahoma, Louisiana, Pittsburgh, and Tampa stand out most prominently and consistently as areas with a higher density of positive natural gas tweets than negative tweets. Looking at the ranges of values, the high-density centers for positive tweets are not as pronounced as the negative density clusters, again harking back to the more

intense clustering of negative tweets. Also, given that more negative tweets exist overall in the dataset, and that the clustering of negative tweets was found to be more significant, it is unsurprising that the most prominent positive sentiment areas do not size up to the magnitude of negative sentiment areas.

Many of the hotspots of positive natural gas sentiment align with natural gas producing regions of the United States. As mentioned, Texas and Pennsylvania are the largest producers of natural gas in the country, and both contain high-density areas of positive sentiment. Louisiana and Oklahoma, which also show favorable natural gas sentiment on the raw sentiment map, are the third and fourth-largest producers (U.S. Energy Information Adminstration, 2019a). The major hotspots of negative sentiment seem concordant with expectations when the sentiment is viewed through a political lens. Seattle, the Bay Area, the Northeast, and Washington, D.C. skew strongly toward the Democratic Party. Given that the Democratic Party favors "green" policies, which generally call for the reduction or complete halt of natural gas production, it is unsurprising to see clusters of negative sentiment in these highly Democratic regions. A similar story applies to negative sentiment observed in Twin Cities, Portland, Detroit, and Los Angeles, which are areas that strongly favor the Democratic Party. Taken together, the current results build an internally consistent representation of "natural gas" sentiment across the contiguous United States that aligns with real-world phenomena, such as political sentiment and regions of natural gas production.

Exaggerated Sentiment Maps

An issue with both the raw sentiment maps and the sentiment cluster maps is that population centers are dominating the results. In one way, this reflects reality; less populated areas contribute far less to the Twitter discourse in terms of the volume of tweets. However, this

leaves a blank in the evaluation of the "natural gas" sentiment, both literally and figuratively. The large swaths of white space on the maps provide no understanding of the sentiment in less populated regions of the country. Although these areas may not have the same level of contribution as more populated areas, there are still many reasons to be interested in the sentiment of these areas.

One way to highlight the sentiment of places with a lower density of tweets is to dampen the magnitude of higher density areas. This can be accomplished by taking the square root of the positive and negative kernel density maps before calculating the difference between them. In doing so, the prevailing sentiment of the classified tweets becomes clear in almost all regions of the contiguous United States (see Figure 3.6). With these "exaggerated sentiment maps," within-state variation in sentiment is now clearly visible. State-by-state, the Ensemble and MaxEnt classifiers show the most agreement in the distribution of sentiment. As expected, the CNB classifier is once again the most different, showing a far greater dispersion of negative sentiment than the other classifiers. C-SVM, which is the most representative of the SVM classifiers used, and may represent the most accurate tweet classifications overall, showed the largest distribution of positive sentiment.

Although these results will not delve into the mapped outcomes state-by-state, it is worthwhile to note that a majority of states show consistent patterns across a majority of the classifiers (excluding CNB). Additionally, many of these patterns make sense with comparison to the locations of natural gas production and the distribution of political viewpoints across the contiguous US. The effect of the relatively poor classification performance by the naïve Bayes classifiers is apparent with this figure – although the negative sentiment results are similar in many locations, the areas of positive sentiment are significantly altered to the extent where

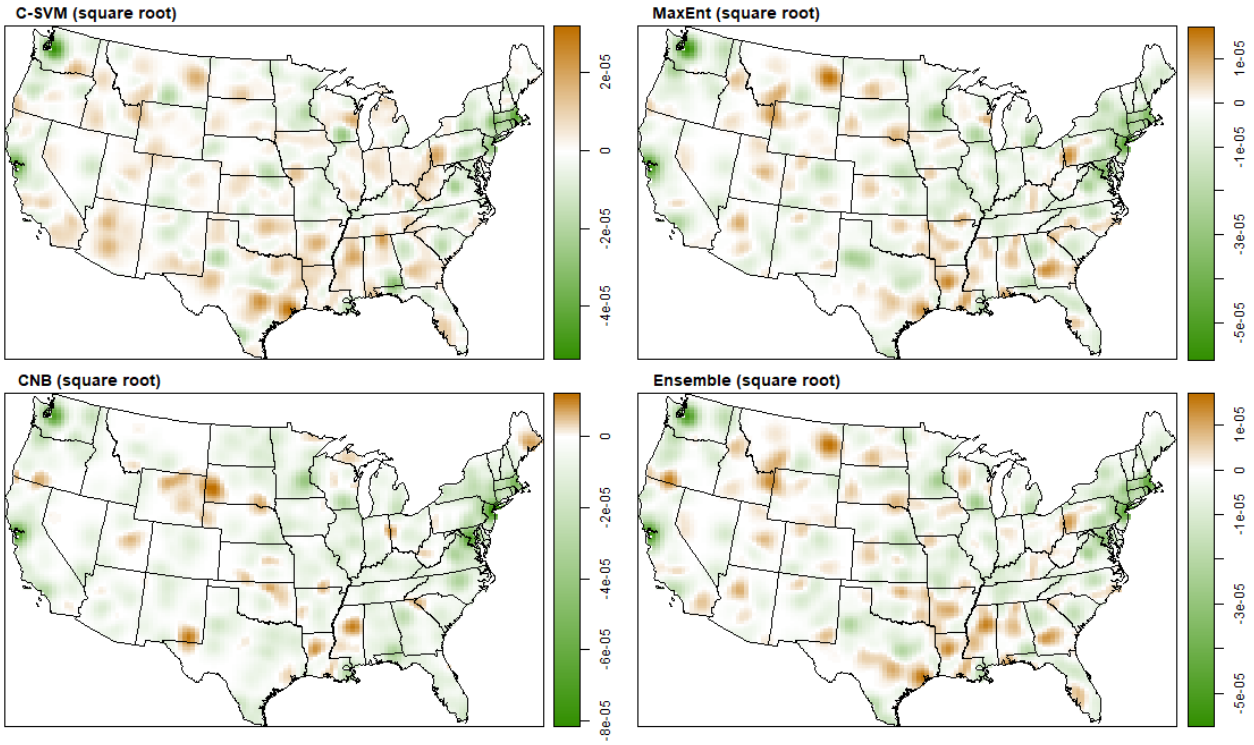

Figure 3.6. "Exaggerated sentiment maps." The difference between the square roots of the density of positively and negatively classified tweets.

state-by-state interpretation of the CNB map would often lead to very different interpretations of the distribution of sentiment. It is a notable example of what could go wrong with the results of sentiment mapping, even with careful hyperparameter optimization and a large set of training data.

One cost of the additional information provided by the exaggerated sentiment maps is that the overall noisiness of the maps increases as well as the number of differences between the results of the four classifiers. Further, any given hotspot of positive or negative sentiment is less reliable, and a sense of the proportion of differences in sentiment intensity is lost. Whatever concerns have developed thus far around the accuracy of the overall sentiment mapping results should be heightened significantly for these plots. Recall also that misclassifications from the

classifiers are not distributed evenly across space. Now would be a good time to reexamine the error maps of Figure 2.6 to recall where the classifiers most consistently made classification mistakes. At this sub-state level of analysis, it is likely best to avoid including the state-level tweets; at best they are adding some useful information along with a lot of noise, and at worst they are biasing the within-state sentiment.

Tweets in Time

The sentiment of the "natural gas" tweets can also be evaluated temporally, both to learn more about the nature of the different sentiments and to evaluate the differences between the classifiers further. Figure 3.7 depicts the density of positive, negative, and neutral tweets over time. Of note is that all classifiers follow a uniform distribution overall, indicating that there was no significant increase or decrease in "natural gas" tweeting over the time period. The two valleys occurring in October are due to two weeks of missed data collection near the beginning and end of the month. Except for CNB, all classifiers agree on three peaks in non-neutral sentiment occurring on June 30th, September 8th, and December 8th. The first and last peaks occur due to high densities of negative sentiment during the periods while the middle peak indicates a high density of both positive and negative sentiment, with positive sentiment rising above negative.

According to the results of these classifiers, it seems that the negative sentiment is not only more highly clustered spatially but also more clustered temporally than the positive sentiment. The C-SVM classifier, which found the lowest ratio of positive to negative tweets among the classifiers plotted, finds these sentiment peaks to be noticeably lower in magnitude. In examining these plots, it is essential to remember that their data are entirely dependent upon the

classifier used, its parameters, and the training data. The spikes observed could be nothing more than artifacts from the classification process.

Many previous studies have linked temporal peaks and valleys in different sentiment categories to news events. For many users, one of Twitter's primary uses is reading and sharing articles and current events. If a topic is trending that brings out significant positive or negative sentiment, this may be reflected in plots such as Figure 3.7. Another contributor to the spikes could be changes in the language used during different periods. If a specific word or phrase is used frequently during a single time window, and this word or phrase is distinctly connected to a particular sentiment category, it is possible that the classifiers were trained to assign non-neutral sentiment to tweets with that text more easily.

For example, the day before the May 30th spike, the US Department of Energy described natural gas as "freedom gas" and "molecules of US freedom" in official statements (Bowden, 2019). This spurred a number of tweets on the subject, the majority of which were of negative or neutral sentiment. What was key in this event's ability to trigger a spike in sentiment is the rarity of words like "molecules" and "freedom" in the natural gas dialogue. If the classifiers learned to associate these words with a significantly higher likelihood of negative sentiment, it would be simple to understand why so many negative tweets were classified because of this event.

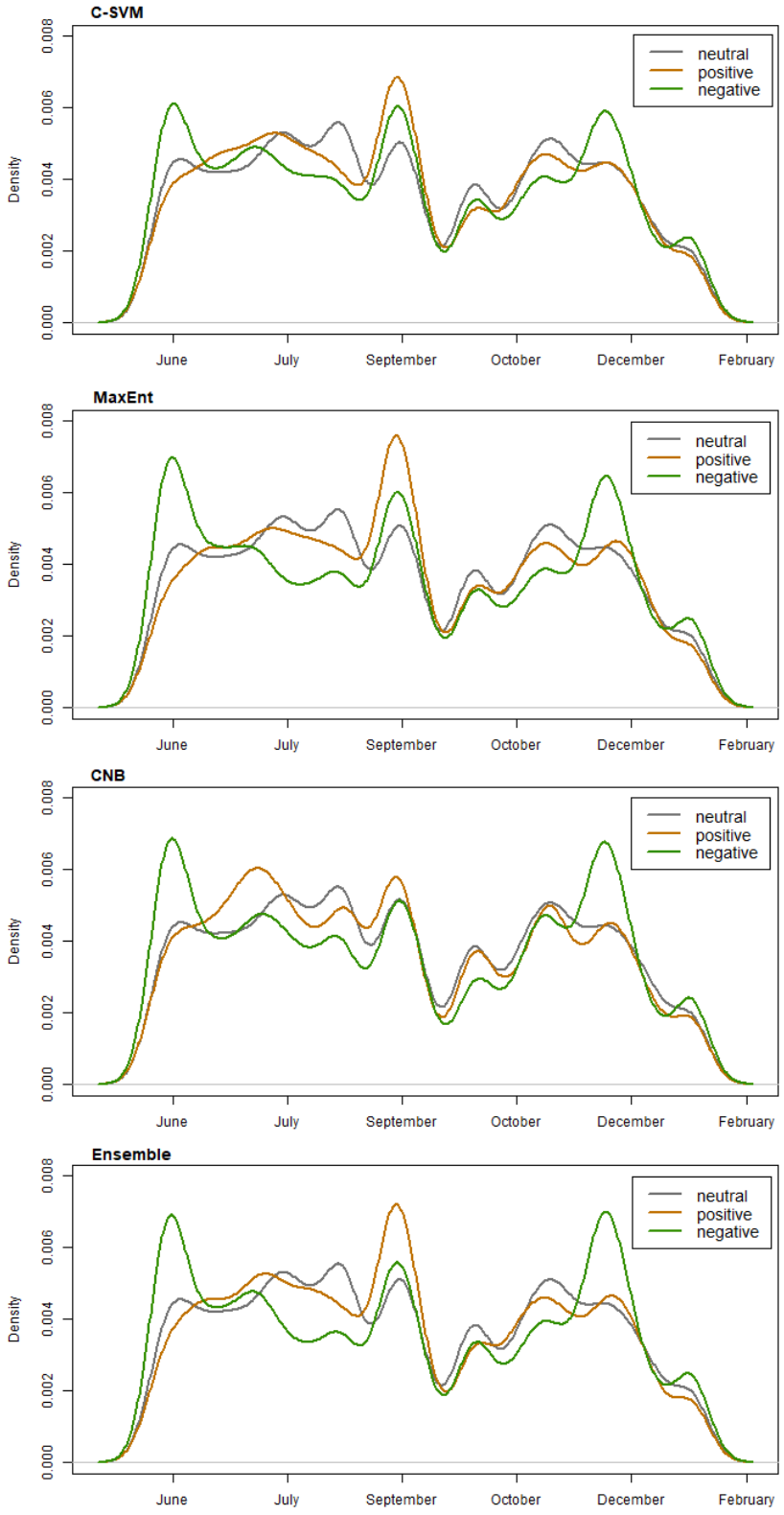

Figure 3.7. Density plot showing the frequency of positive, negative, and neutral tweets over time using four different classifiers. The three most prominent peaks occur on May 30th, September 8th, and December 8th. The bandwidth is set to 7.5 days.

Other temporal analyses are more complex and involve mapping the change in sentiment spatiotemporally. As seen in Figures 3.8 and 3.9 below, which display monthly sentiment cluster maps and monthly raw sentiment maps, this quickly becomes cumbersome as each time slice requires a different map. These data are visualized and interpreted much more easily with interactive maps or gifs with multiple frames. That said, the main takeaway from these plots is still evident from the still plots. There is very high spatial variability in positive and negative sentiment. All visualizations before this section of the results smoothed over this surprisingly high variation. Again, the negative sentiment shows less variability, particularly in Figure 3.9, where the same few coastal hotspots of negative sentiment continue to stand out. Even so, the intensity and locations of negative sentiment hotspots change significantly from month to month. These figures highlight the need for long-term studies on twitter data. This study used an 8-month period and the variability present in these maps and Figure 3.7 suggests that some of the overall spatial patterns in sentiment may be more temporary than previously assumed.


May
June
July
August
September
October
November
December
January

Figure 3.8 (cont'd). Monthly sentiment cluster maps. Created using a spatial scan statistic with data grouped by monthly time slices, using the ensemble classifier. The scan shows elevated levels of positive sentiment (left) and negative sentiment (right) relative to the background rate of neutral sentiment. Each scan was completed with a 150 km radius and 99 simulations, $p < .01$.

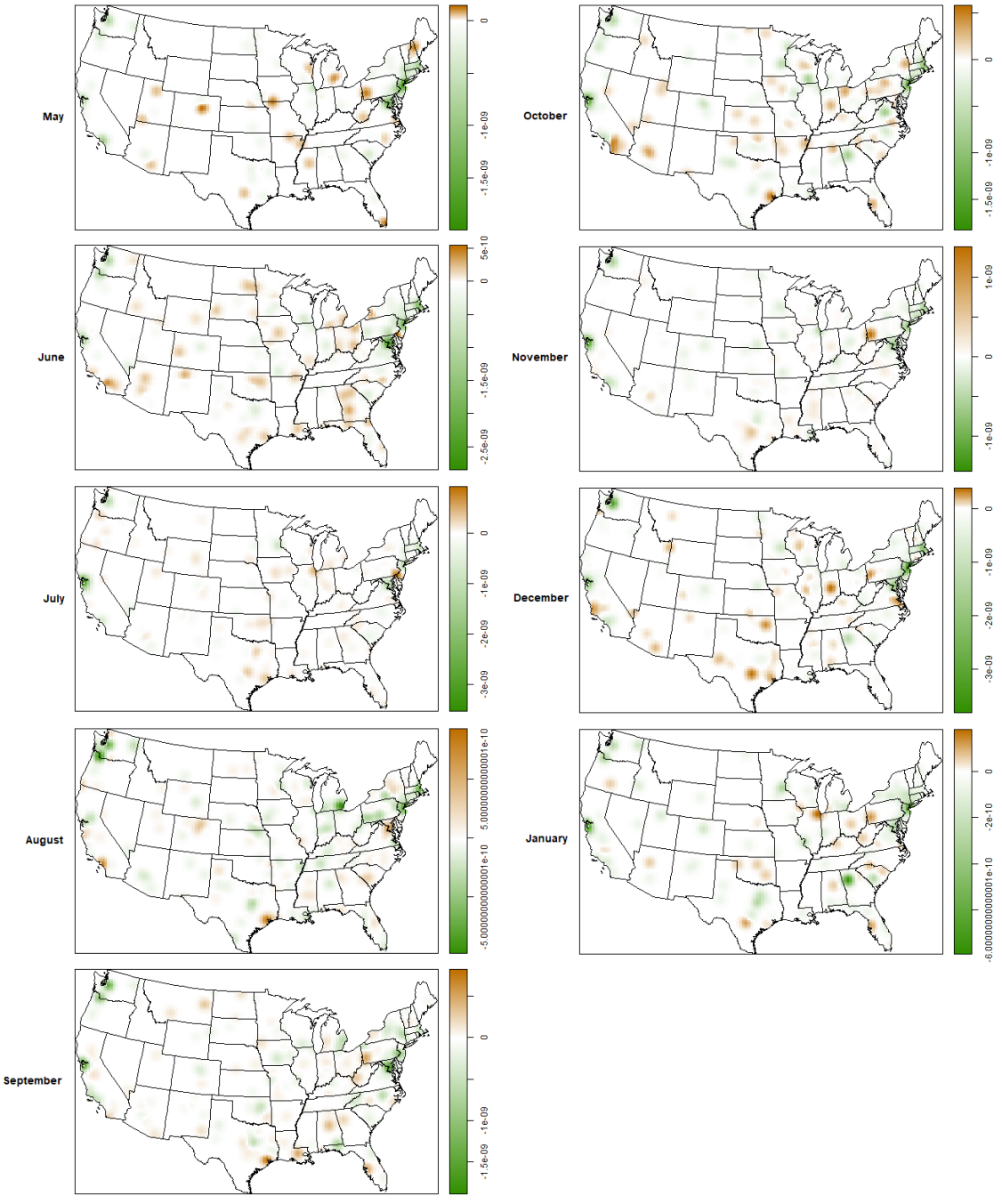

Figure 3.9. Monthly raw sentiment maps. Kernel density difference plots for each month in the data set, using the ensemble classifier. Orange areas have a higher density of positive tweets while green areas have a higher density of negative tweets.

4. CONCLUSION

This paper has provided a demonstration of sentiment mapping, an approach to visualizing the spatial distribution of opinions or sentiment shared on social media. Here, data are pulled from Twitter to analyze sentiment on the platform toward natural gas. The process of sentiment mapping draws from work in two disparate areas: 1) sentiment analysis, a branch of natural language processing focused on identifying the mood or sentiment of the text, and 2) density and distance-based methods of point pattern analysis, which have been applied to a broad range of spatial questions. Although there are many examples to be found of studying sentiment on Twitter or studying the spatiotemporal distribution of tweets, the unique combination of sentiment analysis and point pattern analysis allows for spatiotemporal assessment of Twitter sentiment in ways that have not previously been explored.

In addition to describing the methodology of sentiment mapping, a case study of the natural gas sentiment expressed on Twitter over eight months was examined to evaluate the legitimacy of the results. Although over 300,000 tweets containing the keywords “natural gas” were downloaded, only 121,026 could be geocoded with enough accuracy for use in further spatial analysis. Seven machine learning classifiers were trained on a subset of manually classified tweets to learn the patterns of classifying natural gas tweets into positive, negative, and neutral categories. They were evaluated on their accuracy for this subset using standard training and testing methods, both overall and by individual category. From there, the full dataset was classified, and the sentiment mapping process began.

Three methods of sentiment mapping were applied to the data to examine the spatial patterns of positive and negative sentiment in varying ways. The first, “sentiment cluster maps,”

used a spatial scan statistic to identify spatial clusters of positive and negative sentiment with respect to the background rate of neutral tweets. The second, "raw sentiment maps," depicted the difference between the kernel density maps of positive and negative sentiment. The third, "exaggerated sentiment maps," took the square root of the raw sentiment maps to highlight the sentiment in areas with less tweeting. The maps were created using the classifications output from different machine learning classifiers and were compared to evaluate the range of results resulting from different classifiers.

Overall, the results of the spatial analysis were reasonably consistent between the different classifiers and sentiment mapping techniques applied. The exaggerated sentiment maps, which highlighted local variation, showed the most variability but shared the same big picture trend. Negative sentiment was most densely clustered in three coastal areas: Seattle, Washington, and the northeast near New York City, which are areas that are politically outspoken against natural gas production. Positive sentiment was less densely clustered overall but showed high concentrations in major natural gas producing regions of Texas, Pennsylvania, Louisiana, Oklahoma, and other states.

When considering the sentiment classification results, which were promising and consistent overall, of note is the poor performance of the naïve Bayes classifiers. These results not only run contrary to other research (S. Wang & Manning, 2012) but also highlight the need for more care with the application of sentiment analysis to social media data. Many studies covered in the Introduction relied on only one classifier to perform the sentiment analysis. However, the results here demonstrate that a single classifier, or even a family of classifiers, has the potential to go awry when applied to a large dataset, even with a large sample of training data. It is strongly recommended that multiple classification algorithms are run and compared to

avoid the potential for poor classification and to understand the possible ranges in the spatiotemporal distribution of the data.

Although the sentiment maps told a consistent and cogent story, the spatiotemporal analysis revealed more complexity behind the broad results. There was high temporal variation of both positive and negative sentiment, with two temporal peaks of negative sentiment being responsible for a large portion of the overall negative sentiment. Spatially, the clusters of positive and negative sentiment varied considerably on a month-to-month basis, suggesting that the time over which tweets were collected played an important role in shaping the final results. Interpretation of the results must also consider that the data passes through many filters before the sentiment mapping begins – rather than representing the sentiment of all people in different regions toward a chosen topic, the data can only ever approximate the sentiment of a subset of Twitter users.

That said, the potential applications of sentiment mapping are quite broad. The extremely large sample sizes available through social media can provide information on large populations of people. As Twitter and other social media platforms, some new, expand globally, the potential reach of this research will only continue to grow. Even if the data can only represent people using specific platforms, the capacity to approximate the regional moods or opinions of people, both spatially and temporally, has a broad range of research potential. Many areas of human research, from psychology and sociology to geography and public policy, have lines of research that could be supplemented with the use of sentiment mapping. While it is true that the results can only offer an approximation, this is also true of most broad-scale human research on mood or opinions.

There are many promising avenues for further research in the area of sentiment mapping. Although three types of sentiment maps were presented in this paper, other possibilities exist for using point patterns to analyze and visualize sentiment spatially, such as geographic analysis machine (GAM) (Fotheringham & Zhan, 1996; Openshaw, Charlton, Wymer, & Craft, 1987). Other methods for the temporal analyses may better visualize and/or quantify changes in sentiment temporally and spatiotemporally. A more statistical approach, such as space-time autoregressive modeling, might better quantify the stationarity or non-stationarity of the sentiment in different regions. Also of temporal interest is how the performance of the classifiers varies with time as the lexicon of the conversation around a given topic changes. Analyzing this "concept drift" (Forman, 2006) of the Twitter conversation may be a key part of understanding the temporal variation in sentiment observed in these data. It may also play a role in improving the performance of the classifiers.

Several other areas call for deeper understanding as well. Further evaluation and quantification of the spatial variability across the different machine learning classifiers is important for interpreting these results. Experiments with simulated data and/or more advanced statistical analysis of the spatial deviations of the sentiment maps could provide benchmarks for evaluating the results beyond mere visual interpretation. The results of this analysis tell a fairly consistent story about the broad sentiment distribution toward the topic of natural gas. Still, it is not clear what the likelihood of these results are. It could be that the techniques used simply lend themselves to consistent output so long as the machine learning classifiers are using the same training data and are reasonably well parameterized. Developing a method for hypothesis testing the sentiment map results to compare them to anticipated values or to other topics would be a significant step forward.

There is also further research needed to understand if new results are produced by taking the influence of tweets (such as likes and retweets) into account when measuring their distribution geographically. While this paper utilized only the raw presence of positive and negative tweets in its spatial analyses, creating the sentiment maps with the tweets weighted by their influence may tell a different story, or at the very least, might be used to address different questions.

Both within the area of sentiment mapping and in sentiment analysis of social media data more broadly, there is still more work needed to explore the biases of Twitter users compared to the general population, both in the US and abroad. For example, do the more general biases such as political leaning do well to predict the distribution of sentiment on specific topics, or do more general biases break down when selecting for keywords that isolate a particular subject? With some of these questions addressed, similar sentiment mapping research could contribute to the growing use of social media for disaster management (Houston et al., 2015), public health understanding (Daughton et al., 2018), or even to improve the targeted advertising efforts of businesses (Nair, Shetty, & Shetty, 2017), among many other possibilities.

What this paper offers is a first pass at combining sentiment analysis and point pattern analysis to create sentiment maps, providing some terminology and methods for visualizing and interpreting social media sentiment spatially. The methodology presented in this paper can be refined much further but perhaps it is enough to demonstrate the power infused in these techniques. There is room for improvement in every step of the process, including cleaner location analysis, more accurate machine learning classification, more substantial point pattern analysis, and a more in-depth temporal analysis. The work presented here offers only the

skeleton of sentiment mapping, a novel amalgamation of techniques with great potential for mapping the human social environment.

REFERENCES

REFERENCES

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S. H., & Liu, B. (2011). Predicting flu trends using twitter data. In *2011 IEEE Conference on Computer Communications Workshops, INFOCOM WKSHPS 2011* (pp. 702–707). https://doi.org/10.1109/INFCOMW.2011.5928903

Agarwal, A., Singh, R., & Toshniwal, D. (2018). Geospatial sentiment analysis using twitter data for UK-EU referendum. *Journal of Information and Optimization Sciences*, *39*(1), 303–317. https://doi.org/10.1080/02522667.2017.1374735

Aramaki, E., Maskawa, S., & Morita, M. (2011). Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter.pdf. In *Proceedings ofthe 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 1568–1576).

Ardon, S., Bagchi, A., Mahanti, A., Ruhela, A., Seth, A., Tripathy, R. M., & Triukose, S. (2011). Spatio-Temporal Analysis of Topic Popularity in Twitter. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 219–228). Retrieved from http://arxiv.org/abs/1111.2904

Baddeley, A., & Turner, R. (2005). spatstat: An R Package for Analyzing Spatial Point Patterns. *Journal of Statistical Software*, *12*(6).

Bertrand, K. Z., Bialik, M., Virdee, K., Gros, A., & Bar-Yam, Y. (2013). Sentiment in New York City: A High Resolution Spatial and Temporal View, 1–12. Retrieved from http://arxiv.org/abs/1308.5010

Bivand, R., Rowlingson, B., Diggle, P., Petris, G., Eglen, S., & Bivand, M. R. (2017). Package ‘ splancs .’ *R Package*.

Bivand, Roger, Keitt, T., Rowlingson, B., Pebesma, E., Sumner, M., Hijmans, R., … Bivand, M. R. (2015). Package ‘ rgdal .’ *R Package*.

Bollen, J., & Mao, H. (2011). Twitter Mood as a Stock Market. *Computer*, (October), 91–95.

Bollen, J., Mao, H., & Pepe, A. (2011). Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media Modeling* (pp. 450–453). Retrieved from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=emed5&AN=2001043531

Bowden, J. (2019). Trump energy officials label natural gas “freedom gas.” *The Hill*. Retrieved from https://thehill.com/policy/energy-environment/446004-trump-energy-officials-label-natural-gas-freedom-gas

Carneiro, H. A., & Mylonakis, E. (2009). Google Trends: A Web-Based Tool for Real-Time

Surveillance of Disease Outbreaks. *Clinical Infectious Diseases*, *49*(10), 1557–1564. https://doi.org/10.1086/630200

Chang, C., & Lin, C. (2011). LIBSVM : A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 1–27.

Connors, J. P., Lei, S., & Kelly, M. (2012). Citizen Science in the Age of Neogeography: Utilizing Volunteered Geographic Information for Environmental Monitoring. *Annals of the Association of American Geographers*, *102*(6), 1267–1289. https://doi.org/10.1080/00045608.2011.627058

Conover, M. D., Davis, C., Ferrara, E., Mckelvey, K., Menczer, F., & Flammini, A. (2013). The Geospatial Characteristics of a Social Movement Communication Network. *PLoS ONE*, *8*(3). https://doi.org/10.1371/journal.pone.0055957

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). # Earthquake : Twitter as a Distributed Sensor System. *Transactions in GIS*, *17*(1), 124–147. https://doi.org/10.1111/j.1467-9671.2012.01359.x

Culotta, A. (2010). Towards detecting influenza epidemics. In *Proceedings of the first workshop on social media analytics* (pp. 115–122).

Daughton, A. R., Paul, M. J., & Chunara, R. (2018). What Do People Tweet When They're Sick? A Preliminary Comparison of Symptom Reports and Twitter Timelines. In *ICWSM Social Media and Health Workshop*. Retrieved from www.aaai.org

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., … Seligman, M. E. P. (2015). Psychological Language on Twitter Predicts County-Level Heart Disease Mortality. *Psychological Science*, *26*(2), 159–169. https://doi.org/10.1177/0956797614557867

Erikson, R. S., & Tedin, K. L. (2001). *American public opinion: its origins, content, and impact* (Sixth Edit). New York: Longman. Retrieved from http://catalog.lib.msu.edu/search~S39?/twar+garden+victorious/twar+garden+victorious/1%2C1%2C2%2CB/frameset&FF=twar+garden+victorious+%2F&1%2C%2C2/indexsort=-

Fahrur Rozi, I., Rizky Yunianto, D., Mentari, M., Setiawan, A., Ariyanto, R., & Siradjuddin, I. (2018). Geo-Sentiment Analysis as a Location-Based Opinion Analysis System on Public Opinion Data about Governor Candidates. *International Journal of Engineering & Technology*, *7*(4.44), 110. https://doi.org/10.14419/ijet.v7i4.44.26873

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLoS ONE*, *5*(4). https://doi.org/10.1371/journal.pone.0010271

Forman, G. (2006). Tackling concept drift by temporal inductive transfer. *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2006*, 252–259. https://doi.org/10.1145/1148170.1148216

Fotheringham, A. S., & Zhan, F. B. (1996). A comparison of three exploratory methods for cluster detection in spatial point patterns. *Geographical Analysis*, *28*(3), 200–218.

Frank, E., & Bouckaert, R. R. (2006). Naive Bayes for Text Classification with Unbalanced Classes. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 503–510).

Freitas, A., Fernández, S., Hürlimann, M., Handschuh, S., Davis, B., & Cortis, K. (2016). A Twitter Sentiment Gold Standard for the Brexit Referendum. In *SEMANTICS* (pp. 193–196). https://doi.org/10.1145/2993318.2993350

Gatrell, A. C., Bailey, T. C., Diggle, P. J., & Rowlingson, B. S. (1996). Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers*, *21*(1), 256–274.

Gautam, G., & Yadav, D. (2014). Sentiment Analysis of Twitter Data Using Machine Learning Techniques and Scikit-learn. In *2014 Seventh International Conference on Contemporary Computing (IC3)* (pp. 437–442). https://doi.org/10.1145/3302425.3302492

Gayo-Avello, D. (2012). No , You Cannot Predict Elections with Twitter. *IEEE Internet Computing*, *16*, 91–94. https://doi.org/10.1109/MIC.2012.137

González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in Twitter: A closer look. *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, *2*(2010), 581–586.

Goodchild, M. F. (2007). Citizens as sensors : the world of volunteered geography. *GeoJournal*, *69*, 211–221. https://doi.org/10.1007/s10708-007-9111-y

Hauger, D., & Schedl, M. (2014). Exploring Geospatial Music Listening Patterns in Microblog Data. In *International Workshop on Adaptive Multimedia Retrieval* (pp. 133–146). https://doi.org/10.1007/978-3-319-12093-5

Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 237–246). https://doi.org/10.1109/icimw.2007.4516799

Herbst, S. (1995). *Numbered voices: How opinion polling has shaped American politics.* University of Chicago Press.

Hijmans, R. J., Etten, J. Van, Sumner, M., Cheng, J., Bevan, A., Bivand, R., … Wueest, R. (2015). Package ' raster .' *R Package*.

Houston, J. B., Hawthorne, J., Perreault, M. F., Park, E. H., Goldstein Hode, M., Halliwell, M. R., … Griffith, S. A. (2015). Social media and disasters : A functional framework for social media use in disaster planning , response , and research, *39*(1), 1–22. https://doi.org/10.1111/disa.12092

Hutto, C. J., & Gilbert, E. (2014). VADER : A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Eighth international AAAI conference on weblogs and social media*.

Ikawa, Y., Vukovic, M., Rogstadius, J., & Murakami, A. (2013). Location-based insights from the social web. In *Proceedings of the 22nd International Conference on World Wide Web* (pp. 1013–1016). https://doi.org/10.1145/2487788.2488107

Jurafsky, D., & Martin, J. H. (2019). Logistic Regression. In *Speech and Language Processing*.

Karami, A., Dahl, A. A., Turner-McGrievy, G., Kharrazi, H., & Shaw, G. (2018). Characterizing diabetes, diet, exercise, and obesity comments on Twitter. *International Journal of Information Management*, *38*(1), 1–6. https://doi.org/10.1016/j.ijinfomgt.2017.08.002

Kearney, M. (2019). rtweet: Collecting and analyzing Twitter data. *Journal of Open Source Software*, *4*(42), 1829. https://doi.org/10.21105/joss.01829

Kobayashi, Y., Mozgovoy, M., & Munezero, M. (2016). Analysis of Emotions in Real-time Twitter Streams. *Informatica*, *40*, 387–391.

Kolchyna, O., Souza, T. T. P., Treleaven, P. C., & Aste, T. (2015). Methodology for Twitter Sentiment Analysis.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, *26*(3), 159–190. https://doi.org/10.1007/s10462-007-9052-3

Kulldorff, M. (1997). A spatial scan statistic, *26*(6), 1481–1496. https://doi.org/10.1080/03610929708831995

Kulshrestha, J., Kooti, F., Nikravesh, A., & Gummadi, K. P. (2012). Geographic Dissection of the Twitter Network. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media* (pp. 202–209).

Larsen, M. E., Batterham, P. J., O'Dea, B., Boonstra, T. W., Christensen, H., & Paris, C. (2015). We Feel: Mapping Emotion on Twitter. *IEEE Journal of Biomedical and Health Informatics*, *19*(4), 1246–1252. https://doi.org/10.1109/jbhi.2015.2403839

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Flu : Traps in Big Data Analysis. *Science*, *343*(March), 1203–1206. https://doi.org/10.1126/science.1248506

Lee, E. C., Asher, J. M., Goldlust, S., Kraemer, J. D., Lawson, A. B., & Bansal, S. (2016). Mind the scales: Harnessing spatial big data for infectious disease surveillance and inference. *Journal of Infectious Diseases*, *214*(Suppl 4), S409–S413. https://doi.org/10.1093/infdis/jiw344

Liu, B. (2012). Sentiment Analysis: A Fascinating Problem. In G. Hirst (Ed.), *Sentiment Analysis*

and Opinion Mining (pp. 1–8). Morgan & Claypool.

Loper, E., & Bird, S. (2002). NLTK: The Natural Language Toolkit. *ArXiv Preprint Cs/0205028*.

McGough, S. F., Brownstein, J. S., Hawkins, J. B., & Santillana, M. (2017). Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data. *PLoS Neglected Tropical Diseases*, *11*(1), 1–15. https://doi.org/10.1371/journal.pntd.0005295

Nair, L. R., Shetty, S. D., & Shetty, S. D. (2017). Streaming big data analysis for real-time sentiment based targeted advertising. *International Journal of Electrical and Computer Engineering*, *7*(1), 402–407. https://doi.org/10.11591/ijece.v7i1.pp402-407

Openshaw, S., Charlton, M., Wymer, C., & Craft, A. (1987). A mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems*, *1*(4), 335–358. https://doi.org/10.1080/02693798708927821

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *LREc*, *10*(2010), 1320–1326.

Pebesma, E., & Bivand, R. (2013). Package ‘ sp .’ *R Package*.

Pedregosa, F., Weiss, R., & Brucher, M. (2011). Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pino, C., Kavasidis, I., & Spampinato, C. (2016). GeoSentiment: A tool for analyzing geographically distributed event-related sentiments. *2016 13th IEEE Annual Consumer Communications and Networking Conference, CCNC 2016*, 270–271. https://doi.org/10.1109/CCNC.2016.7444775

R Core Team. (2020). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In *Proceedings of the 20th international conference on machine learning (ICML-03)* (pp. 616–623).

Ripley, B. D. (1979). Tests of “Randomness” for Spatial Point Patterns. *Journal of the Royal Statistical Society. Series B (Methodological)*, *41*(3), 368–374.

Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., & Harabagiu, S. M. (2012). EmpaTweet: Annotating and Detecting Emotions on Twitter. In *Proceedings of the Language Resources and Evaluation Conference* (pp. 3806–3813). Retrieved from http://lrec.elra.info/proceedings/lrec2012/pdf/201_Paper.pdf

Rudis, B., & Thompson, C. (2018). rgeocodio: Tools to Work with the Geocodio “API”. Retrieved from https://github.com/hrbrmstr/rgeocodio

Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. *Information Processing and Management*, *52*(1), 5–19. https://doi.org/10.1016/j.ipm.2015.01.005

Sharag-eldin, A., Ye, X., & Spitzberg, B. (2018). Multilevel model of meme diffusion of fracking through Twitter. *Chinese Sociological Dialogue*, *3*(1), 17–43. https://doi.org/10.1177/2397200917752646

Smith, M. C., Broniatowski, D. A., Paul, M. J., & Dredze, M. (2015). Towards Real-Time Measurement of Public Epidemic Awareness: Monitoring Influenza Awareness through Twitter. *AAAI Workshop on the World Wide Web and Public Health Intelligence.*, *20052*. Retrieved from http://www.cs.jhu.edu/~mdredze/publications/2016_ossm.pdf

The Statistics Portal. (2018). Number of monthly active Twitter users worldwide from 1st quarter 2010 to 4th quarter 2018 (in millions). Retrieved from https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Twitter. (2020). Getting started — Twitter Developers. Retrieved February 15, 2020, from https://developer.twitter.com/en/docs/basics/getting-started

U.S. Energy Information Adminstration. (2019a). Natural Gas Dry Production. Retrieved from http://www.eia.gov/dnav/ng/ng_prod_sum_a_epg0_fpd_mmcf_a.htm%0A

U.S. Energy Information Adminstration. (2019b). Pennsylvania - State Energy Profile Analysis. Retrieved from https://www.eia.gov/state/analysis.php?sid=PA

Velázquez, E., Martínez, I., Getzin, S., Moloney, K. A., & Wiegand, T. (2016). An evaluation of the state of spatial point pattern analysis in ecology. *Ecography*, *39*(11), 1042–1055. https://doi.org/10.1111/ecog.01579

Vo, B.-K. H., & Collier, N. (2013). Twitter Emotion Analysis in Earthquake Situations. *International Journal of Computational Linguistics and Applications*, *4*(1), 159–173.

Wang, S., & Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. *50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference*, *2*(July), 90–94.

Wang, Z., & Ye, X. (2016). Spatial, temporal, and content analysis of Twitter for wildfire hazards. *Natural Hazards*, *83*(1), 523–540. https://doi.org/10.1007/s11069-016-2329-6

Wojcik, S., & Hughes, A. (2019, April). Sizing Up Twitter Users. *Pew Research Center*.

Zandbergen, P. A., & Barbeau, S. J. (2011). Positional accuracy of assisted GPS data from high-sensitivity GPS-enabled mobile phones. *Journal of Navigation*, *64*(3), 381–399. https://doi.org/10.1017/S0373463311000051