8-1-2019

SOCIAL MEDIA FOOTPRINTS OF PUBLIC PERCEPTION ON ENERGY ISSUES IN THE CONTERMINOUS UNITED STATES

David Leifer
Southern Illinois University Carbondale, davleifer@gmail.com

SOCIAL MEDIA FOOTPRINTS OF PUBLIC PERCEPTION ON U.S. ENERGY ISSUES

By
David Leifer
B.A., University of Wisconsin- Eau Claire, 2016

A Thesis
Submitted in Partial Fulfillment of the Requirements for the
Master of Science Degree

Department of Geography and Environmental Resources
In the Graduate School
Southern Illinois University Carbondale
May 2019

THESIS APPROVAL

SOCIAL MEDIA FOOTPRINTS OF PUBLIC PERCEPTION ON U.S. ENERGY ISSUES

By

David Leifer

A Thesis

Submitted in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in the field of Geography and Environmental Resources

Approved by:

Dr. Ruopu Li

Dr. Justin Schoof

Dr. Guangxing Wang

In the Graduate School
Southern Illinois University- Carbondale
May 12, 2019

AN ABSTRACT OF THE THESIS OF

David Leifer, for the Master of Science degree in Geography, presented on May 11, 2019, at Southern Illinois University Carbondale.

TITLE: SOCIAL MEDIA FOOTPRINTS OF PUBLIC PERCEPTION ON U.S. ENERGY ISSUES

MAJOR PROFESSOR: Dr. Ruopu Li

Energy has been at the top of the national and global political agenda along with other concomitant challenges, such as poverty, disaster and climate change. Social perception on various energy issues, such as its availability, development and consumption deeply affect our energy future. This type of information is traditionally collected through structured energy surveys. However, these surveys are often subject to formidable costs and intensive labor, as well as a lack of temporal dimensions. Social media can provide a more cost-effective solution to collect massive amount of data on public opinions in a timely manner that may complement the survey. The purpose of this study is to use machine learning algorithms and social media conversations to characterize the spatiotemporal topics and social perception on different energy in terms of spatial and temporal dimensions. Text analysis algorithms, such as sentiment analysis and topic analysis, were employed to offer insights into the public attitudes and those prominent issues related to energy. The results show that renewable energy sources were consistently more positive than either nuclear or coal, which holds true when viewed temporally or spatially. The study is expected to help inform decision making, formulate national energy policies, and update entrepreneurial energy development decisions.

ACKNOWLEDGMENTS

I would like to acknowledge the Department of Geography and Environmental Resources for their continued support of my research and teaching. I would also like to acknowledge the Advanced Coal and Energy Research Center for the financial support of my Thesis. I especially would like to acknowledge Dr. Ruopu Li for his invaluable advice and my parents for their financial and emotional support. I would finally like to thank Dr. Justin Schoof and Dr. Guangxing Wang for their instructions and serving as members of my thesis committee.

TABLE OF CONTENTS

CHAPTER PAGE

ABSTRACT……i

ACKNOWLEDGMENTS……...……...….ii

TABLE OF CONTENTS……...iii

LIST OF TABLES……..vi

LIST OF FIGURES……..……vii

CHAPTERS

CHAPTER 1 – Introduction…….....1

1.1 Problem Statement……2

1.2 Research Questions ……...……3

1.3 Research Significance……..……...……3

1.4 Structure Overview……4

CHAPTER 2 – Literature Review……...5

2.1 Using Social Media to Understand Energy Perception…..……5

2.2 Data Mining Techniques…….....6

2.2.1 The General Workflow……..…….....7

2.2.2 Sentiment Analysis……..…….10

2.2.3 Topic Analysis……..……14

2.3 Spatiotemporal Dimensions of Social Media Data…….16

CHAPTER 3 – Methods……24

3.1 Study Area…….……..24

3.2 Twitter Web Crawler……..25

3.3 Geocoding Twitter Users' Location……...27

3.3.1 Setting up the Geocoding Script……..……...27

3.3.2 Importing the Libraries……...28

3.3.3 Setting up an Iterative Loop and Collecting Data……..28

3.4 Geocoding Process…….…….…29

3.5 VADER Sentiment Analysis…….30

3.6 Accuracy Assessment……31

CHAPTER 4 – Results and Discussion…….……..33

4.1 Overall Social Sentiments on Energy…….....33

4.1.1 Social Perception on Solar…….....34

4.1.2 Social Perception on Coal……...35

4.1.3 Social Perception on Nuclear……..35

4.1.4 Social Perception on Wind……..35

4.2 Energy Topic Analysis…….....36

4.2.1 Coal Topic Analysis……36

4.2.2 Solar Topic Analysis……37

4.2.3 Nuclear Topic Analysis……39

4.2.4 Wind Topic Analysis……40

4.3 Who tweeted about Energy?…….....41

4.3.1 Who tweeted about Coal?..…….....42

4.3.2 Who tweeted about Solar?..…….....43

4.3.3 Who tweeted about Nuclear?..…….....45

4.3.4 Who tweeted about Wind?……..…....46

4.4 Spatial Distribution of Energy-related tweets……47

4.4.1 Spatial Distribution of Coal-related tweets…….48

4.4.2 Spatial Distribution of Solar-related tweets……....49

4.4.3 Spatial Distribution of Nuclear-related tweets……....49

4.4.4 Spatial Distribution of Wind-related tweets……....50

4.5 Spatiotemporal Patterns of Social Perception…….51

4.5.1 Spatial Patterns for Coal Sentiments.....…….53

4.5.2 Spatial Patterns for Solar Sentiments…….....54

4.5.3 Spatial Patterns for Nuclear Sentiments…….....56

4.5.4 Spatial Patterns for Wind Sentiments……..…....57

4.6 Accuracy and Uncertainties……....58

4.6.1 Geocoding Uncertainties…….....58

4.6.2 Sentiment Accuracy……..….....59

4.7 Policy and Entrepreneurial Implications……..60

CHAPTER 5 – Conclusion……62

5.1 Summary of Findings……...62

5.2 Limitations and Recommendations…….....63

5.3 Conclusions and Final Remarks……...66

REFERENCES……..……....68

VITA……..75

LIST OF TABLES

TABLE PAGE

Table 1 – Hashtags for my focused energy types.....26

Table 2 - A table displaying coal and solar sentiment accuracy scores…….....59

LIST OF FIGURES

FIGURE PAGE

Figure 1 - A chart of my workflow……..…….....26

Figure 2 - The percent of electricity produced by non-fossil fuel sources..……..……...34

Figure 3 - The coal word frequency chart……...…….....37

Figure 4 - An example of Trump's tweet regarding coal.……...…….....37

Figure 5 - The solar word frequency chart.…….....….…….....38

Figure 6 - A tweet by user @mikehudema about jobs in solar energy.…….....…….....39

Figure 7 – The most frequent words from the nuclear dataset.....……..…….....40

Figure 8 - An example tweet from user @NEI.....…….…….....40

Figure 9 - The most frequent words for the wind energy type.……...……....41

Figure 10 - The top Twitter users from my coal dataset.……..…...……...42

Figure 11 - An example tweet from @ Bwillisful.…….…….....43

Figure 12 - The top power users for the hashtag solar dataset.…….…….44

Figure 13 - A recent tweet by @solargenerator2.....……..44

Figure 14 - The user frequency chart for nuclear energy.....…….45

Figure 15 - The user frequency chart for wind energy type..…….46

Figure 16 - An example tweet from RenewableSearch.....……...47

Figure 17 - The coal location frequency chart.....……...….48

Figure 18 - The location frequency chart for solar.…….....…….49

Figure 19 - The location frequency for nuclear energy.....…….50

Figure 20 - The location frequency for wind.....…….51

Figure 21 – A timeseries graph of solar, coal, nuclear, and wind sentiment…….…….52

Figure 22 - Coal location kernel density and county spatial join.……..……...54

Figure 23 - Five months of solar data plotted by kernel density and county spatial join.…….…...55

Figure 24 - Five months of nuclear data plotted by kernel density and county spatial join.….....56

Figure 25 - Five months of wind data plotted by kernel density and county spatial join..……...57

CHAPTER 1

INTRODUCTION

Energy research has transitioned from a research area dominated by physical energy systems to one focusing on social dimensions. The success of programs on energy development and consumption depend on how the public perceives those related issues, such as environmental pollution and anthropogenic climate change. A common way to measure social perception is through phone and structured mail surveys (e.g., Farhar et al., 1980; Farhar et al., 1994; Bolsen and Cook, 2008; Sütterlin and Siegrist, 2017). Although these approaches statistically represent the surveyed population, they are subject to a few limitations, such as its lack of temporal scales, intensive labor costs, low response rates, and limited space for free expression (Ceron et al., 2014). Social media-based data mining offers an alternative insight into people's opinions on a variety of energy issues.

Social media datasets are within the scope of big data, which can be characterized by four V's: volume, variety, velocity, and value. Volume refers to the amount of data, variety is the different forms of data, velocity is the analysis of streaming data, and value is the ability to use the data as actionable information (Kaisler et al., 2013). Accessing this large stream of information poses challenges because the data needs to be filtered and further condensed for analyses using traditional methods to occur. Twitter specifically states that over 500 million tweets are created each day by hundreds of millions of accounts (Twitter for Business). Because of the size, variability, type, and uncertainty presented by social media data, we employed scripting languages such as Python to scrape, condense, analyze, and measure the uncertainty of the data.

1.1 Problem Statement

Energy has been developed and consumed at an unprecedented pace, as the world is becoming more populated and third world countries are rapidly developing. According to the Intergovernmental Panel on Climate Change (IPCC)'s Summary for Policymakers (2014), the global average temperature is expected to increase by 1.5-2.0 °C by the end of the 21st century. It is widely accepted by the scientific community that climate change is directly linked to anthropogenic energy-related activities, i.e., the release of carbon dioxide and other greenhouse gasses from the burning of coal, natural gas, and other fossil fuels. As developing nations, such as India and China, have increased in population and industrialization, their demands for energy have reached all-time highs. Despite having a smaller amount of the world's population, the United States consumes a disproportionately large amount of energy. As we begin the transition from a fossil fuel energy economy to one dominated by renewable sources, how people perceive different types of energy needs to be examined to better inform decisions on energy development.

Social perception of energy affects how government officials set national, state, and local energy agendas. Social perception on energy has been found to be an important factor behind social and political campaigns that sway decision making on energy development, consumption, and policy. (Batel et al., 2013). Therefore, it is necessary to understand the public preference and opposition towards different energy types and related concerned issues. Public perception towards energy sources will significantly influence our energy future (Wüstenhagen et al., 2007).

Social media platforms, such as Facebook, Twitter, and YouTube, have emerged in the last decade as incredibly popular means of sharing opinions and connecting with other individuals. These uncensored media have increasingly been used to study public opinions on

social issues (Petkov et al., 2011; Bertot et al., 2012; Sivarajah et al., 2015). Social media also provides rich textual information that is largely limited or unavailable in traditional surveys. Since social media has been routinely used to study a variety of social research issues, it is necessary to apply these established methods to study today's "energyscape". Social media provides a backdrop of temporal, spatial, and contextual information along with a freedom of expression not found in traditional surveys. Many contemporary issues, such as social images of coal and the transition to carbon-zero energy sources, can thus be studied by what people post on the internet. For example, Santoianni (2013) revealed that industries powered by coal have been facing increasing public opposition on social media.

1.2 Research Questions

This thesis focuses on two research questions:

1. What are the public perception and concerned issues towards different energy sectors in the U.S. based on social media?

2. Can we gain insights of spatiotemporal differences in social perception on energy issues using social media-based data mining approaches?

1.3 Research Significance

This study contributes to the knowledge of location-based perception towards energy resources by leveraging uncensored and free social media conversations. Additionally, this study adds to the plethora of literature regarding the applications of social media data mining techniques. It is my intention that the outcome of this research be used by policy makers and entrepreneurs to support strategical decisions about energy development in the United States.

1.4 Structure Overview

Chapter 2 is dedicated to the literature review of social perception towards energy and potential data mining techniques, including past research using social media location, sentiment analysis, and topic analysis on very large text corpora. Chapter 3 outlines the methodology for scraping Twitter data, extracting geolocations from user profiles, and analyzing the data using machine learning techniques. Chapter 4 presents and interprets the results and discusses technical and social issues related to the research outcome. Finally, chapter 5 summarizes the research findings and proposes future work.

CHAPTER 2

LITERATURE REVIEW

This chapter outlines the literature review or methods that will be employed throughout the study. In section 2.1, we examined the literature behind using social media to understand energy perception. In section 2.2, we covered the workflow and various data mining and sentiment analysis techniques. Finally, in section 2.3, we analyzed the literature on spatiotemporal characteristics of social media datasets.

2.1 Using Social Media to Understand Energy Perception

Although there has been research on social perception towards a variety of subjects on social media, we have identified a gap when it comes to the perception of energy sources on social media. Existing studies on energy sources from social media include energy and environment (Santoianni, 2013), energy conservation domestically (Petkov et al., 2011), and energy efficiency and savings (Russel et al., 2013). For example, Santoianni (2013) found that pages on Facebook supportive of coal were far outnumbered by pages that were anti-coal. Furthermore, 76% of posts containing the word “coal ash” were unfavorable. Similarly, Russel et al. (2013) found that Twitter data was useful in discovering communication campaigns about energy-related posts.

A study conducted by the Pew Research Center (Americans' Opinion on Renewables and Other Energy Sources, 2016) found interesting results about the perception of the American people towards various energy types. They conducted a traditional survey and had respondents include their political affiliation. Coal was very controversial, with only 41% in favor of

expanding coal while 57% oppose of it. 89% of all respondents favored more solar panel farms while just 9% oppose it. Nuclear was also more controversial, with 43% favoring more nuclear and 54% opposing it. 83% favor more wind turbine farms while just 14% oppose them. The study also found divisions when controlling for political affiliation. Although both Republicans and Democrats favor more renewable resources, Republicans support more coal mining with a 73% favor rate while just 14% of Democrats favor it. 57% of Republicans favor more nuclear plants and 40% of Democrats favor it.

Although traditional surveys provide an important background on American perception on energy sources, they often lack, or avail at formidable costs, important spatiotemporal dimensions. Most of surveys can be regarded as 'snapshots' of social conditions in terms of the time periods. A national survey that covers a wide range of geographic areas is usually extremely expensive. Additionally, as the respondents fill in bubbles to predefined questions, the contextual background and unstructured information related to the surveys may be unavailable. Most importantly, these traditional surveys often limit spatial extents due to formidable costs.

2.2 Data Mining Techniques

As this study relies on data mining, it is important to address the body of literature that pertains to a wide range of data mining techniques. As previously noted, social media platforms in the age of Web 2.0 supplies a wealth of social conversations. However, a majority of the generated information was designed in such a way that reduces storage space in an effort to speed up the call and response time of the web interface. Thus, we are left with a term coined as "messy data", or data that needs to be cleaned using preprocessing techniques.

2.2.1 The General Workflow

The general workflow of analyzing social media includes collecting, cleaning, quality checking, and analyzing the data. To remedy the variable quality of social media data, it is imperative to employ data cleaning methods developed by Wickham (2014). This was an attempt to create a framework for making these complicated tabular sets of data easier to be analyzed by examining five common problems. This includes column headers stored as values instead of names, multiple variables stored in a single column, variables being stored in columns and rows, many different types of observational units stored in the same dataset and finding a single observational unit in multiple datasets. Approaches to remedying this include the following and align respectively with the five problems: melting columns into rows, splitting columns, rotating melted rows into columns, splitting a single dataset into multiple datasets, and concatenating multiple datasets into a single dataset.

Once the data is collected, it is important to discover how the data will be implemented to answer questions. He et al. (2015) created a social media analytics framework to analyze sentiment for business competitors. Their approach used IBM SPSS to extract information and perform cluster analysis on the data. The problem with building such a framework is that since the programs are written to utilize existing APIs and web crawlers, if the API or web crawler is updated by Twitter or another organization, the tool breaks and becomes worthless. It would take a dedicated team to constantly keep the tool up to date with the APIs and web crawlers. This is beyond the scope of my own survey.

After the data is collected and properly cleaned, there are myriad ways of progressing with the analyzation. One such approach examined by Hu et al. in (2015) created an agent-based model to simulate how people interact with each other and mass news on social media. This

model consisted of committed agents, who cannot change their opinions, influencing uncommitted agents, who update their opinions at each time step. The uncommitted agent is also influenced by mass news media sources, which acts as a weight that can be increased or decreased. Similarly, Anderson et al. (2015) attempted to create an agent-based model represented as a social graph to keep track of how disaster news about people's whereabouts is disseminated as it breaks in real time. This is highly dependent on the amount of information a person creates about themselves in the minutes and hours after a disaster to feed the study's support vector machine and random forest classifier algorithm. Without provided data, it is largely difficult to estimate their status. A limit of both of these approaches is the lack of literature linking quantitative models to real world humans making decisions. The real world is so stochastic that it is difficult to account for every minuscule detail in a model.

A comprehensive literature review of the tools available was examined by Adeoyin-Olowe et al. (2014). The aspect-based/feature-based approach determines the polarity of an overall review based on whether the text was positive or negative. Homophily clustering links the sentiment of the same opinion with other similar opinions as a series of nodes. Opinion definition/summarization is a machine learning approach that learns the polarity of text using support vector machines. Sentiment orientation rates a text on a one to five-star basis. Aspect rating analysis extracts each aspect of a text via probabilistic latent semantic analysis for each subsequent text's score. Sentiment lexicon is an approach that uses a dictionary of predefined words (also known as a text corpus) and scores the text based on if the predefined word appears in the text. Unsupervised classification discovers adjectives and adverbs to classify the text using average semantic orientation. This approach does not distinguish between test and training data, using all supplied data to build a prediction model. Supervised and semi-supervised approaches

split the data into training and testing data to evaluate the model's predictive powers in classifying text as positive or negative. Topic detection or topic analysis uses support vector machines or naïve Bayesian classifiers to specify words as either positive or negative.

Rather than relying on state-of-the-art opinion estimation algorithms, Russell et al. (2013) analyzed the content of social media regarding energy consumption. The purpose of the study was to decipher the feasibility of using Twitter streams as a data source and analyze the collected data for word frequency, periodicity, and context. The analysis portion was concerned mainly with co-occurring hashtags and clusters of words rather than the actual analysis of the text's contents. This focus provides a framework for discovering similar hashtags on Twitter for use in data collection.

The validity of using social media as a data source was questioned in prior research applications, such as in Adedoyin-Olowe et al. (2014) and discussed in Tufekci (2014) and Oliveira et al. (2017). Tufekci (2014) specifically set out to address the over emphasis of using only one platform, Twitter, in academic research to summarize the consensus of social media. Existing frameworks in prior research fails to accommodate the sampling bias of using only hashtags in data collection. One side of an argument might use a particular hashtag more than the other, providing another basis of sampling bias. Such shortfalls of using Twitter include the one-way graph between users; one user does not have to confirm the friendship of another to follow or follow back. Additionally, Twitter is confined to short texts, lacks private communication, and the demographics of its user base are primarily young, white adults. Less than 20% of adults in the United States use Twitter on a frequent basis (Tufekci, 2014). Measurable statistics such as retweets can also lead to incorrect conclusions about the data. Trending news might accumulate a mass of retweets and comments but has no bearing on whether the news was either positive or

negative or was even factual. To further address the question of validity of using a social media microblogging service like Twitter, Oliveira et al. (2017) compared the results of sentiment analysis and traditional opinion polls during the Brazilian presidential race. They collected data from Twitter during the same time period as the traditional polls. Opinion mining software such as DiscoverText, RapidMiner, and Scup was utilized and got an 81% correct classification rate on their Twitter data while error from traditional polls was 95%. The candidate Dilma had more positive tweets than did Aecio, which coincides with the traditional opinion poll stating that Dilma had taken the lead over Aecio.

Another problem with using Twitter as a data source is that only a small percentage of tweets have attached geolocation information. Resulting from my preliminary study, less than 1% of all tweets had attached geolocated coordinates. This statistic is confirmed when analyzing the literature and is most likely because people have to opt into activate this feature, besides the fact that the scale at which the geocoding occurs varies widely (Crampton et al., 2013). Stefanidis et al. (2011) attempted to rectify this lack of available data by collecting what was coined as ambient geospatial information. tweets were included if they contained a keyword about a physical location and an example about Tahrir square during the Arab Spring in 2011 was examined. They modeled the connection of users via a graph system of interconnected nodes.

2.2.2 Sentiment Analysis

Sentiment analysis is critical to the general emotional tendency of people towards the concerned issues, resulting in numerical scores representative of positive, negative, and neutral

attitudes. There are many studies that focus on the algorithmic and applied aspects of the sentiments. Due to its importance, we make it a separate section to dedicate to its review.

The ease of use for a sentiment analysis tool was examined by self-proclaimed novice technologists in Yoon et al. (2013). Rather than using complex machine learning algorithms, this was an attempt to provide a comprehensive approach to sentiment analysis. They attempted to classify sentiment towards 17 keywords relating to physical activity. They imported the tweets using NodeXL, a free and open source Microsoft Excel application that analyzes social media networks and visualizes the results. This application reduced dimensionality and removed special characters, and then broke the words down into n-grams, which are a chain of text words starting with unigrams, progressing to bigrams, and eventually forming chained-together words. To analyze the preprocessed data, the information was deciphered using topic detection, sentiment analysis, and hot topics. The information was then displayed as the frequency of content changes over time by week. For example, 77% of bicycling related tweets were classified as positive and 40% of running related tweets were classified as negative. This study provides an important backdrop to the ease of use in using some of these sentiment analysis tools.

A different approach to sentiment analysis was carried out by Williams et al. (2015), which examined the polarizing sentiment of attitudes towards anthropogenic climate change on social media by exploring the various echo chambers that arise from five Twitter hashtags. Using each hashtag, the mean Sorenson similarity between user populations was calculated for ten-day intervals. This resulted in 15 networks from followers to people that they retweet and were visualized as directed graphs created by the ForceAtlas2 layout algorithm. ForceAtlas2 refers to a combination of a variety of techniques used by Gephi, a network visualization software, to visualize social networks. The users themselves were classified by a panel of researchers as

activist, skeptic, neutral, or unknown and the tweets were classified by the same panel as positive, negative, neutral or unknown. Homophily was then measured between similar users and dissimilar users and a Louvain method algorithmically detected which group a user belonged to. The results confirmed that there are many more skeptics and activists than neutral or unknown, indicating the polarizing nature of the subject. There was also little retweeting and inmixing by the users on opposing sides, indicating that they largely ignored each other. Although hand labeling the sentiment is an interesting approach, it falls beyond the scope of my study.

Another example of sentiment analysis on data collected from Twitter was proposed by Cody et al. (2015) in regard to climate change. They employed the Hedonometer, a previously created social media metrics tool, to analyze the change in sentiment on Twitter. The Hedonometer assigns scores based on the positive or negative sentiment and draws from a corpus of 10,222 most frequently used words in social media. The scores ranged from one to nine and created word shift graphs comparing the happiness of two texts by ranking the content of their words. Four events were analyzed in this study and included three natural disasters and a climate rally. They discovered that the frequency of the word climate decreased over the period of study from 2008 to 2014. The average happiness score in regard to climate was 5.84/9.0, while the average of all tweets during this time was 5.99/9.0. It was also realized from the data that climate change deniers used the term “global warming” more than climate change activists.

A more complex study on Twitter compared three different approaches to sentiment analysis: lexicon-based approach, machine learning, and a hybrid-based approach (Chen et al., 2016). This study looked to explore how various governments use social media by examining 20 different city government accounts over eight months. Three different lexicons or text corpora were used: Taboada et al., Valence Aware Dictionary and sEntiment Reasoner (VADER), and

the National Research Council Emotion Lexicon. The text processing library used was the Natural Language Tool Kit (*nltk*). The machine learning implementation they used was Weka, a library written in the Java programming language. The text was transformed into a vector using StringToWordVector and then used the Naïve Bayes, K-nearest Neighbors, and Random Forest algorithms. Training data was taken from Sentiment140. A final approach combined the two methods by using Senti-Strength to classify short, informal text. The hybrid approach also classified emoticons. An Analysis of Variance (ANOVA) test was used to show dips and spikes as the government announced music festivals or other expositions. This study is specifically useful because they employed the VADER corpus, which is a widely adopted method in exploring short social media texts.

Lexicon-based approaches to sentiment analysis include Natural language processing (NLP) steps to preprocess the data and have data cleaning, stop word removal, hanging feature removal, and character disabling (Jeong, 2015). Machine learning approaches, such as those included in Go et al. (2009), Mikolov and Le (2014) and Lau and Baldwin (2016), employed Naïve Bayes, Maximum Entropy, Support Vector Machines, and doc2vec algorithms to analyze the sentiment. Accuracy for these methods ranged from 80% to 82% accuracy. The results were different when they attempted to account for two or more words in a sequence, known as bigrams. When unigrams were included with bigrams, Maximum Entropy accuracy improved but Naïve Bayes and Support Vector Machines did not. Compared with bag of word models, doc2vec far outperformed the accuracy by representing words as vectors in feature space. Although these are significant contributions to the field of NLP, the accuracy results proved inconclusive to replicate (Lau and Baldwin, 2016).

2.2.3 Topic Analysis

Topic analysis, also known as topic modeling or feature extraction of text corpora, is important to understand social media discussions in terms of generalized common topics. Dai et al. (2015) used the algorithm doc2vec to learn paragraphs of Wikipedia and arXiv as embedded paragraphs. This approach was compared with Latent Dirichlet Allocation (LDA). The paper attempted to see if the methods could find similar paragraphs based on a given input word. The Paragraph Vector Model works as follows:

1) Insert the memory vector into a standard language model.

2) The paragraph vector is then averaged with local context to predict the next word.

3) Backpropagation is issued to tune paragraph vectors (known as distributed bag of words).

Word embeddings were jointly trained with paragraph vectors. Triplet pairs of items that were close to each other were then constructed. 10 epochs of training using hierarchical softmax as a Huffman tree classifier occurred. Cosine similarity was used as the similarity index. LDA was applied with Gibbs sampling and 500 iterations, where alpha was set at .1. Posterior topic proportions with Hellinger distance was used to compute similarity. The word embeddings for each word was then averaged. They visualized 915,715 words using the t-SNE algorithm. They found the nearest neighbors for the term ‘Machine Learning’. Deep learning is more similar to machine learning than to a computer network. They found that the term ‘Google’ is closer to ‘Facebook’ than ‘Walmart. They also found that paragraph vectors outperformed LDA in the similarity index.

Kusner et al. (2015) measured the similarity of text documents using the Word Mover’s Distance (WMD) algorithm, which is the distance the words need to travel in space to reach

other words. They approached the subject as though it is a transportation problem. Their algorithm was based on word2vec's shallow neural network architecture and is as follows:

1) An input layer, projection layer, and output layer are used to predict close words.

2) Word vectors are then trained to max log probability of nearby words.

An example of this input/output sequence would be 'Japan' minus 'sushi' plus 'Germany' would equal the output term 'bratwurst'. The Euclidian distance was found between similar words to produce a dissimilar index. Another example would be that transforming 'Illinois' to 'Chicago' is closer than 'Japan' to 'Chicago'. Each document is represented by its weighted mean vector. WMD, albeit slow in performance, outperformed all other attempts with the lowest test error. It was also found that the model improved with sample size.

In the paper by Zhang and Wallace (2015), they attempted to conduct an analysis of a single layer convolutional neural network (CNN) with altered parameters in the sentence classification task. They converted a sentence to a vector matrix for input from word2vec or GloVe. They acknowledge that tuning the hyperparameters of the CNN is like a black art. They took a significantly long time to train, thus it is important that the hyperparameters are tuned correctly to begin with. There were multiple steps to counteract this, as described in the paper. First, perform convolution with linear filters, where the filters were parameterized by the weight matrix. The inclusion of a bias term and activation function to create a feature map, using 1-max pooling to extract a scalar from each feature map. Next, they were concatenated into a fixed length feature vector and fed into a softmax layer. Regularization was then applied via dropout or L2 norm. The aim of this is to minimize cross-entropy loss by measuring the parameters. Optimization is done using stochastic gradient descent and back propagation. Instead of reporting on the accuracy, they instead measured the area under the curve. They found that a

CNN is not suitable when the dataset is small. The best result for filter region depends on the sentence length. 1-max pooling outperformed all others while dropout outperformed L2 norm for regularization when set to a small value. Finally, it was recommended to use word2vec or GloVe rather than one-hot vectors for topic analysis.

An interesting approach to sifting through the roughly 10% of robot posts on Twitter was examined in Chen et al. (2017). They define this spam as low-quality content and developed an expectation maximization algorithm on tweets by dividing tweets into four categories: low quality advertisements, automatically generated content, meaningless content, and clickbait. They then surveyed 211 participants in an attempt to define low-quality content. 100,000 tweets were labeled using this approach. Support Vector Machines and random forest classifiers were then applied to classify more tweets into categories. Cohen's Kappa coefficient was used to evaluate the agreement of the surveyed labeling. From their results, words like 'click' and 'free' appeared more in low quality content. From building a blacklist dictionary, about 95% of tweets were correctly categorized as low-quality content. This number improved to 97.11% when word level analysis was implemented.

2.3 Spatiotemporal Dimensions of Social Media Data

Social media is rich in textual information with spatiotemporal dimensions. It has been in the research spotlight along with the 'geospatial revolution' in the recent decade. Goodchild and Sui (2011) curated an article discussing the new challenges faced by geographic information systems (GIS) as social media emerges as a new topic. The rise of websites like Google Maps, Bing Maps, and Yahoo Maps have drawn in users by the millions and created new GIS/social media communities, resulting in real life meetups. This mapping has expanded to other social

media outlets that employed location-based services to track its users such as Facebook and Twitter. The rise of big data from these location-based media sites has opened up new frontiers for GIS studies, resulting in three categories of data interaction: people who generate data, people who collect data, and people who can analyze data. A foreseen pitfall in future studies on big data is discovering how people without mobile phones can be included in the study. New developments in data analyzation frameworks will need to be created to account for this influx in available data collection.

Since the introduction of smartphones equipped with 5-15-meter positional accurate global positioning satellites (GPS) and highly accurate clocks, the spatiotemporal variability of human movement patterns has been examined by researchers. Hasan et al. (2013) used social media data to track the timing component as well as spatial distribution of social media entries in Los Angeles, New York City, and Chicago. By breaking up the Twitter data down to users who had more than 25 geolocated check-ins, the study categorized the likelihood an individual was either at home, work, eating, being entertained, recreating, or shopping. The temporal component was crucial in impacting the spatial aspect of the activity. For example, entertainment was usually logged in the evening and eating was geographically centered. This work could be employed to probabilistically track a person choosing a destination. The idea of using the spatiotemporal configuration for location recommendation is expanded upon by Hu et al. (2013). They developed a supervised learning model to predict the likelihood of an individual choosing a destination based on the Monte Carlo Expectation Maximization. Their model outperformed similar state of the art approaches.

One of the first applications of this new geospatial technology from social media arose from Crandall et al. (2010), which examined the online data of people to see if those who were

located near to each other in space-time knew each other. The photo-sharing website Flickr was utilized to scrape data with a time stamp and geolocated coordinates. About 85 million such photos were crawled and filtered down to 38 million photos from 490,000 users. The Earth was subsequently divided into a grid and if users appeared in the same grid, they were marked as co-occurring. To decide if they knew each other, their friend list was examined. The results concluded that the probability of a connection increased with the presence of a co-occurring photo and the temporal range decreased. Bayesian inference was used to examine the likelihood of knowing each other based on the aforementioned information.

Further exploration of the value of spatiotemporal data was examined by Hossain et al. (2016) in deciphering alcohol consumption patterns based on geolocated Twitter data. This study collected geo-tagged tweets from rural, urban and suburban areas in New York state. By employing the payment-based Amazon Mechanical Turk, the tweets were connected to actual users who said they were drinking at the time of the tweet. The dataset was preprocessed and used to build a hierarchical support vector machine, a type of supervised machine learning algorithm, which avoids overfitting by restricting the dataset. The model was 80% accurate in determining if a person was tweeting from their home and drinking, indicating a high degree of accuracy in predicting block level geolocation.

An interesting approach to the spatiotemporal variations of human movements in real world spaces along with virtual places was in Gao et al. (2018). This was an attempt to better understand the human dynamics at play in geotagged tweets. They defined three edges in a spatiotemporal network (STN): physical edges, social edges, and physical-social edges. This study took Michael Phelps tweets during the 2016 Olympics from three time periods: after the 4X100M medley relay, after Schooling beat him, and after a return to the United States. They

also analyzed 30 retweets from Phelps after he won and 20 after he lost. Most of the retweets came from the user's original location. For example, most of Phelps' retweets came from the United States while Schooling's came from Singapore and from the University of Texas-Austin, where Schooling is a student. This study is important because of the implications from studying human movement along with social or behavioral science as it relates to computational modeling.

In an attempt to examine the spatiotemporal movement patterns in the aftermath of a hurricane, He et al. (2015) took 418 million geotagged tweets over a six-month period and split the dataset into temporally related to Hurricane Matthew, spatially related to Matthew, and evacuation behavior analysis categories. They employed a regression analysis from the normalized Twitter activity by county and the length from the county centroid to the track of the hurricane. They found that 60% of tweets came from Florida and peaked on the day of the hurricane, then declined rapidly. They established a threshold of ten tweets to discover the user's home location. From the tweet dispersal, it is estimated that 54% of people actually evacuated, with most evacuees (45.6%) residing in South Carolina.

Perhaps a more important application of this new highly accurate temporal data can be found in a De Choudhury et al. (2013) study in predicting depression from social media posts. First, a list of Twitter users that were diagnosed with depression was compiled and used to scrape their tweet history for the past year. Amazon Mechanical Turk was used to administer a clinical survey to analyze their depression history. The applicants were not told the nature of the study, just that it was a study on behavioral psychology. From a pool of 1,583 crowdsourced individuals, only 171 met the criteria of having depression, were willing to share their Twitter posts, and passed the validity tests. Temporally, it was discovered that depressed individuals had

peak Twitter usage late at night and early in the morning while the opposite is true for non-depressed individuals. Depressed users were also less likely to post at all and were more likely to post about their symptoms than non-depressed individuals. A limitation in using this study is the lack of a spatial component to make predictions.

Huang et al. (2014) used an approach to geolocating tweets using no less than 50 geotagged tweets with a spatial accuracy of 20 by 20 meters. The study discovered spatial clustering by utilizing a density-based approach with a noise algorithm. Temporally, the noise algorithm was employed to find the nearest neighbor instead of the Euclidean distance. The point level data of 1,259 geotagged tweets around St. Louis, Missouri was then joined with a land-use map. Eleven high representative activity zones with more than six tweets were identified by looking at the cluster of tweets in space and four zones were identified along the temporal component. It was apparent from the data that areas of traffic jams often had numerous tweets.

Making accurate predictions based on a user's movements is a valuable resource in studies, however, it raises numerous privacy concerns about how that data will be used in the future. Swift and Weidemann (2013) attempted to address these problems by creating an ArcGIS add-in tool called Twitter2GIS that converts geolocated tweets into geospatial data. The tool also had the ability to geocode tweets from a specified region. From this tool, 15 million total tweets were collected based on all hashtags. About 3.5% of this information had precise coordinates and 23.5% could be located down to street level accuracy. 2.2% of the tweets had enough information from the Text to geolocate. Because this tool uses the legacy API, the tool was not available for access when the authors were contacted.

Liu et al. (2019) explored the uncertainty of what type of activity a user is engaged in using digital footprints from social media. They employed a multi-scaled algorithm Density-

based spatial clustering of applications with noise (DBSCAN) to build upon previous research. This algorithmic research is essential to understand what people are doing while they tweet and is used in many industry standard products such as recommendation systems in popular content aggregate webpages and mobile applications.

Mahmud et al. (2012) contribute to the examination of inferring home locations of Twitter users using tweets to triangulate their location based on contextual information. They collected tweets from July 2011 to August 2011 from the 100 most populated cities in the United States using a bounding latitude and longitude box. They then selected the 200 most recent tweets from a user from their location. The users were then classified using three techniques. Content-based statistical classifiers, which bases the user's home location on words, hashtags, and place names. Content-based heuristics classifiers, which uses contextual information such as if a user mentions a home state or city more often than other states or cities. And finally, behavior-based time zone classifiers, which uses the time from the user tweet to find their home location. The study created a list of heuristic and statistical classifiers to build upon their accuracy assessments.

Yaqub et al. (2018) used information in the form of tweets gathered during the 2016 United States presidential election to test how sentiment scores varied for each candidate by state and who was mentioned more in subjective tweets. They also attempted to examine how tweets from state locations correlated with real-world sentiment of the public in both candidates. Although this study expressed reprehension in using Twitter data because of doubts about representativeness, Twitter is still used increasingly to post a candidate's opinion and influence traditional media outlets. The research thus used Flask and Python as the backend to create a web application that analyzes the sentiment of users and plots their location on a web map. They also

used TextBlob for the NLP sentiment analysis portion of the application. Tableau was used for the mapping portion of the application. Again, the issue of only 1-2% of tweets having precise geolocation information arose and was circumnavigated by using the location information provided in a user's profile. Some studies cited in the paper have found an 80% correlation between Twitter sentiment analysis and actual public opinion surveys. It was found that Donald Trump had a higher subjectivity score in the top 10 most populous states. This effort could be expanded to divide the United States into rural and urban areas to define location better. This application could be further refined to study other elections.

Hamstead et al. (2018) used geolocated social media data as an indicator of visitors to a park along with deciphering where equal park access took place in New York City (NYC). They did this by geolocating Flickr and Twitter data to NYC parks. Usage rate was specified as Flickr user days and Twitter user days and combed through 51.3 million geotagged tweets from 2012 to 2014. Pictures taken at NYC parks from Flickr were then divided into local residents and tourists. They found that 76% of photos were from people residing outside of NYC. They focused on three predictors for park visitations. These included park facilities, accessibility, and neighborhood characteristics. Accessibility was defined as parks within a ten-minute walk from a user's location. Pearson's correlation coefficients were used to validate the regression model that tested whether neighborhood characteristics explained park visitation. They found that 95% of parks had 0-13 user days, seven parks had 537-894 and six had 7,291-50,384. They used 35 explanatory variables to perform backward stepwise regression. Wi-Fi was a positive predictor and might indicate that people go to parks to find free Wi-Fi, or that they upload more when free Wi-Fi is available. A racial divided was also discovered with more Whites located near large

parks while Hispanics and Asians located near smaller parks. This study provides an excellent framework of explanatory variables to run statistical tests.

CHAPTER 3

METHODS

This chapter outlines the methodological techniques employed to use a lexicon-based approach to classify massive amounts of text data into positive, negative, and neutral categories. Section 3.1 provides for an outline of the area of interest, the lower continental 48 states of the United States. Section 3.2 describes how we scrapped Twitter tweets from the website using a web crawler designed using Python libraries. Section 3.3 describes how we accessed profile location information from the website. Section 3.4 introduces the computing environment used to scrape and geocode the tweets. Section 3.5 details how the *geocoder* Python library was used along with a *nltk* implementation of the VADER sentiment analysis algorithm to classify sentiment. Section 3.6 outlines the steps taken to analyze the geocoded and classified sentiment scores and the performance of the model using accuracy assessment equations.

3.1 Study Area

The study area focuses on the conterminous United States of America, consisting of the lower 48 states and ranging from the Northeast corner at 44°48′55.4″N 66°56′59.2″W to the Southwest corner at 32°32'20.8"N and 117°7'16.2228"W (Latitude and Longitude Finder, 2019). The lower 48 states have a combined area of 7,663,942 square kilometers (Central Intelligence Agency, 2018). There are large urban collections throughout the Eastern and Western portion of the United States. It is heavily populated on the coastal regions (Central Intelligence Agency, 2018). It is the third largest country in the world in terms of both population and country size.

3.2 Twitter Web Crawler

The first step was to set up a web crawler to use customized functions to access Twitter's API. Since the programming language Python is widely accepted in research studies due to its ease of readability and reproducibility, we built the web scraper using these programming guidelines. An example of my workflow can be described in Figure 1. The script first imported the module *tweepy*, which is a Python library sponsored by Twitter that can access the API. Other libraries imported were *os* and *sys* for basic operating and system functionalities as well as the library datetime which kept track of when and how long the operation took place. The file name created by this script included the date and time of the first tweet captured and was saved with the extension .txt. The actual data captured was saved as a JavaScript Object Notation (json) file, which is a datatype that is an extension of JavaScript. Next, the script needed to authenticate with the API that it was indeed registered with Twitter. This required the creation of a custom application on the Twitter Developer website, which provided access credentials, including access_token, access_token_secret, consumer_key, and consumer_secret. We also created a .log file that logged when an error was found for debugging purposes. Next, a class called MyListener was setup to write each received tweet to a file. The limit on this file was for 10,000 tweets per file in order to make the file size manageable for future processing. We used functions to help define code that would be reused within the script. We used two functions to define how the tweets were written to a file. A third and fourth function were created to log the exceptions when MyListener raised an error. To avoid the interruption of script execution due to rate limits, we created another function to close the current python file and open another duplicate python file to resume data collection. Finally, the stream was filtered along 13 hashtags predefined as

being relevant to energy consumption. These hashtags were chosen after careful deliberation of the most popular energy related hashtags on Twitter.

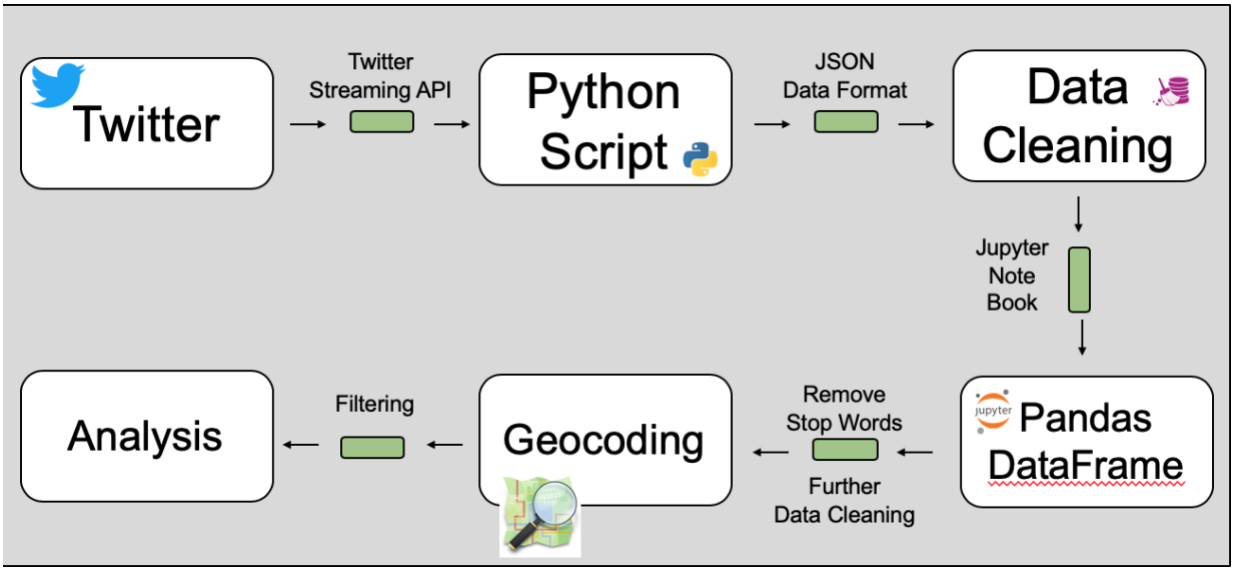

Figure 1: A chart of my workflow.

Table 1: Hashtags for my focused energy types.

Energy Type	**Search Keywords**
Coal	Coal, coalmine
Nuclear	Nuclearenergy, nuclear
Wind Energy	WindEnergy, windfarm, WindTurbine, THE
Solar Energy	Solarfarm, solar, SolarPower, SolarEnergy, SolarPanels

3.3 Geocoding Twitter Users' Locations

This section details the first half of the geocoding script. Since we had such a small number of tweets with precise geolocated coordinates, we needed a method to assign a general location to the tweets. Twitter profiles allow the inclusion of a city and state name in the location field, which can be accessed via Python scraping. These named places can then be sent to a geocoder to collect geolocation coordinates of the named place. In order to complete this task, we needed to set up the script environments. Next, the appropriate libraries for the script were imported. Then, the location profile scraping data structure was setup to capture the relevant location information. The location information from the scraped profile was included in the data structure. Finally, the data was saved as a newly created json file.

3.3.1 Setting up the Geocoding Script

A problem emerged when attempting to geolocate the tweets using the social media website's built-in geolocation feature. Since this feature needs to be activated by the user, only about 1-2% of tweets contained precise coordinates (Crampton et al., 2013). This meant that of the roughly 6,000,000 tweets scraped, only about 60,000 tweets had precise geolocation information. Since the web scrapper included 53 hashtags, this meant that each hashtag only had about 1,132 tweets. To remedy this lack of location data, a second script was written in Python and was designed to access user specified location information located on a Twitter user's profile. This script was written for Python version 2.7. Since we were using the Anaconda Python environment, we needed to first activate this version by specifying "source activate

py27". We also needed to place the list of json files, each containing roughly 10,000 tweets, into the same directory as the Python script file.

3.3.2 Importing the Libraries

The script itself first imported *json* and *os* libraries to read the json file and run rudimentary operations. The libraries *pandas* were imported for reading additional data and *glob* was imported for creating lists of files. The library *geocoder* was imported as a wrapper for accessing Open Street Map's geocoding API known as Nominatim. Multiple alternative geocoding APIs, such as ESRI or Google Maps, are available within this *geocoder* library; however, they either required a paid account or were subject to daily limits. Open Street Map's Nominatim is also rate limited to one geocode per second, however, there is no limit to the number of geocodes per day. Next, the libraries time and datetime were imported to both keep track of the amount of time it was taking to geocode and to limit the *geocoder* API to one second. The libraries *collections*, *numpy*, and *unicodedata* were imported to do some basic arithmetic on the *pandas* dataframe. Finally, stopwords and subjectivity were imported from *nltk.corpus*, and *SentimentAnalyze*r and *SentimentIntensityAnalyzer* were imported from *nltk.sentiment* and *nltk.sentiment.vader*.

3.3.3 Setting up an Iterative Loop and Collecting Data

The next step of the script required the path of the current directory to be discovered and a list created with the library *glob* to grab all of the json files, demarked by the file ending with .txt. A loop was then setup to loop over the list of files in the directory. Each file was opened using the "with open" denotation and a json datatype list was created to hold the features. A nested loop was then inserted to loop over each line in the json file, loading each line into a

variable named “tweet”. Next, an if statement to test if the concatenated tweet’s user and id existed was established. We gave each tweet’s users some data to find them by. This was done by creating a user_data dictionary that included the user_id and other field such as name, id, screen_name, tweets, location, text, and created_at. A series of if, else if, and else statements were then setup to scrape the tweet coordinates, tweet place, or User location and append them to the data structure. This created an output list of json files in the hundreds. This portion of the script was adapted from a blog by Mikael Brunila (Brunila, 2017). Finally, the amount of profile location information scraped was printed out as a percentage and the file was saved with the json extension.

3.4 Geocoding Process

Open Street Map’s open source geocoding API Nominatim has a download rate limit of one per second. Since we had nearly 6,000,000 tweets to geocode, it was essential that we took advantage of multiple computers located in Faner Hall of Southern Illinois University Carbondale. Thus, two PCs had the appropriate Python libraries installed on their version of Python 2.7 and the scripts were run overnight. Two Linux terminals also had the same Python libraries installed and the scripts again ran overnight. Because of starts and stops of the Python script due to network interruptions, this took roughly 12 days to geocode the tweets. The process was repeated twice due to errors on the original script appending the geocoded location information to the wrong tweet.

After the profile location information was scrapped from the first half of the script, this location data was then sent to the *geocoder* API of Open Street Map’s Nominatim. This began the same way as the first half of the script. The path was generated and *glob* was used to grab all

the files with location information and tweets with the ending json. Another loop was setup to iterate over all of the files in the directory. Next, *pandas* were used to read in each file and create a *pandas* dataframe. Since the features were oriented incorrectly in the aforementioned 'data' dictionary, it was necessary to correct this with a few lines of code.

A nested loop was setup to iterate over each index in the *pandas* dataframe. This loop had two portions: a try and an except. In the try portion of the loop, we printed the location name from the tweet's profile, delayed the script by 1.01 seconds, initialized the *geocoder* Open Street Map function on the 'location' field, and found the *geocoder's* latitude and longitude. Finally, we printed out the newly created coordinates for the location field called 'geo' and appended the 'geo' variable using the panda's library 'at' function.

3.5 VADER Sentiment Analysis

The next step was to split the newly geocoded coordinate column into 'x' and 'y' columns to ease the processes of plotting of the tweets in a Geographic Information Systems (GIS) desktop application. This first required the coordinate numbers to be converted to a string. Then, the function 'strip' was employed to remove the brackets enclosing the newly converted string column. Two columns named 'x' and 'y' were created and the built-in 'split' function was used to split the coordinate column along the comma character, creating two new 'x' and 'y' columns. We then printed that the data had been geocoded. Next, since we were receiving an error when attempting to remove the stopwords, it was essential to specify that the 'text' column of the *pandas* dataframe was 'notnull'. Stopwords include words that are not necessary to derive the context of the sentence. Stopwords from the *nltk.corpus* stopwords library were used by way

of a function and established a dictionary of stopwords to remove from the text corpus. We finally set up a code to remove words like 'The', 'RT', '&amp', '-', 'https:', '.', and '@'.

The final portion of the geocoding script involved setting up the compound, negative, neutral and positive columns to place the appropriately classified tweets into each column. Next, the SentimentIntensityAnalyzer function was set as a variable named 'sid'. A loop that iterates over each tweet was then created with a try-except structure. In the try clause, the variable 'sentence' was set to normalize the newly created column 'tweet_without_stopwords' along Unicode data standards. The variable 'ss' was created to hold the polarity scores of the sentence. Next, the columns named compound, negative, neutral, and positive were populated with the appropriate scores. In the except clause, we printed out the user and subsequent 'tweet_without_stopwords' variable as a sanity check when the sentiment analyzer failed. Finally, the data file that was analyzed was printed out and saved as a json file.

3.6 Accuracy Assessment

The data was contained in hundreds of json files containing 4,000 to 5,000 tweets per file. To ease the analysis process, these files were concatenated to create one single file. This was completed by utilizing Python's Jupyter Notebook environment. First, the libraries *os* and *json* were imported to read json files and perform rudimentary operations. Next, *pandas* and *glob* were imported to create dataframes and create file lists of directories. The code that ran was setup to specify the path of the current working directory. *Glob* was used to make a list of all the files ending with json. *Pandas* was used to read in each file using a loop and each *pandas* dataframe was concatenated to form a single dataframe. Finally, the tweets index was reset and subsequently saved into a single file.

In an attempt to see how accurate the VADER sentiment analysis algorithm was at classifying tweets, an accuracy assessment was introduced. We used a random sampler to create a dataset of 100 tweets. Next, the categories True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) were set up to see if the hand label category matched with the compound score provided by the VADER sentiment analysis algorithm. Once totals for each category was tallied, the following equations were used to achieve accuracy, precision, recall, and F1 score (see e.q. 1-4).

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \qquad \text{e.q.1.}$$

$$Precision = \frac{TP}{TP+FP} \qquad \text{e.q.2.}$$

$$Recall = \frac{TP}{TP+FP} \qquad \text{e.q.3.}$$

$$F1\ Score = 2 * (Recall * Precision)/(Recall + Precision) \quad \text{e.q.4.}$$

CHAPTER 4

RESULTS & DISCUSSION

Chapter 4 examines the results and discusses their implications. Section 4.1 begins by assessing the sentiment analysis results of coal, solar, nuclear, and wind energy hashtags. This section includes an accuracy assessment for coal and solar. Word frequency is examined in section 4.2 for these energy types. User frequency for the four energy hashtags is then examined in section 4.3 to identify power users in the "Twitterscape". Section 4.4 then delves into the frequency of location for each of the four energy types. Section 4.5 constructs a time-series graph to display the temporal dimensions of the data and we also create a kernel density map of each energy type with an overlaid choropleth map to decipher patterns of positive and negative perception. Section 4.6 displays the result of our topic analysis using word frequency charts. Finally, section 4.7 discuss the policy implications for lawmakers and entrepreneurs.

4.1 Overall Social Sentiments on Energy

Careful consideration was placed towards generating sentiment scores for four types of energy hashtags via a lexicon-based approach to sentiment analysis, as described in Hutto and Gilbert (2014), Adeoyin-Olowe et al. (2014), and Chen et al. (2016). We used a python implementation of VADER, which is found in the *ntlk* python library. These hashtags included solar, which contained 22,219 filtered tweets, coal, which contained 5,625, nuclear, which contained 23,843, and wind, which contained 10,269. As part of the exploratory data analysis steps that took place towards the beginning of the study, we constructed histograms ranging from -1 (negative sentiment) to +1 (positive sentiment). It was found that each of the energy types is

dominated by neutral tweets, characterized by the sentiment score of 0. This may be because the sentiment analysis algorithm we employed defaults to neutral when none of the words available in its lexicon appear in the analyzed social media text.

4.1.1 Social Perception on Solar

In an attempt to quantify the perception of the public toward solar panel generated power in the United States, the mean and median sentiment scores were generated for the hashtag of solar. The overall mean on Twitter was .20 while the median was 0. It is suspected that the cause of the median being neutral was because VADER errors towards categorizing text as neutral when no keywords are found to classify. Since solar, wind, and nuclear are some of the primary sources of non-fossil fuel production of energy in the United States, it is helpful to see a breakdown of each state's percent of electricity produced by non-fossil fuel energy sources (Figure 2).

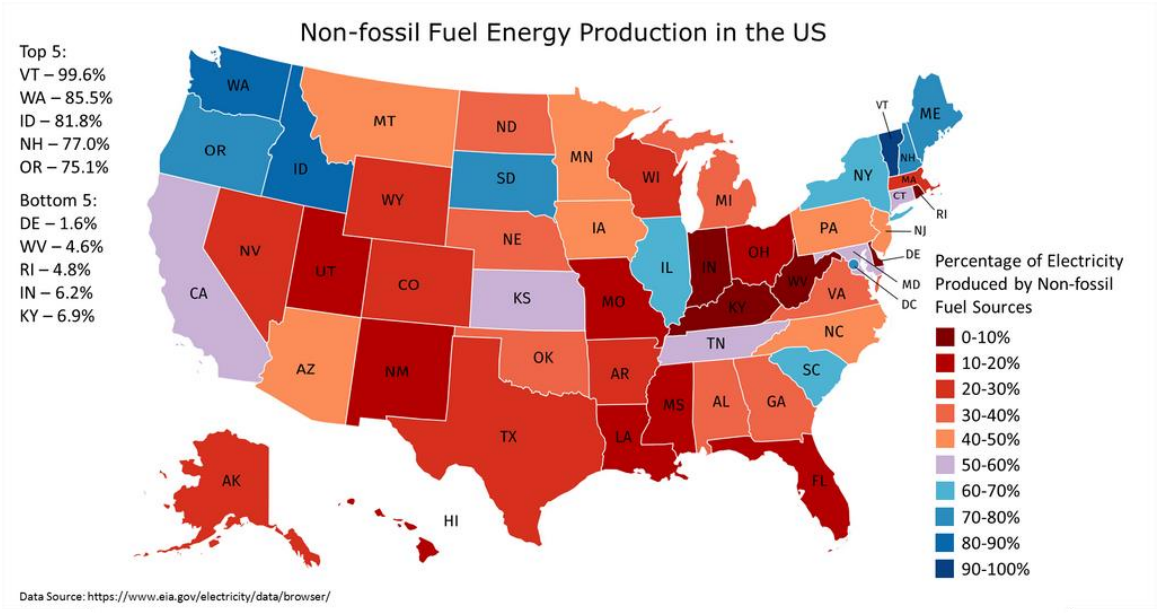

Figure 2: The percent of electricity produced by non-fossil fuel sources.

4.1.2 Social Perception on Coal

Although many states from Figure 3 display areas that use coal or other non-renewable resources, there are still multiple states in the Northwest and Northeast that use alternative resources, such as solar, wind or nuclear, to get their electricity. Conversely, states like Delaware, West Virginia, and Kentucky all have almost no electricity produced by methods besides fossil fuels. These locations will specifically be interesting to view the sentiments on the various energy types. The mean score for all tweets about coal was 0.07 while the median was again 0. This indicates a generally more negative score towards coal than solar.

4.1.3 Social Perception on Nuclear

Nuclear was slightly less in sentiment than coal. The mean sentiment score for this data was 0.05 while the median was 0. This is again because of the large amount of neutral sentiment scores given by the VADER sentiment analyzer. From these scores, it can be thought that the sentiment of Americans on Twitter is mixed at best. This is most likely because of the perception of the danger of nuclear meltdowns, as happened at Chernobyl and Fukushima. There is also the issue of storing the waste products of nuclear energy after the core has been spent.

4.1.4 Social Perception on Wind

There is a large number of neutral tweets for wind energy because of how the VADER sentiment analysis scores tweets. The mean was the second highest score out of the four energy types at .18 while the median was 0. Since very few people had negative things to say about the wind energy type on Twitter, it can be thought that users view wind as being a positive energy

type. Reasons why Twitter users think of wind as negative is because of the intermittent nature of the energy collection. When the wind is not blowing, energy is not being produced, which disrupts usage on America's energy grid. There needs to be a way to efficiently store wind captured energy in batteries from times of extreme wind energy production so that the energy stream can continuously serve the American people.

4.2 Energy Topic Analysis

There are more advanced algorithms to extract topics from text corpus, however, algorithms such as doc2vec and LDA were not employed in this. From these methods, LDA processed words are weighted by way of graphing similar meaning words in vector space and discovering how often they appear in the text corpora. My preliminary analysis found that both methods were incapable of producing meaningful clusters of topics. Instead, term frequency was employed to count how many of the same words appear in the text corpora. From this, topics are exposed and can be further analyzed to decipher the overall meaning of the topics.

4.2.1 Coal Topic Analysis

Word frequency was used to extract the main categories of the four datasets. After careful evaluation and further removal of stoppage words, these main categories for coal can be viewed in Figure 3. An interesting phenomenon extracted from using this approach of topic analysis emerges from examining how often Trump is mentioned in the tweets about coal. An example of one of Trump's tweets can be examined in Figure 4. Although not appearing in the top 10 most frequent words, he does appear in the top 50. This example is included to illustrate how divisive a single topic word like #coal can be using the data collection techniques. Instead of a black and

white blanket statement about how coal is positive or negative, there instead emerges a spectrum of views on the subject from various organizations and leaders.

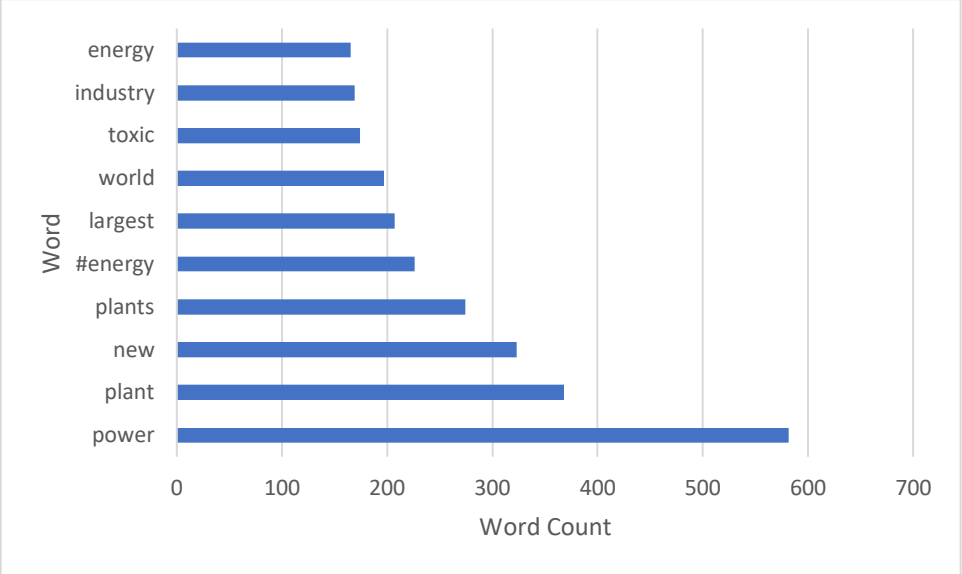

Figure 3: The coal word frequency chart.

Figure 4: An example of Trump's tweet regarding coal.

4.2.2 Solar Topic Analysis

The same process was applied to the solar dataset in an attempt to extract meaning from the tens of thousands of tweets. To do this, we counted how often the dataset had various keywords. This word chart can be viewed in Figure 5. As with coal, specific topics were deciphered from this word chart. One of the most tweeted about subjects was energy, which

indicates the data collection techniques were correctly collecting data about solar energy related tweets. Another frequently tweeted about topic was about wind, which indicates that renewable energy was typically included in tweets about solar energy. Jobs was also found quite frequently in the dataset which indicates solar energy tweets included figures about how many jobs were being created or disrupted because of solar energy. Finally, user @mikehudema was included frequently in the top words for solar. This could be because he is a renewable energy advocate for Greenpeace who consistently tweets about energy related topics. An example tweet about solar energy and the jobs created by this technology is included in Figure 6.

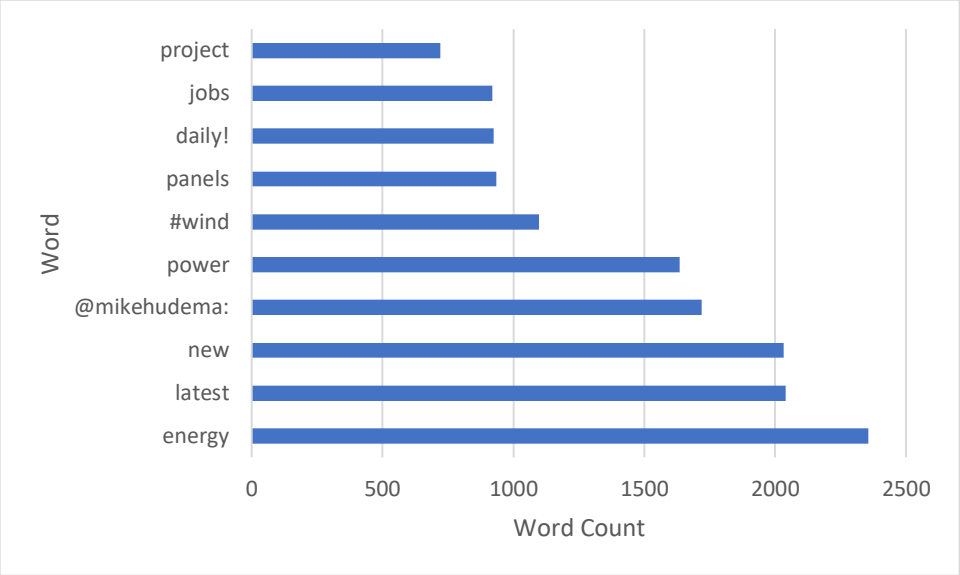

Figure 5: The solar word frequency chart.

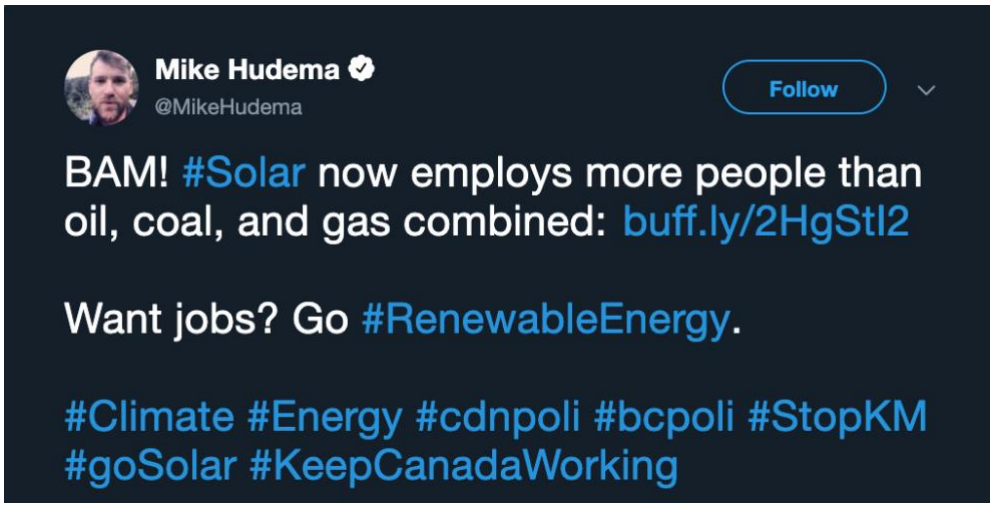

Figure 6: A tweet by user @mikehudema about jobs in solar energy.

4.2.3 Nuclear Topic Analysis

The most frequent words were again counted for the nuclear hashtag and filtered by removing the stop words or other irrelevant words from the dataset. This chart is displayed in Figure 7. The most frequently used word from this dataset was for the user @NEI. Located in Washington, D.C., the NEI is an account for the Nuclear Energy Institute. This is an account that wishes to spread information about the benefits of using nuclear energy to the English-speaking world. An example tweet from them is found in Figure 8. Another frequently tweeted about user was @nuclear_matters, a national coalition that was setup to create tweets about nuclear energy and the benefits therein. Although energy was a frequently tweeted about word, indicating that most tweets were about nuclear energy, there were also outlier words. These included words like weapons, Russia, and North Korea, which indicates that some of the tweets were actually about nuclear warfare and not nuclear energy. Data filter techniques should be developed to further clean the dataset to get the information desired.

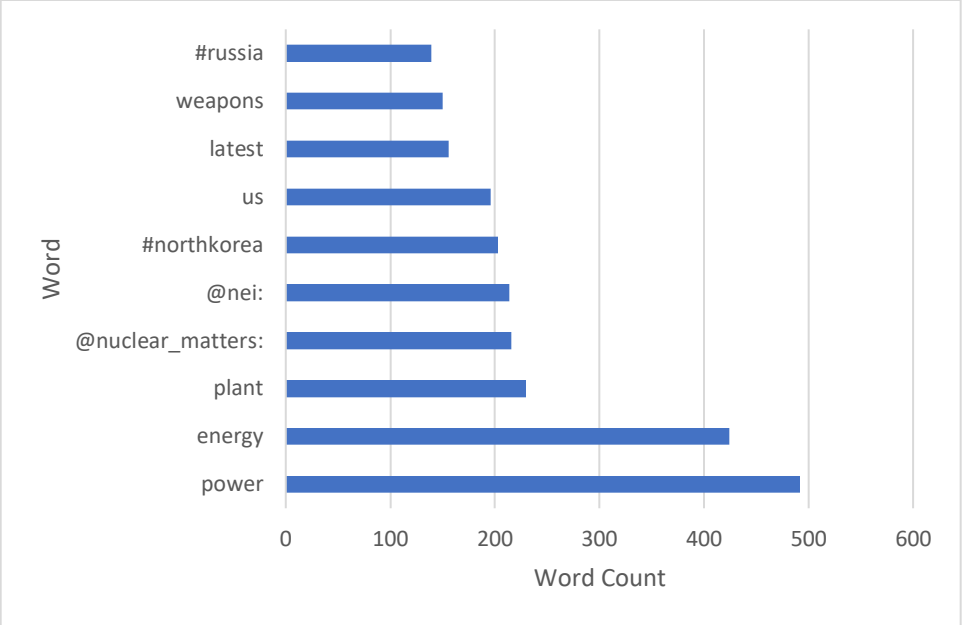

Figure 7: The most frequent words from the nuclear dataset.

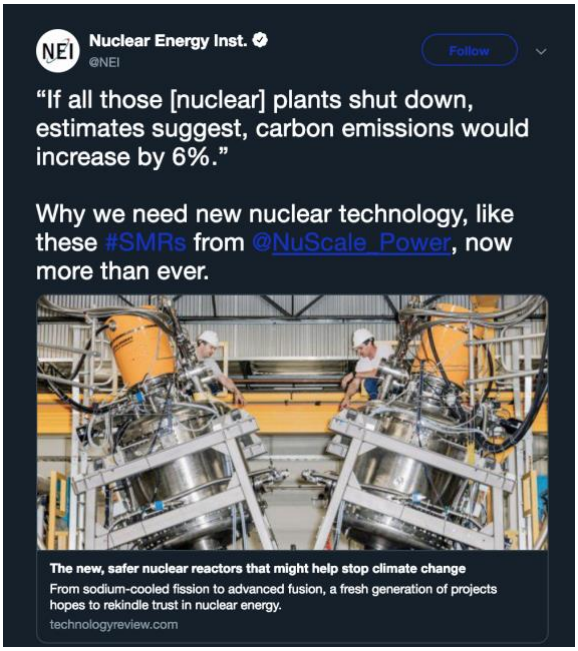

Figure 8: An example tweet from user @NEI.

4.2.4 Wind Topic Analysis

The most frequent words were again counted for the hashtag of wind energy type. The stop words were again filtered, and irrelevant words were removed from the dataset. The chart

can be viewed in Figure 9. Two of the most frequently mentioned terms was ‘windfarm’, which is indicative of tweeters writing about new installations of wind power plants, along with ‘solar’, which is indicative of tweeters writing about renewable energy types.

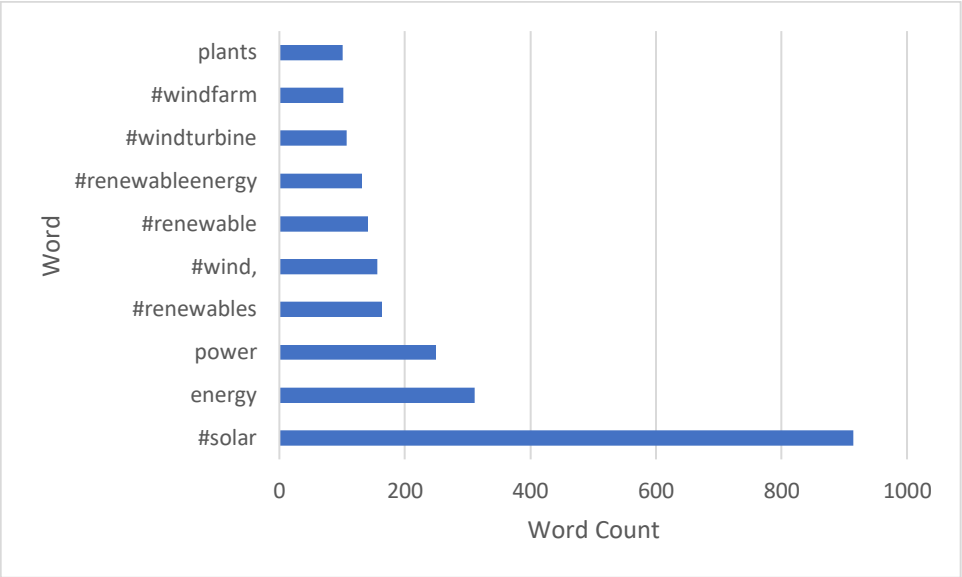

Figure 9: The most frequent words for the wind energy type.

4.3 Who tweeted about Energy?

In an attempt to identify power-users within the dataset, the top tweeters were counted using a Python script and subsequently plotted in Excel. Power-users were further broken down in the dataset by combining all of the tweets by a single user and counted it as one tweet. The sentiment score for the power-user was estimated using both mean and median sentiment scores by combining all tweets from an individual user. This limited the influence of the robots that were consistently tweeting positive or negative tweets.

4.3.1 Who tweeted about Coal?

One of the top tweeters in the coal dataset was the user @BeyondCoal, with 32 tweets (Figure 10). This account, run by director Mary Anne Hit, is an attempt by the Sierra Club to get countries to move beyond using coal as their primary energy source. User Bwillisful also had 32 tweets from this time and is an account for the Beyond Coal's press secretary, named Brian Wilson (Figure 11). This account also tweets and retweets about using alternative, renewable energy sources.

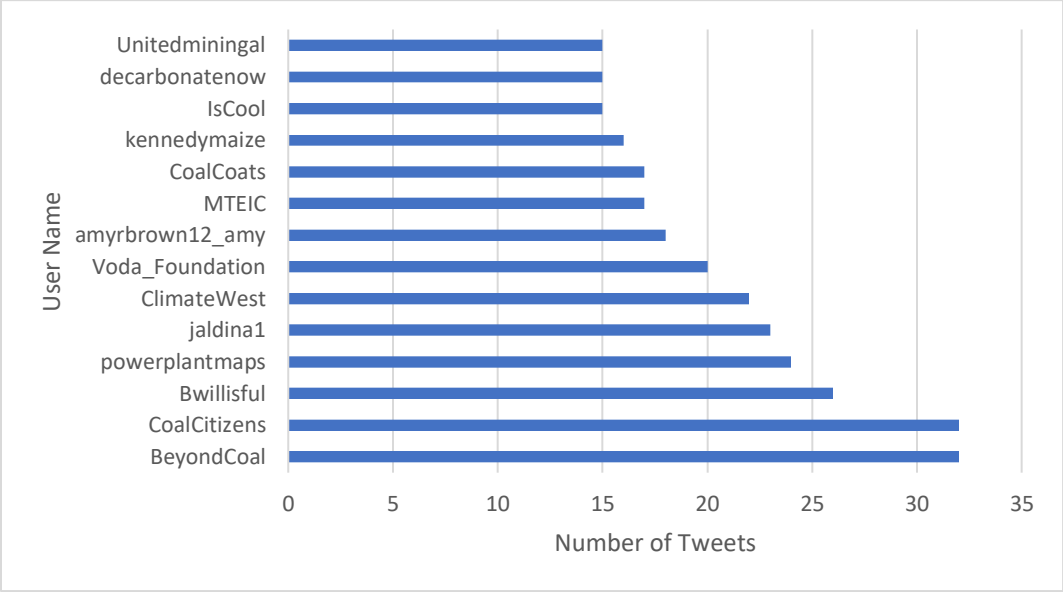

Figure 10: The top Twitter users from the coal dataset.

Figure 11: An example tweet from @Bwillisful.

4.3.2 Who tweeted about Solar?

To further identify users within the solar portion of the dataset, the amount of times each user tweeted was again counted using a Python script and plotted the data in Excel. The top user that tweeted the most about #solar was user @solargenerator2, with 145 tweets over this time period (Figure 12). An interesting aspect of the top frequently tweeting user named @solargenerator2 is that when the account is searched for using Twitter's search bar, nothing comes up. However, when the user is searched for using Google, the account pops up with a warning that the user's content has been temporarily restricted. From my own investigation, it appears that the user is a bot that tweets anything related to solar energy. This is evident because of the massive amounts of tweets per day that the account is putting out. An example tweet about solar is included in Figure 13.

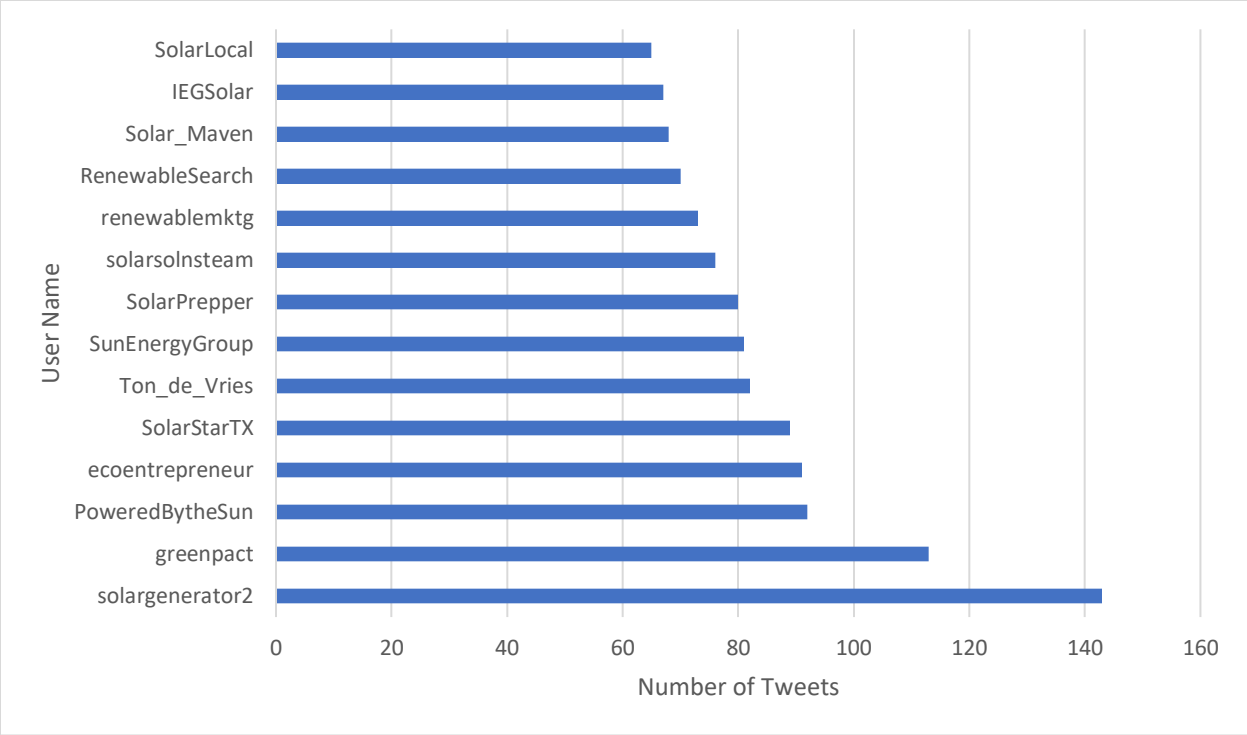

Figure 12: The top power users for the hashtag solar dataset.

Figure 13: A recent tweet by @solargenerator2

4.3.3 Who tweeted about Nuclear?

To continue to examine the top users within the dataset to discover power users, the users screen names were again counted for the nuclear energy hashtag. This is included in Figure 14. The top user from this time period was @discardedbacon, a Japanese account for a person named Kiyoshi Hara. Although most of this user's tweets were in Japanese, their actual location was from New York, New York. An example tweet from this user in English was, "Fukui weighs new wave of reactors to protect status as Japan's '#nuclear capital' (1/21 JapanTimes) https://t.co/Q9IfgaQHrP #jishin_e," which received a score of .3818 and indicates a positive tweet. The next most tweeted profile was @nuclearwebinfo, an information site located at North Myrtle Beach, South Carolina. An example tweet from this user links to Nuclear Marketing Daily, an SEO for digital marketing that includes Nuclear energy. Since this tweet was largely thanking the people that work there, the tweet received a score of 0.4926.

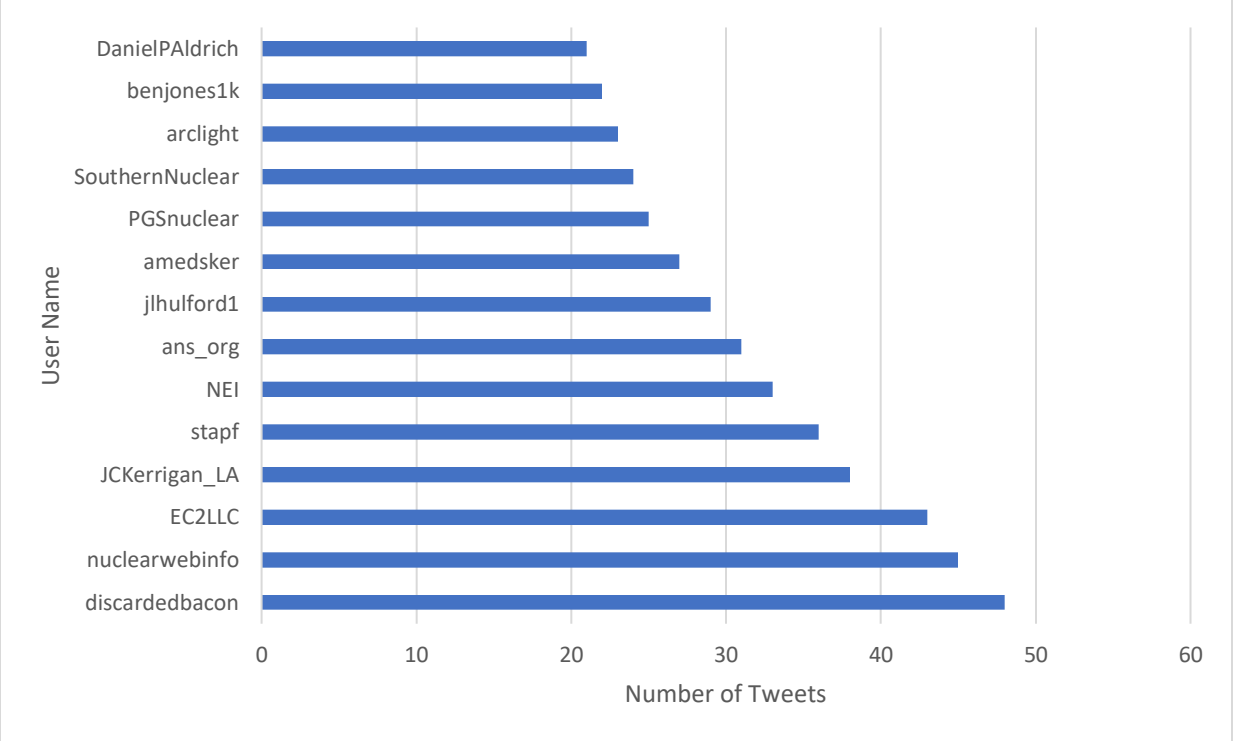

Figure 14: The user frequency chart for nuclear energy.

4.3.4 Who tweeted about Wind?

The same methodology was applied to the wind energy dataset to extract the top power users from the data. The top users are included in Figure 15. The top power user from this hashtag was @renewablesearch. Located in Hartford, Connecticut, this account was started by Tom DeRosa, the CEO of the same company. This is a company designed to spread news about Cleantech, especially for energy storage, solar, and wind. An example tweet can be found below in Figure 16. Another top tweeter was @revolutionsolar, an account for Ken Bradley from Minneapolis, Minnesota. An example tweet from this user was, "Broadcasting Bradley is out information about #solar, #evs, #energy, #wind, #energyefficiency! Thanks for your work!" which received a sentiment compound score of .5399, indicating it was a very positive tweet about #wind.

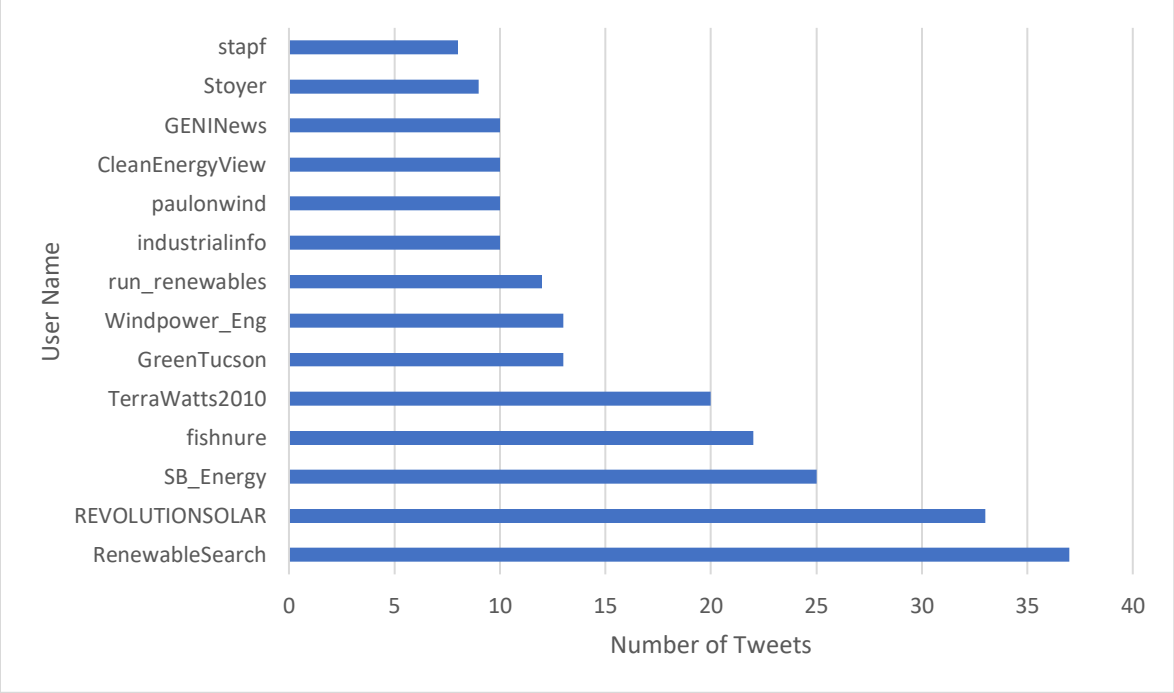

Figure 15: The user frequency chart for wind energy type.

Figure 16: An example tweet from RenewableSearch.

4.4 Spatial Distribution of Energy-related tweets

The next aspect that was examined was how often a particular city was tweeting about each subject. This was done by writing a script in python to count the occurrences of each of the four hashtags. It is hypothesized that places such as Appalachian Mountains, particularly in West Virginia, Southern Illinois, and North Dakota would tweet frequently about coal and fossil fuels. This falls in line with Figure 2, which displays that West Virginia has 0-10% of their energy resources coming from non-fossil fuel sources. It is also hypothesized that cities in places like California, Oregon, and Washington would tweet more about Solar and Wind energy types. This is supported by Figure 2, which displays those three states having most of their energy production come from non-fossil fuel sources. Since Illinois is a top nuclear energy producer, it would be expected that tweeters from this state would have more tweets about nuclear energy production.

4.4.1 Spatial Distribution of Coal related tweets

This location frequency chart for the coal topic is included in Figure 17. Some of the top locations that most people were tweeting from were New York, New York, Los Angeles, California, and Chicago, Illinois. It is no surprise that that top three most populous cities in the United States also correspond with the most tweeted from places. An interesting location that was routinely tweeted from was Ely, Minnesota. Although there are multiple copper and nickel mines in the area, there are no coal mines that can be discerned. The presence of this many coal tweets cannot be explained by examining the data.

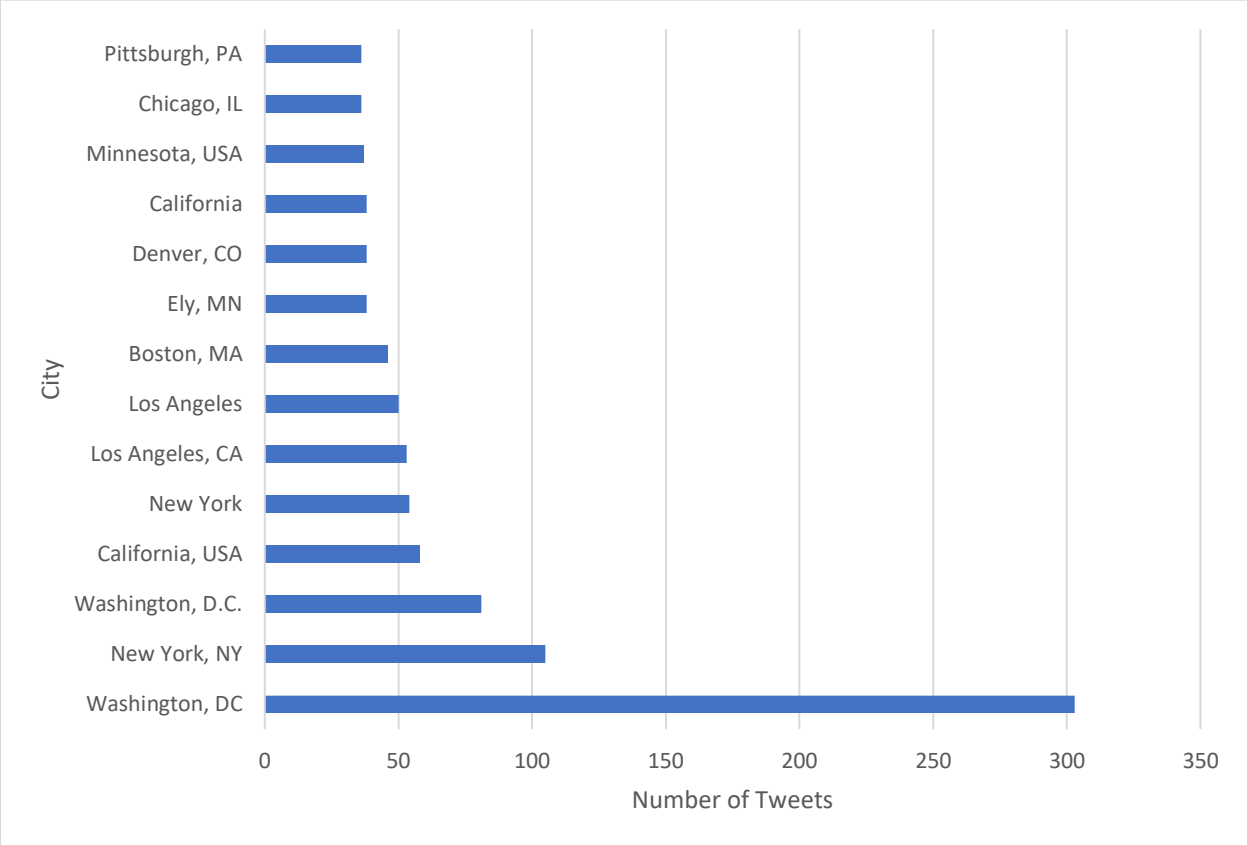

Figure 17: The coal location frequency chart.

4.4.2 Spatial Distribution of Solar-related tweets

The next aspect examined was the location frequency of how people tweeted about solar energy. This is included in Figure 18. Surprisingly, the top three most frequently tweeted from places around the country were Washington, DC, New York, New York, and San Diego, California. Although New York is to be expected because it is the most populous city in the country, Washington, DC and San Diego are 20th and 8th respectively (United States).

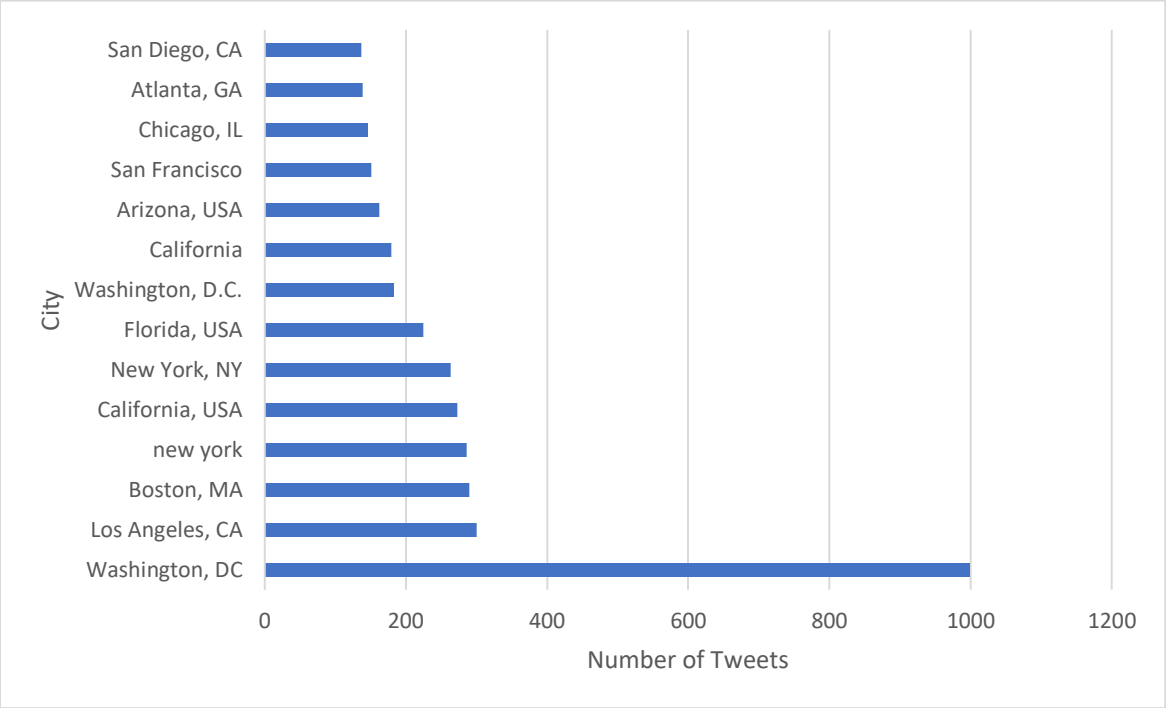

Figure 18: The location frequency chart for solar.

4.4.3 Spatial Distribution of Nuclear-related tweets

The location frequency was then counted for the nuclear energy type. The results are displayed in Figure 19. The top locations that were tweeted from were Washington, D.C., Chicago, Illinois, and New York, New York. Different characterizations of Washington D.C. appear so that the total amount of tweets coming from Washington D.C. is even greater than it appears at first glance. Being the United States capital, nuclear energy is apparently routinely on

the mind of people living in that city. The location marker Chicago, Illinois is the third most populous city in the United States and can thus be explained by that reasoning.

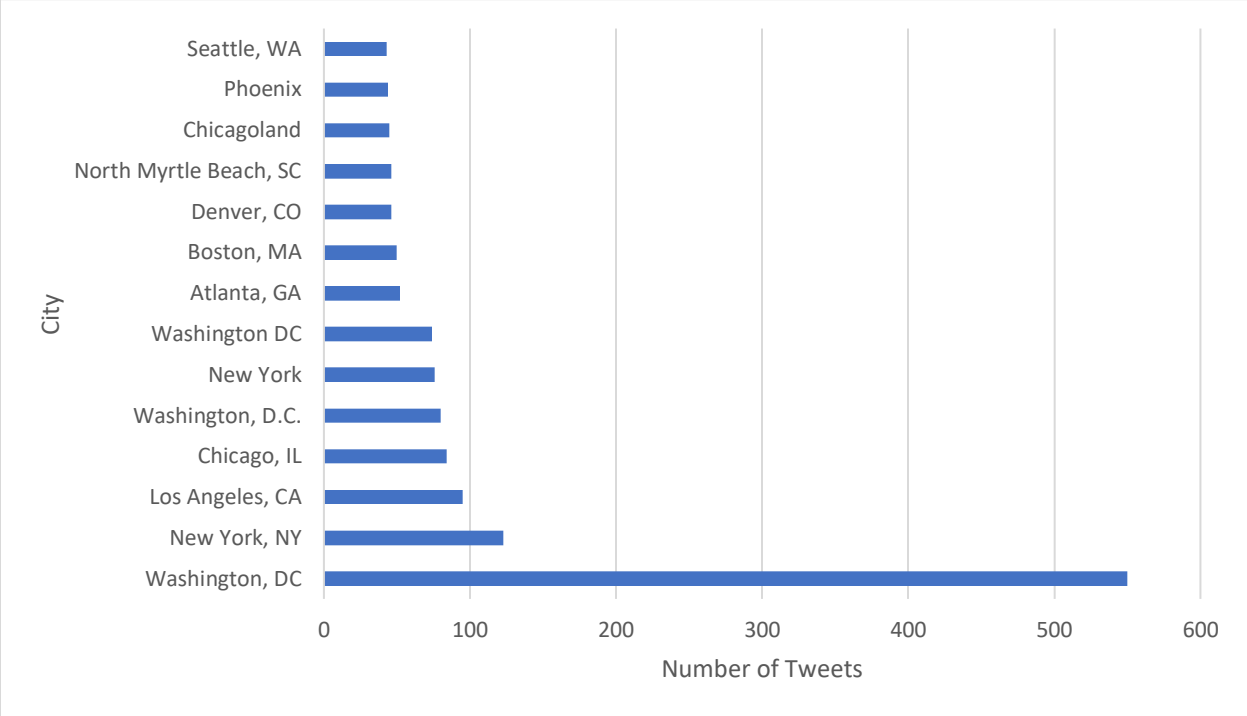

Figure 19: The location frequency for nuclear energy.

4.4.4 Spatial Distribution of Wind-related tweets

Finally, the location frequency was counted for wind tweets. This graph is displayed in Figure 20. The top cities identified from this location frequency chart was New York, New York, Washington, D.C., and Minneapolis, Minnesota. The most populous city in the United States again is New York, thus it makes sense that the most tweets are coming from this area. Washington, D.C. again is the nation's capital and many of our energy related decisions are made there. Minneapolis, Minnesota is a place that generates a fair amount of wind energy and

can thus be explained that way.

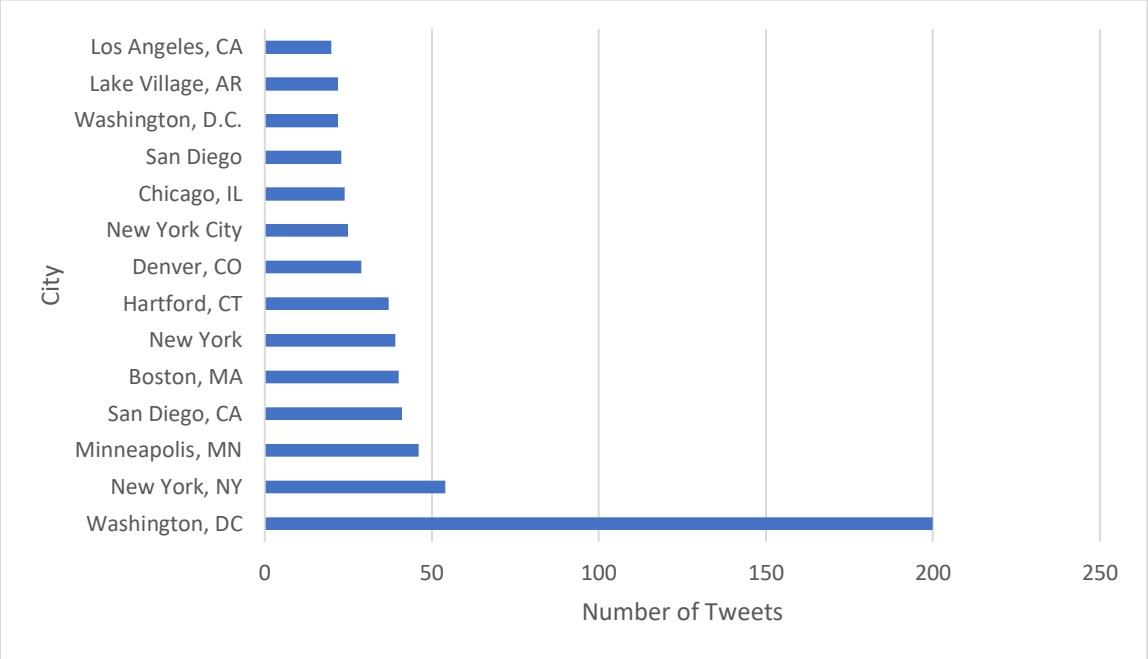

Figure 20: The location frequency for wind.

4.5 Spatiotemporal Patterns of Social Perception

The sentiment of people's opinion toward both solar, coal, nuclear, and wind vary across space. Similarly, the sentiment of people's opinion also varies across time. Thus, it is pertinent to include a time series graph displaying how people's opinion changes over time. This is included in Figure 21. Firstly, it should be noted that the temporal dimensions of this dataset are not complete. That is, there are two significant gaps in data collection: one during winter break for three weeks and one during spring break for a single week. To make a whole time series graph, mean imputation was used to fill in the missing values. As can be viewed over time, the sentiment of solar is almost always more positive than the other three categories. The exception to this rule is during mid-February, when coal sentiment spiked, and in early April when wind spiked. For the most part, nuclear has the most negative sentiment, closely followed by coal.

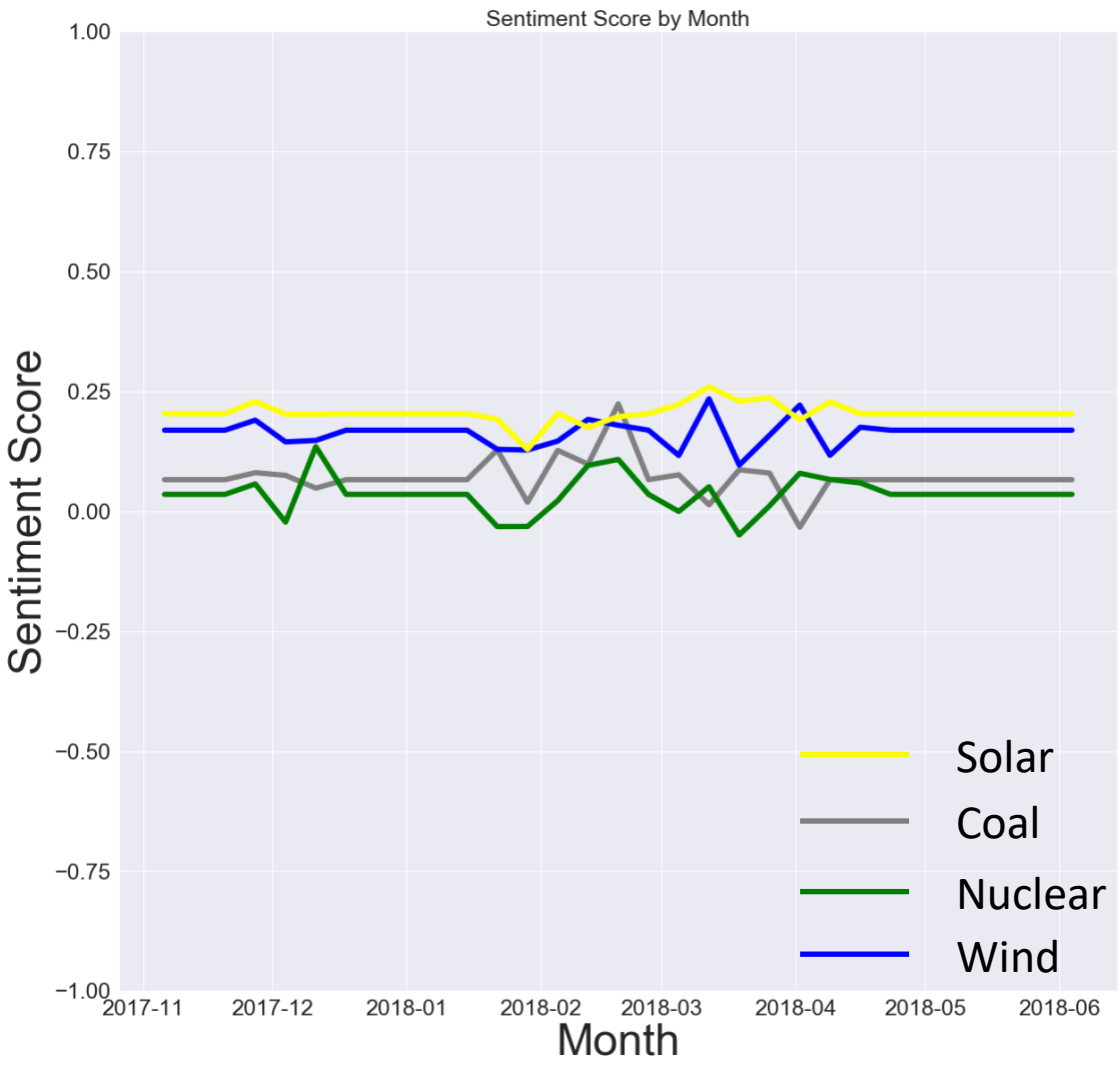

Figure 21: A timeseries graph of solar, coal, nuclear, and wind sentiment.

A possible explanation for the reason coal sentiment spiked in mid-February of 2018 was because of the closing of the Big Brown coal fired plant in Dallas, Texas. Heralded by conservations such as Mary Anne Hit as a step forward in reducing deadly air pollution, the plant was also a large emitter of CO2. Mary Anne Hit specifically tweeted, “Big news that will literally save lives - nation, single largest source of deadly SO2 pollution, the Big Brown #coal plant closed.” This tweet was classified as having a 0.4939 compound score, which is indicative of a very positive tweet.

A possible explanation for the spike in Wind sentiment towards the beginning of April 2018 occurred when user @wradv tweeted, "73% of respondents to the recent @GallupPoll said they preferred an emphasis on alternative energy, such as #windpower and #solarenergy, over fossil fuels. #cleanenergy". This tweet had a sentiment score of 0.2732, which is indicative of a fairly positive tweet. Additionally, this tweet was retweeted three times. The rest of the time series was consistent in the sentiment score. Renewable energy types were consistently higher than nuclear or coal.

Hot spot analysis was originally used to cluster specific areas of positive and negative tweets together spatially. However, this did not work as intended and led to the discovery of erroneous spatial patterns. Therefore, spatial join was used to join tweets to county level data provided by the U.S. Census. These tweets were the summarized using mean and median scores. Finally, kernel density estimation was used as an underlaid map layer to define how frequently various areas were tweeted from.

4.5.1 Spatial Patterns for Coal Sentiments

When the resulting locations of coal or solar are plotted on a map with their sentiment scores included as a choropleth range, no decipherable pattern was observed. Therefore, the use of kernel density was used to examine the spatial density of tweets along with joining the tweets to each county spatially to attempt to explain some patterns. This was done for five months of data. This figure is included in Figure 22. The map has a few patterns consistent with my hypothesis. For example, in the heavily populated (and usually more liberal) areas like the San Francisco Bay Area and the city of Los Angeles, these are extremely negative counties towards coal. Toward more sparsely populated, rural counties in the central valley of California, there are

more neutral and even positive tweets about coal. Surprisingly, much of Oregon and Washington have neutral or positive counties of tweets. In the Appalachian mountain ranges, and specifically in West Virginia, these counties are either neutral or positive towards coal. The Northeast coastal counties seem to be more mixed about coal with positive and negative counties doting the landscape.

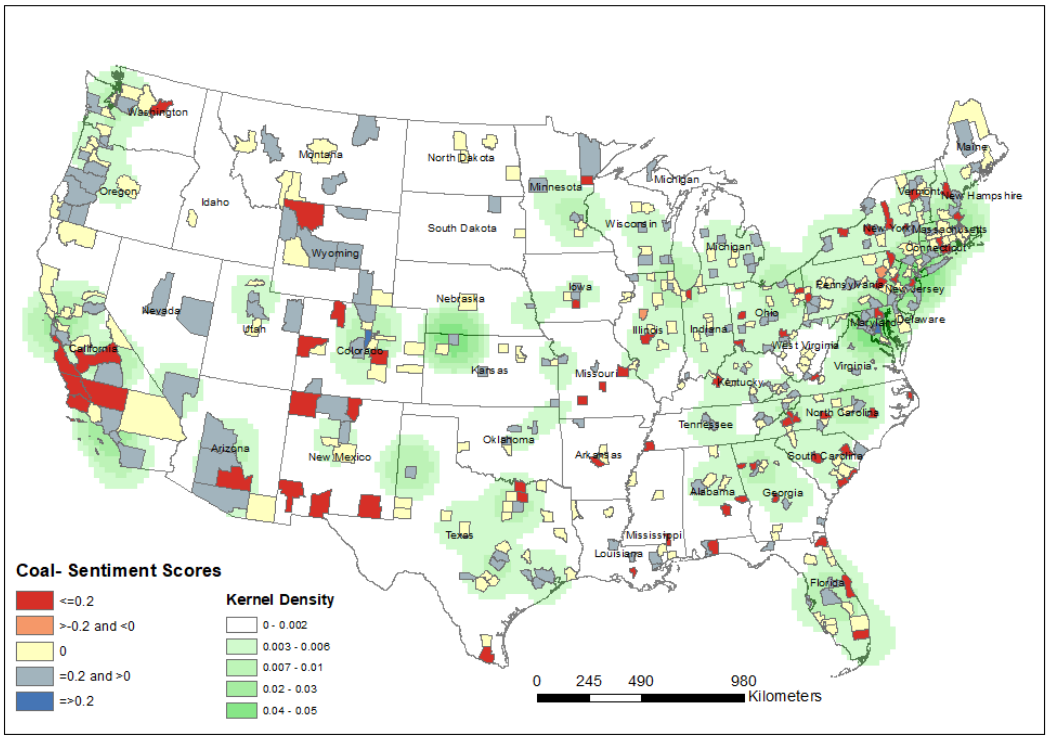

Figure 22: Coal location kernel density and county spatial join.

4.5.2 Spatial Patterns for Solar Sentiments

The same analysis was used to plot the solar sentiment analysis data by employing kernel density and spatial join. Figure 23 displays the sentiment clusters from the sentiment analysis

locations. This map appears to have many more positive counties than the coal map. For example, California can be viewed as being almost entirely populated by counties that have positive tweets about solar energy. The coastal counties of Oregon and Washington are also primarily positive, even though these counties receive very little sunlight. Northeast coastal states also have a more positive outlook on solar with an exception of Maine, which was neutral to negative. Surprisingly, the counties in the Appalachian Mountain range, especially West Virginia, were more positive towards solar.

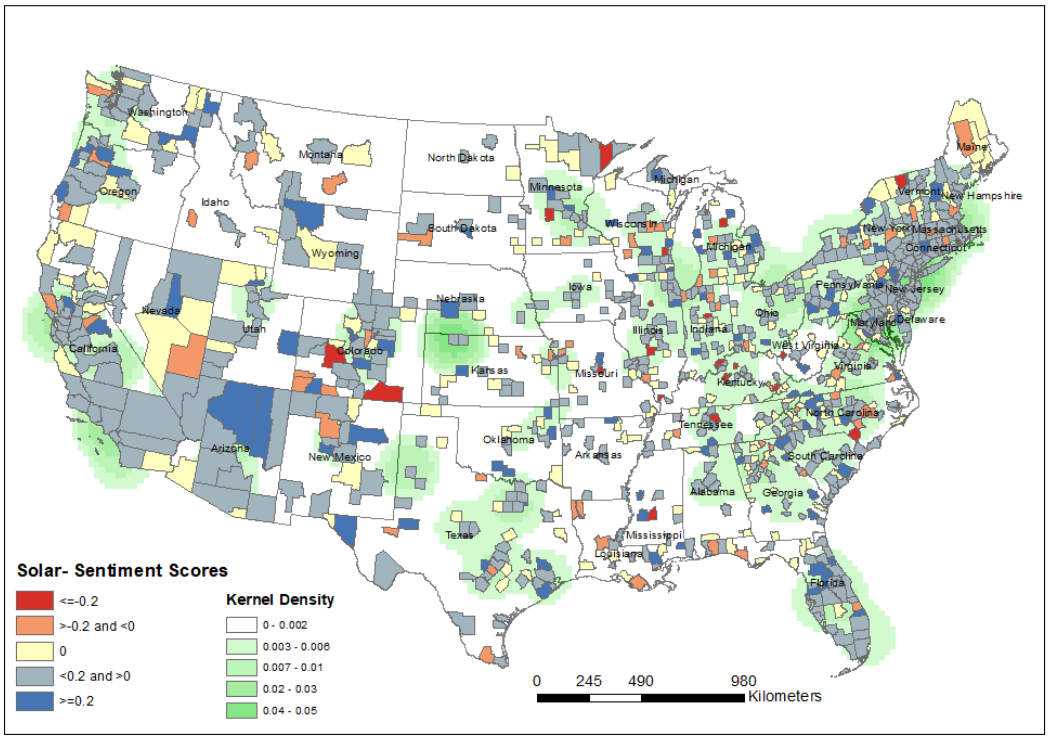

Figure 23: Five months of solar data plotted by kernel density and county spatial join.

4.5.3 Spatial Patterns for Nuclear Sentiments

The same methodology was applied to the nuclear dataset of sentiment in the U.S. and was plotted in Figure 24. Although the overall data is sparser than coal or solar, it can be seen that there are multiple instances of negative counties located throughout the U.S. Again, examining California, it can be seen that most counties are either neutral or negative towards nuclear energy. Counties along the Oregon and Washington coasts were very divisive with either extreme positive or negative sentiments towards nuclear energy. Data was sparser for the Appalachian Mountain range. Northeastern counties were generally negative with a few anomalies.

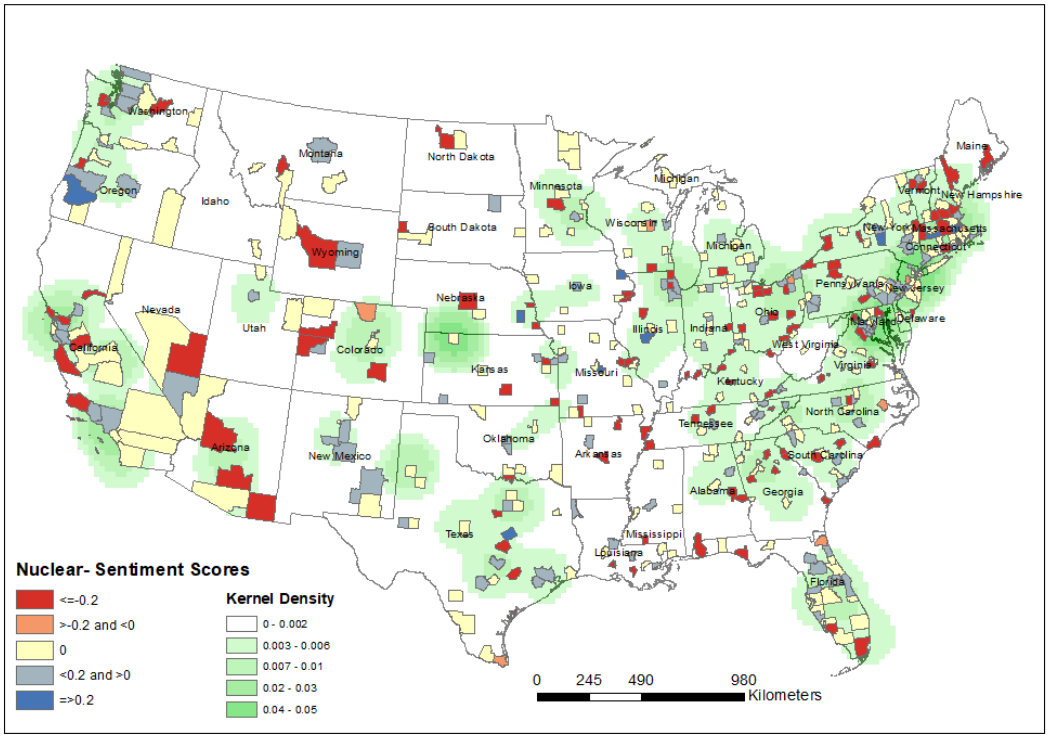

Figure 24: Five months of nuclear data plotted by kernel density and county spatial join.

4.5.4 Spatial Patterns for Wind Sentiments

The final kernel density and county spatial join map was created for the wind dataset. This map is included in Figure 25. The data seems to be generally positive with rare instances of counties with negative posts. California in particular was especially positive towards wind, with only one instance of a neutral county in the Southeastern portion of the state. All the counties in Oregon and Washington that had data for Wind tweets were positive. Data was almost nonexistent for the Appalachia Mountain area. The Northeastern counties were primarily positive with six negative counties interspersed throughout.

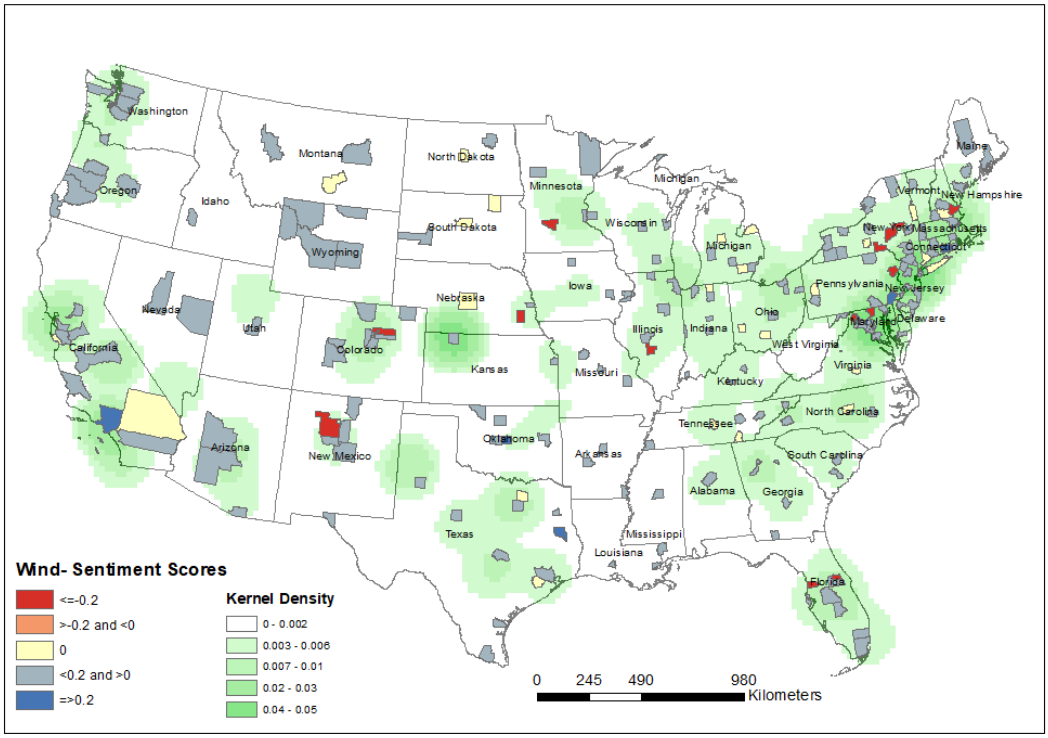

Figure 25: Five months of wind data plotted by kernel density and county spatial join.

4.6 Accuracy and Uncertainties

There are numerous data quality and data uncertainty issues when using social media data. For example, representativeness of the population is difficult because Twitter users are usually younger than the general populace (Longley et al., 2015). Also, geotagged tweets do not follow a random statistical sampling framework. The people that have the knowledge or will to turn on geolocated coordinates is relatively low and represents a specific type of person, one who cares less about privacy. There are clear issues of social media's representativeness (Sang and Bos, 2012; Anstead and O'Loughlin, 2015). Additionally, it has been found that many people describe their location embedded within their profile as nonspecific state or country names, making it impossible to accurately geocode. Finally, the accuracy of using VADER for sentiment analysis, while consistent in how it categorizes text, may be subjective to a human operator.

4.6.1 Geocoding Uncertainties

There are numerous data quality and uncertainty issues present in using social media data to understand social perceptions in terms of spatial patterns. For example, the scrapped user provided location information was used to aide in the geolocation of numerous tweets. There were several pitfalls in using this method, exemplified by how many people used general place names like "Earth", "California", or "Illinois" to describe their location. In order to better understand how often this occurred, it was discovered that 77% of Twitter profile location information had city/state names while 23% had general place names.

4.6.2 Sentiment Accuracy

In order to properly assess the data for integrity, it is essential to establish an accuracy report of the data. From the existing literature, VADER sentiment analysis implementation achieved a correlation to ground truth of 0.881. Overall precision was 0.99, recall was 0.94, and F1 score was 0.96 (Hutto and Gilbert, 2014). To measure the accuracy of the data, one hundred random sample tweets were extracted from the dataset using Python. Each tweet was then hand labeled as either positive, neutral, or negative, based on the contents of the text column. The columns, true positive, true negative, false positive, and false negative were then setup and the totals were calculated by hand. Finally, the equation for accuracy, precision, recall, and F1 Score were specified. The results for these two hashtags are displayed in Table 2.

	Coal	Solar
Accuracy	0.86	0.83
Precision	0.83	0.89
Recall	0.98	0.92
F1 Score	0.90	0.90

Table 2: A table displaying coal and solar sentiment accuracy scores.

While completing the accuracy assessment, it became clear that classifying sentiment by hand is a rather subjective exercise. What might be considered a positive statement for one individual might be neutral or negative and vice versa. Thus, it is essential to rely on a less biased classifier like VADER to classify text because the preconceived biases are known, unlike a human classifier.

4.7 Policy and Entrepreneurial Implications

These massive datasets on the social perception of energy issues provide important information on energy policy. For example, seeing which counties could use informational campaigns on the benefits and drawbacks of renewable technologies such as wind and solar is an important aspect of examining social media data. Similarly, informational campaigns could be targeted at counties that displayed a fondness for nonrenewable fossil fuel technologies. Other policy considerations could examine the backlash of public opinion on the Energy and Natural Resources act of 2017 (ENRA), Energy Independence and Security Act of 2007 (EISA), and Energy Policy Act of 2005. There are a few ways opinion on energy policy could be updated. Recommendations could be developed by amending existing gaps between public opinion and policy implementation. Additionally, the gap between energy development, consumption, and public views on energy could be more efficiently communicated. The opinions on social media could then be immediately measured to find the successfulness of these campaigns. In an effort to expand solar and wind installations, marketing strategies could be used on test audiences and the effectiveness of their campaigns could again be measured before and after.

According to Sovacool (2009), there is an apparent disconnect between how energy is produced and how it is socially perceived. Permeating the fabric of society in the U.S. is general apathy and misinformation about the benefits and drawbacks of various energy sources, thus leading to the view that renewable sources such as solar and wind are unneeded because people do not comprehend why such changes to the electrical grid are necessary. Furthermore, Soytas et al. (2007) found that income growth alone does not cause increased carbon emissions in the U.S., but increased energy use does. They found that continued income growth, as purported by

multiple agencies as the ultimate solution to climate issues, may not actually be a solution to the problem of run-away carbon emissions.

CHAPTER 5

CONCLUSION

This chapter is a summary of all key findings, limitations, and future research. Section 5.1 answers the research questions posed in section 1.2 of this paper. Next, section 5.2 describes the limitations of this study and includes recommendations for the next steps in research. Finally, section 5.3 concludes the paper by providing some final remarks.

5.1 Summary of Findings

In an attempt to better comprehend the perception of energy consumption on Twitter, it is essential to examine the original two research questions asked in section 1.2. Firstly, what is the public perception towards various energy sectors? This question can be answered by investigating four energy types: coal, solar, nuclear, and wind, using geocoding, frequency, sentient, and spatial analyses. Specifically, when viewing the timeseries chart, solar was almost always rated as more positive than coal, nuclear, and wind. Similarly, wind was almost always rated higher than coal and nuclear. This indicates that the public sentiment is higher towards renewable energy than either fossil fuels or nuclear energy. Coal was almost always rated higher than nuclear, indicating that there is a large disconnect between the public's perception on nuclear and the safety of using such an energy source. This low ranking for nuclear over time could also be attributed to how nuclear picked up some tweets about nuclear weapons.

The second question is concerned with if the public perception on energy issues exhibit spatiotemporal variation. It can be extracted from this study that coal, solar, nuclear, and wind perceptions from Twitter vary across both time and space. Different areas of the country have

starkly different characteristics, as specified from their different cultural backgrounds. From coal's location analysis, it was found that there were multiple counties around highly populated areas in California that had negative sentiment towards coal. The Appalachian Mountain range counties had neutral or even positive sentiments towards coal. Solar was more positive across almost all counties, especially on the West coast. Nuclear, on the other hand, was almost always thought of as negative or neutral across the conterminous U.S. Wind, although sparse in data, was also generally positive across the U.S. Someone tweeting from the Appalachia Mountain range has very different views than someone from California because their economy depends on the use of coal. Conversely, solar and wind seem to have similar spatial configurations as many people from California agree that energy production should come from these sources. A final item to note would be how positivity toward an energy subject can vary in time while different events occur. This is exemplified best by the closing of the Brown coal plant in Dallas, Texas, which sparked numerous tweets praising the move and subsequently increasing the positivity towards coal.

5.2 Limitations and Recommendations

The use of social media displayed several limitations and critical issues. Although the total volume of tweets on energy related subjects is high, there is a limited amount of defined data available for specific metropolitan areas. Challenges posed by this methodology of data aggregation can be summarized into three main sections: demographic representativeness, data quality, and algorithmic accuracy. For demographic representativeness, there are concerns about the statistical representativeness of the general population behind social media. Twitter users have been found to be much younger (Longley et al., 2015), thus it is difficult to glean attitudes

of the general public on Twitter. Additionally, geotagged tweets do not follow a consistent statistical sampling pattern, leading to errors in the representativeness of the dataset. Opinions on social media also tend to be inflammatory click-bait written to generate the largest number of likes and follows.

Data quality is another issue that has been found to be critical in the analysis. Even though there are petabytes of data generated on the subject of energy in the United States, a small amount of that data is actually actionable. Approximately 1% of all data has precise geotagged coordinates and 5% of data has general location information embedded in user profiles. These numbers are consistent with previous studies (Longley et al., 2015). It would be trivial for a profile location information to be spoofed. It was even found that broad location identifiers such as state names, country names, and even planet names were used. The location spread is also concentrated around major metropolitan areas and are sparse around smaller townships.

Twitter bots are also difficult to filter out since these autonomous entities provide hashtags that might coincide with the subject. Since it is estimated that about 10% of all users on Twitter are robots (Chen et al., 2017), the Twitter scrapper might be capturing the perception of autonomous entities rather than actual people. This fact was present in the user frequency charts created to examine power users in the study. The top power user for solar energy @solargenerator2 was almost certainly a robot that was generating text and retweeting anything that contained the hashtag “solar”. This can be discerned from the high frequency of the tweets, along with the fact that the account had been flagged by Twitter as an account with suspicious activity. It was elected to keep these robots in the dataset because it still represents activity on Twitter and might influence public opinion. However, for future studies examining human perception, it might be wise to remove these erroneous users by employing a threshold limit on

the tweets per day value. A spam filter for low-quality content as described by Chen et al. (2017) requires a large hand-labeled dataset and might be ill-suited for use because of the costs involved.

It is recommended that future studies examine the relationship between user supplied profile location information and the precise geolocated coordinates supplied when a user turns on the geolocation feature. This approach would find the percentage of tweets that have accurate profile location information by examining the relationship between user-supplied location information and the user's phone coordinates, which is a much more technologically challenging location track to spoof. A mobile virtual private network (VPN) application could be used to spoof the phone's accurate coordinates, but the percentage of people technologically capable of such a spoof is perceived to be relatively low. Perhaps additional studies could survey how many Twitter user's use a VPN to funnel their internet traffic through a server located elsewhere.

Current NLP algorithms are only so accurate at gleaning the sentiment of the writer through text analyses. Potential pitfalls include double negative concurrent words and sarcasm. The unsupervised lexicon-based NLP algorithm VADER (Hutto and Gilbert, 2014) was used to tag the positive, negative and neutral sentiments of tweets with accuracy rates ranging between 83-86%. Hopkins and King (2010) present a supervised, nonparametric statistical approach that might be used to improve upon accuracy in future studies.

The degrees of uncertainty in trusting social media as a source of information should be approached with caution, as there are many pitfalls of using such a source. Further studies examining the representativeness of demographics on social media are needed to address these data quality concerns. There is a clear gap in knowledge of social media's representativeness that is consistent with existing literature (Sang and Bos, 2012; Anstead and O'Loughlin, 2015).

Energy research on social media presents important opportunities and relevant challenges. Rich, discussion-filled content can be acquired with spatiotemporal dynamics from social media on energy development and consumption. Conversely, demographic representativeness, data quality, and algorithmic accuracy present a study with significant limitations. It is recommended that social media be used as a supportive data source coupled with traditional surveys. Using this data source can complement the perception of energy issues gleaned from traditional surveys by providing in-depth discussions of polarized opinions.

5.3 Conclusions and Final Remarks

There have been multiple positive aspects identified of using social media as a data source to study social perception of Twitter on energy sources based on this study and literature review. Social media has been found to be a dynamic, spatiotemporal data source that has a wide range of discussions on a host of renewable and nonrenewable energy types. Geolocated social media conversations can provide insights into how people view energy issues and thus be used as a precursor to traditional surveys to narrow down spatial clusters of positive, negative, and neutral areas of the country. Since spatiotemporal dynamics are difficult to record with traditional surveys, social media can provide important initial discoveries and inform survey designers. It is thus recommended that location-focused surveys be designed to improve upon statistical significance.

An additional positive aspect of using social media to study energy perception is that freeform conversations can take place that would be impossible to record with traditional surveys. Social media influencers can positively or negatively impact their following's perception on issues related to energy and can even mobilize support for various causes. Anyone

connected to these social media influencers has a chance of being influenced (Ceron et al., 2014). Social network analysis can further identify these influencers to see where their posts, likes, follows, and retweets impact opinion.

REFERENCES

"Americans' Opinion on Renewables and Other Energy Sources." Pew Research Center Science & Society. 2016. Accessed March 19, 2019. http://www.pewresearch.org/science/2016/10/04/public-opinion-on-renewables-and-other-energy-sources/#strong-public-support-for-more-wind-and-solar-closer-divides-over-nuclear-and-fossil-fuels.

“Charlottesville Wind Power Installation – Charlottesville Wind Turbine Installers.” DASolar. Accessed March 19, 2019. https://www.dasolar.com/home-wind-power/virginia/charlottesville.

“Get Latitude and Longitude.” Latitude and Longitude Finer. Accessed March 19, 2019. https://www.latlong.net/.

“The Four V’s of Big Data.” IBM Big Data & Analytics Hub. Accessed April 05, 2019. https://www.ibmbigdatahub.com/infographic/four-vs-big-data

“The World Factbook: United States.” Central Intelligence Agency. February 01, 2018. Accessed March 19, 2019. https://www.cia.gov/library/publications/the-world-factbook/geos/us.html

“Topic Modeling in Python with Gensim.” Machine Learning Plus. December 04, 2018. Accessed March 27, 2019. https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/.

“Twitter for Business | Twitter Tips, Tools and Best Practices.” Twitter. Accessed April 10, 2019. https://business.twitter.com/.

“United States.” Data. Accessed March 19, 2019. https://data.worldbank.org/country/united-states.

Adedoyin-Olowe, M., M. M. Gaber, F. Stahl. 2014. A Survey of Data Mining Techniques for Social Media Analysis. *Journal of Data Mining & Digital Humanities* 18.

Anderson, C., P. Breimyer, S. Foster, K. Geyer, J.D. Griffith, A. Heier, A. Majumdar, D.C. Shah, O. Simek, N. Stanisha, F.R. Waugh. 2015. A Network Science Approach to Open Source Data Fusion and Analytics for Disaster Response. *18th International Conference on Information Fusion*.

Anstead, N., O’Loughlin, B., 2015. Social media analysis and public opinion: The 2010 UK general election. *Journal of Computer-Mediated Communication* 20, 204–220.

Baker, K.S., S.K. Sahu. 2015. spTimer: Spatio-Temporal Bayesian Modeling Using R. *Journal of Statistical Software* 63(15).

Bolsen, T., Cook, F.L., 2008. The polls--trends: Public opinion on energy policy: 1974-2006. *Public Opinion Quarterly* 72: 364–388.

Batel, S., Devine-Wright, P., Tangeland, T., 2013. Social acceptance of low carbon energy and associated infrastructures: A critical discussion. *Energy Policy* 58: 1–5.

Bertot, J. C., P. T. Jaeger, and D. Hansen, 2012. The impact of polices on government social media usage: Issues, challenges, and recommendations. *Government Information Quarterly* 29 (1): 30–40.

Ceron, A., Curini, L., Iacus, S.M., Porro, G., 2014. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society* 16: 340–358.

Chen, H. M., P. Franks. 2016. Exploring Government Uses of Social Media through Twitter Sentiment Analysis. *Journal of Digital Information Management* 14(5): 290-301.

Chen, W., C.K. Yeo, C.T. Lau, B.S. Lee. 2017. A study on real-time low-quality content detection on Twitter from the users' perspective. *PLoS ONE* 12(8). https://doi.org/10.1371/journal.pone.0182487

Cody, E.M., A.J. Reagan, L. Mitchell, P.S. Dodds, C.M. Danforth. 2015. Climate change sentiment on Twitter: an unsolicited public opinion poll. *PLoS One* 10(8).

Cohen, J.J., Reichl, J., Schmidthaler, M., 2014. Re-focusing research efforts on the public acceptance of energy infrastructure: A critical review. *Energy* 76: 4–9.

Crampton, J.W., M. Graham, A. Poorthuis, T. Shelton, M. Stephens, M.W. Wilson, M. Zook. 2013. Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. *Cartography and Geographic Information Science* 40(2): 130-139.

Crandall, D.J., L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. 2010. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Science of the United States of America* 107(52).

Dai, A.M., C. Olah, Q.V. Le. 2015. Document Embedding with Paragraph Vectors. *ARXIV*. Accessed January 17, 2018. https://arxiv.org/abs/1507.07998

De Choudhury, M., M. Gamon, S. Counts, E. Horvitz. 2013. Predicting depression via social media. *AAAI Conference on Weblogs and Social Media*.

Farhar, B.C., Unseld, C.T., Vories, R., Crews, R., 1980. Public opinion about energy. *Annual Review of Energy* 5: 141–172.

Farhar, B., 1994. Trends: Public opinion about energy. *The Public Opinion Quarterly* 58(4): 603-632. Retrieved from http://www.jstor.org.libproxy.unl.edu/stable/2749611.

Gao, S., H. Chen, W. Luo, Y. Hu, X. Ye. 2018. Spatio-Temporal-Network Visualization for Exploring Human Movements and Interactions in Physical and Virtual Spaces. *Human Dynamics Research in Smart and Connected Communities* 67-80.

Go, A. R. Bhayani, L. Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford* (1)12.

Goodchild, M., D. Sui. 2011. The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science* 25(11): 1737-1748.

Hamstead, Z.A., D. Fisher, R.T. Ilieva, S.A. Wood, T. McPhearson, P. Kremer. 2018. Geolocated social media as a rapid indicator of park visitation and equitable park access. *Computers, Environment and Urban Systems*.

Hasan, S., X. Zhan, and S.V. Ukkusuri. 2013. Understanding Urban Human Activity and Mobility Patterns Using Large-scale Location-based Data from Online Social Media. *UrbComp '13 Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing* 6.

He, W., H. Wu, G. Yan, V. Akula, J. Shen. 2015. A novel social media competitive analytics framework with sentiment benchmarks. *Information & Management* 52(7): 801-812.

Hopkins, D.J., King, G., 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science* 54: 229–247.

Hossain, N., T. Hu, R. Feizi, A.M. White, J. Luo, H. Kautz. 2016. Inferring Fine-grained Details on User Activities and Home Location from Social Media: Detecting Drinking-While-tweeting Patterns in Communities. *ICWSM 2016*.

Hu, B., M. Jamali, and M. Ester. 2013. Spatio-temporal topic modeling in mobile social media for location recommendation. *2013 IEEE 13th International Conference on Data Mining*.

Hu, H., J.J.H. Zhu. 2015. Social networks, mass media and public opinions. *Journal of Economic Interaction and Coordination* 1-19.

Huang, Q., G. Cao, C. Wang. 2014. From Where Do tweets Originate? - A GIS Approach for User Location Inference. *LBSN'14 Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks* 1-8.

Hutto, C.J., E. Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media.*

International Panel on Climate Change. 2014. Climate Change 2014 Synthesis Report. *Summary for Policymakers* 10.

Kaisler, S., F. Armour, J.A. Espinosa, W. Money. 2013. Big Data: Issues and Challenges Moving Forward. *2013 46th Hawaii International Conference on System Sciences*.

Kim, Y., S.R. Jeong. 2015. Opinion-Mining Methodology for Social Media Analytics. *KSII Transactions on Internet and Information Systems* 9(1).

Kulkarni, Y. 2017. Sentiment Analysis of Twitter Posts on Channai Floods using Python. https://www.analyticsvidhya.com/blog/2017/01/sentiment-analysis-of-twitter-posts-on-chennai-floods-using-python/. (accessed November 29, 2017).

Kusner, M.J., Y. Sun, N.I. Kolkin, K.Q. Weinberger. 2015. From word embeddings to document distances. *ICML'15 Proceedings of the 32nd International Conference on Machine Learning* 37: 957-966.

Lau, J.H., T. Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. *Proceedings of the 1st Workshop on Representation Learning for NLP* 78-86.

Lee, T.M., E.M. Markowitz, P.D. Howe, C.Y. Ko, and A.A. Leiserowitz. 2015. Predictors of public climate change awareness and risk perception around the world. *Nature Climate Change* 5: 1014-1020.

Lesic, V., W.B. Bruin, M.C. Davis, T. Krishnamurti, and I.M.L. 2018. Azevedo. Consumers' perceptions of energy use and energy savings: A literature review. *Environmental Research Letters* 13:3.

Line, Bottom. "The 10 States That Run on Nuclear Power." NBCNews.com. February 23, 2012. Accessed March 19, 2019. https://www.nbcnews.com/businessmain/10-states-run-nuclear-power-169050.

Lith, A., J. Mattsson. 2010. Investigating storage solutions for large data: A comparison of well performing and scalable data storage solutions for real time extraction and batch insertion of data. (Masters' Thesis, Chalmers University of Technology), 12-15.

Liu, X., Q. Huang, S. Gao. 2019. Exploring the uncertainty of activity zone detection using digital footprints with multi-scaled DBSCAN. *International Journal of Geographic Information Science*.

Longley, P. A., Adnan, M., Lansley, G., 2015. The geotemporal demographics of Twitter usage. *Environment and Planning* 47(2) 465–484.

Mahmud, J., J. Nichols, C. Drews. 2012. Where Is This tweet From? Inferring Home Locations of Twitter Users. *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*.

Mikael. "Scraping, Extracting and Mapping Geodata from Twitter." Mikael Brunila. April 22, 2017. Accessed March 19, 2019. http://www.mikaelbrunila.fi/2017/03/27/scraping-extracting-mapping-geodata-twitter/

Mikolov, T., Q. Le. 2014. Distributed Representations of Sentences and Documents. *Proceedings of the 31st International Conference on Machine Learning (ICML)* 1188-1196.

Newman, D., A. Asuncion, P. Smyth, M. Weling. 2009. Distributed Algorithms for Topic Models. *Journal of Machine Learning Research* 10: 1801-1828.

Petkov, P., F. Köbler, M. Foth, H. Krcmar. 2011. Motivating domestic energy conservation through comparative, community-based feedback in mobile and social media. *Proceedings of the 5th International Conference on Communities and Technologies* 21-30.

Russell, M.G., J. Flora, M. Strohmaier, J. Poeschke, J. Yu, N. Rubens, M.A. Smith. 2013. Semantic analysis of energy-related conversations in social media. *L. Kahle and E.G. Atay (eds.), Sustainability and Lifestyle Marketing, M.E. Sharpe: Armonk, NY*.

Sang, E.T.K., Bos, J., 2012. Predicting the 2011 Dutch senate election results with Twitter, in: Proceedings of the Workshop on Semantic Analysis in Social Media, EACL '12. *Association for Computational Linguistics, Stroudsburg, PA, USA* pp. 53–60.

Santoianni D.A. 2013. The Social media battleground: How public perception, science communication, media coverage, and politics are shaped by social media. *2013 World of Coal Ash (WOCA) Conference*. Online Retrieved at http://www.flyash.info/2013/093-Santoianni-2013.pdf

Sivarajah, U., Fragidis, G., Lombardi, M., Lee, H., & Irani, Z. 2015. The use of social media for improving energy consumption awareness and efficiency: An overview of existing tools. *European, Mediterranean & Middle Eastern Conference on Information Systems 2015* (EMCIS2015), Athens, Greece.

Sovacool, B.K. 2009. The cultural barriers to renewable energy and energy efficiency in the United States. *Technology in Society* 31(4): 365-373.

Soytas, U., R. Sari, B.T. Ewing. 2007. Energy consumption, income, and carbon emissions in the United States. *Ecological Economics* 62(3-4): 482-489.

Stefanidis, A., A. Crooks, J. Radzikowski. 2011. Harvesting ambient geospatial information from social media feeds. *GeoJournal* 78(2): 319-338.

Sütterlin, B., Siegrist, M., 2017. Public acceptance of renewable energy technologies from an abstract versus concrete perspective and the positive imagery of solar power. *Energy Policy* 106: 356–366.

Swift, J. and C. Weidemann. 2013. Social Media Location Intelligence: The Next Privacy Battle – An ArcGIS add in and Analysis of Geospatial Data Collected from Twitter.com. *International Journal of Geoinformatics* 9(2).

Oliveira, D.J.S., P.H.S. Bermejo, P.A. Santos. 2017. Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls. *Journal of Information Technology & Politics* 14(1): 34-45.

Tufekci, Z. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *ICWSM '14: Proceedings of the 8the International AAAI Conference on Weblogs and Social Media*.

Twitter. "Docs". Developer Twitter. https://developer.twitter.com/en/docs. (accessed November 29, 2017).

Vraga, E.K., A.A. Anderson, J.E. Kotcher, E.W. Maibach. 2015. Issue-Specific Engagement: How Facebook Contributes to Opinion Leadership and Efficacy on Energy and Climate Issues. *Journal of Information Technology & Politics* 12: 200-218.

Wagner, K. How many people use Twitter every day? *Recode*. July 27, 2017. Accessed November 07, 2017. https://www.recode.net/2017/7/27/16049334/twitter-daily-active-users-dau-growth-q2-earnings-2017.

Wickham, H. 2014. Tidy Data. *Journal of Statistical Software* 59(10).

Williams, H.T.P., J.R. McMurray, T. Kurz, F.H. Lambert. 2015. Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change* 32: 126-138.

Wüstenhagen, R., Wolsink, M., Bürer, M.J., 2007. Social acceptance of renewable energy innovation: An introduction to the concept. *Energy Policy* 35: 2683–2691.

Yang, W. and M. Lan. 2015. GIS analysis of depression among Twitter users. *Applied Geography* 60: 217-223.

Yaqub, U., N. Sharma, R. Pabreja, S.A. Chun, V. Atluri, J. Vaidya. 2018. Analysis and Visualization of Subjectivity and Polarity of Twitter Location Data. *Proceedings of the 19th Annual International Conference on Digital Government Research*.

Yoon, S., N. Elhadad, S. Bakken. 2013. A practical approach for content mining of tweets. *American journal of preventive medicine* 45(1): 122-129.

Zhang, Y., B. Wallace. 2015. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification.

VITA

Graduate School
Southern Illinois University

David Leifer

davleifer@gmail.com

University of Wisconsin- Eau Claire
Bachelor of Arts, Environmental Geography, May 2016

Special Honors and Awards:

- Gamma Theta Upsilon International Honorary Society in Geography
- Graduate Professional Student Council Career Development Reimbursement for Travel, 2018
- American Society of Mining Reclamation Travel Grant, 2018
- Open Source Undergraduate Geospatial Technical Skills National Science Foundation National Geospatial Technology Center Award, 2016
- AAG Cyberinfrastructure Specialty Group Robert Raskin Student Competition Finalist, 2016
- UW-Eau Claire Department of Geography Student Travel for Presentation, 2016
- UW-Eau Claire Office of Research Sponsored Programs Student Travel for Research Presentation, 2016

Thesis Paper Title

Social Media Footprints of Public Perception on U.S. Energy Issues

Major Professor: Ruopu Li