Proceedings of The National Conference
On Undergraduate Research (NCUR) 2020
Montana State University, Bozeman, MT
March 26-28, 2020

Sentiment Analysis and Visualization for Citizen Opinion in Social Media

Nicholas Alvino
School of Computer Science and Technology
Kean University
1000 Morris Avenue
Union, New Jersey 07083 USA

Faculty Advisor: Dr. Daehan Kwak

Abstract

Over the past few years, social media has become the political space for campaigning and governing, changing the way information is sent from one population to another. Governments have been trying to focus more on a citizen-centric model of society, making priorities and services driven more by citizen needs rather than the governments' capability. It has become a daily routine for political leaders to use the social media platform, Twitter, to send out messages, called tweets, to influence campaigns and interact with the public. Twitter provides a plethora of potentially useful information that could be collected and analyzed to weigh the public opinion. It is essential to monitor the public opinion on social media for negative responses, petitions, and any other concerns citizens have. However, there are limitations on how to quantify, visualize, and look at the big picture regarding public opinion. In this research, a tweet mining system is developed to monitor, analyze, and visualize the negative citizen sentiment via Twitter. This mining system will obtain the tweets based on geographic location and negative sentiment, such as #angry or #mad, and search for recurring words that are used to determine the topic of negative public option, such as #GasPrice or #PublicTransportation. Using a sentiment analyzer that quantifies the negative sentiment on a numeric (1-10) and color scale, the tweets are analyzed and visually displayed in a word cloud which is embedded on a Google Map to pictorially understand the topic of negative public sentiment for a specific area. A case study is conducted on tweets collected for 2019 as proof of concept. The proposed tweet mining sentiment analysis system has significance to easily visualize common issues citizens have, improve government services, reshape political agenda, and personalized campaigning.

Keywords: Sentiment Analysis, Data Visualization, Public Opinion, Twitter

1. Introduction

Every day, Twitter has on average around 500 million tweets posted along with its nearly 126 million daily active users on the site [1]. Upon Twitter's creation in March of 2006, it quickly grew to one of the largest social media platforms we have today, alongside Facebook, Instagram, and YouTube. Tweets are versatile in the sense that it could be anything the user wants to post about; they could retweet someone else's post, share videos, photos, poems, or anything they want to share. To some, Twitter is considered the main source of information leading many high-profile people, such as politicians, celebrities, and analysts to move their operations over to the site.

This research aims to conduct an analysis of United States citizens' sentiment by mining tweets from Twitter and applying a sentiment value to locations in order to provide public opinions of the areas. Sentiment can be swayed by many factors: political agendas, emergency response, local or federal governmental changes, natural disasters, etc. A sentiment map is created to provide a visual of common topics of interest in specific locations. From this map, a conclusion could be drawn to show common issues that arise in these areas. The overall objective of this project is to determine if tweets can be used to create an accurate sentiment analysis of public opinion to represent an area.

2. Background

Social Media's uprising has made a dramatic impact on the publics' opinion on important topics; some may be positive influences; however, many opinions may be swayed by misinformation. This impact has affected every part of the world, from political views in the United States to consumer purchases in China. There have been many studies conducted showing the effect social media has on sensitive topics, such as one discussing the impact WeChat has on college students' political agendas [2]. This study focuses on understanding how political discussions are swayed due to the use of social media. Additionally, there have been studies focusing on the public reactions to certain events in our history, such as same-sex-marriage cases [3]. This specific study focuses on the behavioral aspect of current events, using Twitter to monitor witnesses' reactions to crucial court hearings. While these studies primarily research specific topics or focus on specific methodology, our study will build on these views showing many genres as well as incorporate computer science techniques such as web scraping to make collecting data more efficient and less labor-intensive, minimizing the amount of work we would manually have to do. In order to do so, there are aspects of Tweets that will be useful in determining the users statistics, such as the location, date/time, and hashtags.

Twitter, after becoming the data powerhouse it is today, provides many tools that researchers and data scientists use. In order to gain access to these tools, a user must apply for a Twitter developer account to access the Twitter API. Once your application has been accepted, you will gain access to your unique access token, secret access token, consumer key, and secret consumer key. As some of the titles suggest, these numbers are a string of characters that should be kept "secret". These passcodes give you access to the library of Twitter APIs and it is important not to lose these keys as they will be used when connecting to the Twitter API data. In addition, if your keys end up in the wrong hands and are used maliciously, Twitter reserves the right to ban your account and take action. There are many libraries for coding languages such as Python that use a user's tokens and keys to pull data from Twitter; this pulling of data is known as scraping. The scraping of tweets and other data provided by Twitter API is crucial for the success of research projects such as ours, allowing for more purposeful data-driven research to thrive.

For our study, we focus on one important aspect of tweets, i.e. geotags. Geotags, or location tags, are provided in most tweets that allow the user to share the location in which the tweet took place (*Figure 1*). With this place of origin, we are able to better decipher where the user is tweeting from and we can connect this data to the information within the tweet. Along with the location, we will be taking a look at the other metadata of the tweet, such as the date, time, and hashtags of the tweet.

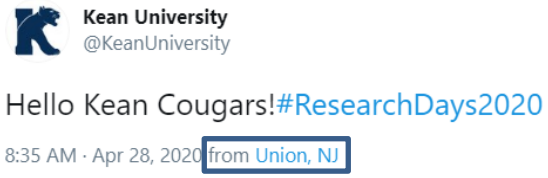

Figure 1. Example of a Twitter user with Geotags enabled. This tweet from the user Kean University indicates the message was posted from Union, NJ as shown in the geotag. Additional information such as the date of April 2020 as well as the hashtag #ResearchDays2020 can be used to further classify the tweet.

3. Experiment Setup

Before beginning the experiment, we decided to study two major cities, Union, New Jersey and New York City, New York. One-thousand tweets were collected from each location starting from February 2020 to March 2020, resulting in two months of data. For data collection, we used Python coding language to use and interact with two API's: Twitter API and the IBM Watson tone analyzer [4, 5]. When a database was needed to store results for comparison, a MySQL database was utilized. In regards to the IBM Watson tone analyzer, there were a total of seven tones in which we used only two for this study, Anger and Joy. These were the two we felt would be most helpful for our purpose of assigning sentiment to public opinion. Once the sentiment was determined, the result would impact the color of the text on our final map.

4. Procedures

4.1. Data Mining

A Twitter mining system (Figure 2) was set up to collect the proper tweet geotag using a Python library called Tweepy [6, 7]. With Tweepy, we were able to direct the Twitter API to crawl only the tweets with our specific location request and return the results. Using Python, the results returned were condensed and placed in a comma-separated values (.csv) file. This CSV file is saved to a server database running on the Linux operating system. Once prompted, the server would communicate with the IBM Watson API and return the rating from 1-5, if the tweet was happy, it would be closer to 5 and if the tweet was aggressive or negative, it would be closer to 1. The artificial intelligence within the Watson API will make this determination based on linguistic analysis that focuses on emotional tones and language used in the text. These results were sent back to the server and stored within a MySQL database for further calculations, if necessary, and easy access for visualization later.

The overall process is shown in Figure 2 which visualizes all the processes that communicate with the server database. First, the server queries Twitter API using Tweepy and returns results relevant to our specific geotag entries, New York City, New York and Union, New Jersey. The data collected is stored in the CSV file format and sent back to the database server. This server then queries IBM Watson's API to return a sentiment score and stores it locally.

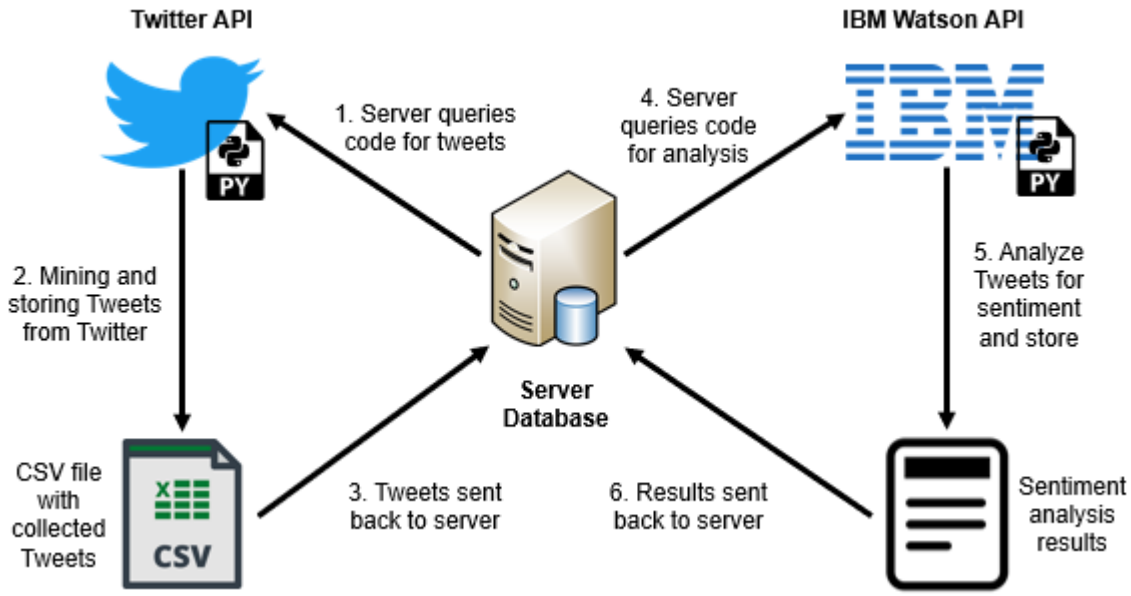

Figure 2. Diagram showing the process in which data is scraped using Twitter's API and sent to IBM Watson tone analyzer for sentiment analysis, resulting in usable data being sent back to the server.

4.2. Sentiment Analysis Visualization

To succeed in the readability of this study, the visualization of the data was essential. In order to do this, our Python code would communicate with our server and send the results from prior analysis to Google Maps API [8]. With this, we were able to display a word cloud using d3.js [9] in combination with photo editing software, Adobe Photoshop and Gimp, to shape the "cloud" to our liking. This process had minimal complexity but the ability to customize our findings was a never-ending discovery. If a result was deemed negative, the text would be in red while if the tweet was a positive one, it would be shown in green or blue as shown in Figure 3. If data was unable to be recognized by IBM Watson API, the data was sent back independently of other data and reprocessed by the Watson API so it would generate meaningful results.

As demonstrated in Figure 4, after the results are sent back to the server database following our datamining procedures, they are sent to Google Maps API which allows for the clear designation of the areas we are investigating. With these regions shaded properly, the word cloud is placed on top of the location, resulting in a meaningful yet concise image visualizing the data from our analysis.

Figure 3. Legend for Sentiment Analysis Colors. Happiness (5) would be displayed as a light blue color while anger or negativity (1) would be shown as a darker maroon or red.

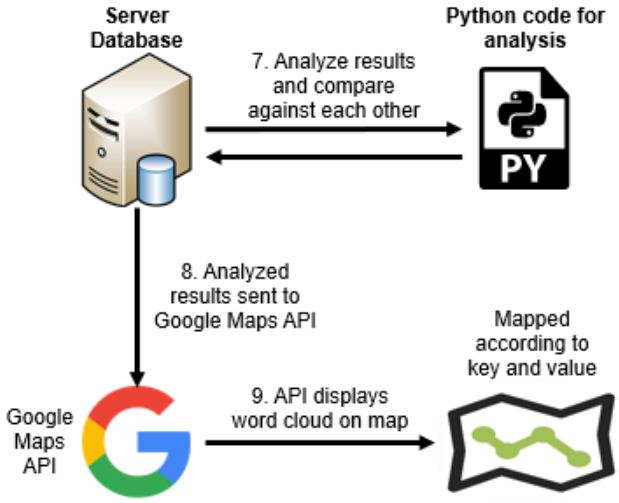

Figure 4. Diagram shows the process in which the data collected from sentiment analysis is displayed over a map to stress importance. The usage of API's was crucial in the autonomous nature of this setup.

5. Results

The resulting images are shown in Figures 5 and 6. We were visibly able to see the amount of negatives as the red text outweighs the amount of green text in both pins. Due to the Coronavirus, COVID-19, outbreak that occurred during this study, most of the tweets involved a topic relating to this. As seen in both visuals, this term along with topics about cleanliness or cleaning appears most because in the word cloud produced, the more a word occurs the larger the size of the text. Figures 5 and 6 are the cloud representations of topics colored by the sentiment from Union, New Jersey and NYC, New York, respectively. The main common public opinions for both cities were the coronavirus, cleanliness, and gas prices being high. Some differences between the two cities were that New Jersey had public opinions about "NJ transit" and "quit littering", whereas New York had "I like Cuomo" (NY's governor) as their most popular topics. Our one-thousand tweet threshold was met much faster with tweets originating from New York as the city generated nearly twice as many tweets as New Jersey per minute.

Figure 5. Visualization of sentiment analysis for Union, New Jersey, USA

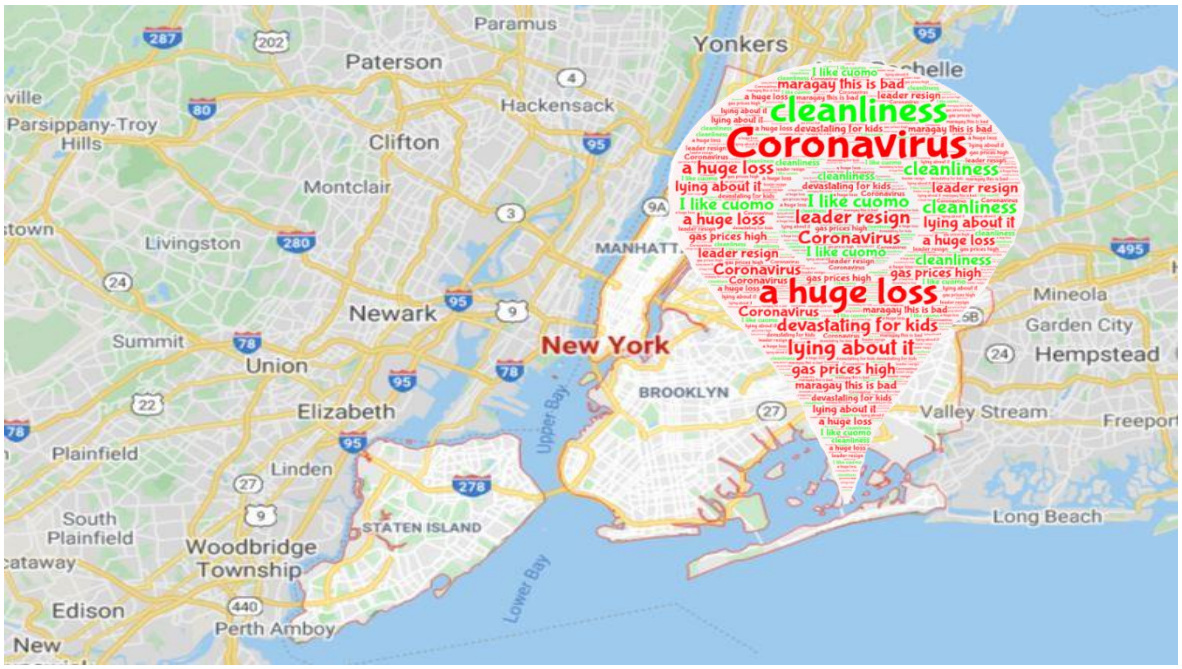

Figure 6. Visualization of sentiment analysis for New York City, New York, USA

6. Conclusion

After the conclusion of this study, it is apparent that many people provide useful information to Twitter that could greatly benefit other users if properly visualized or displayed to be understood more clearly. By searching for tweets using the two geotags of Union, New Jersey and New York City, New York, we were able to determine a generalized

sentiment value on public opinions which could be used by others to gauge the importance of issues in these areas. For example, it was very clear that cleanliness was important to most in both cities and the coronavirus had a negative impact on both communities. With this type of analysis in combination with other software and applications such as Google maps, we have the capability to visualize the common issues citizens have and to look at the big picture regarding public opinion and thus improve government services, reshape political agenda, and personalized campaigning. It is also important to note that there are mostly negatives shown on both maps and that most of the tweets received were not positive. One could infer this means the general morale of the location is negative and can use this information to perhaps try to change this tenor.

7. Future Work

With the system we have in place, there is still a high level of human interaction that must take place for the results to be finalized and displayed. To improve this, we plan to automate the system so the user can input a location and then be provided with topics of interest that are visually understandable based on the system we currently have in place. Another way to improve would be to further research sentiment analysis methods for these areas [10, 11] and research other ways to visualize information onto a map [12, 13]. Twitter is an amazing resource; however, there have been many other means by which people express their opinions. Additionally, our study and Twitter itself is unable to 100% control the amount of misinformation shown on their site. Due to the interconnectivity of social media, information can travel very fast from one nation or state to another and it has become increasingly difficult for a user to determine whether a source is reliable or not. With this said, it would be interesting to try and develop or train a form of artificial intelligence to search other websites on the internet and determine whether a news story or statement in a tweet can be reliable or not.

8. Acknowledgements

This research is based upon work supported by Kean University's Students Partnering with Faculty (SpF) Summer Research Program and Louis Stokes Alliances for Minority Participation Program (LSAMP).

9. References

1. Hamza Shaban, "Twitter reveals its daily active user numbers for the first time." The Washington Post, February 7, 2019. https://www.washingtonpost.com/technology/2019/02/07/twitter-reveals-its-daily-active-user-numbers-first-time/.

2. Pang, H. (2018). Is mobile app a new political discussion platform? An empirical study of the effect of WeChat use on college students' political discussion and political efficacy. Plos One, 13(8). https://doi.org/10.1371/journal.pone.0202244.

3. Clark, T. S., Staton, J. K., Wang, Y., & Agichtein, E. (2018). Using Twitter to Study Public Discourse in the Wake of Judicial Decisions: Public Reactions to the Supreme Court's Same-Sex-Marriage Cases. Journal of Law and Courts, 6(1), 93–126. https://doi.org/10.1086/695423.

4. Opesanya, Bayo. "Getting Started with IBM Watson Tone Analyzer." Codeburst.io, March 27, 2018. https://codeburst.io/getting-started-with-ibm-watson-tone-analyzer-3aa3386cff15.

5. IBM Watson's tone analyzer: www.ibm.com/watson/tone-analyzer

6. "Extended Tweets." Extended Tweets - tweepy 3.8.0 documentation, n.d. http://docs.tweepy.org/en/latest/extended_tweets.html#streaming.

7. Garcia, Miguel. "How to Make a Twitter Bot in Python With Tweepy." Real Python, September 22, 2019. https://realpython.com/twitter-bot-python-tweepy/.

8. Google Maps APIs: https://developers.google.com/maps/documentation

9. D3.js - Data-Driven Documents: https://d3js.org/

10. Ali, F., Kwak, D., Khan, P., Islam, S.R., Kim, K.H. and Kwak, K.S., 2017. Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling. Transportation Research Part C: Emerging Technologies, 77, pp.33-48.

11. Ali, F., Kwak, D., Khan, P., El-Sappagh, S., Ali, A., Ullah, S., Kim, K.H. and Kwak, K.S., 2019. Transportation sentiment analysis using word embedding and ontology-based topic modeling. Knowledge-Based Systems, 174, pp.27-42.

12. Kwak, D., Liu, R., Kim, D., Nath, B. and Iftode, L., 2016. Seeing is believing: Sharing real-time visual traffic information via vehicular clouds. IEEE Access, 4, pp.3617-3631.

13. Kwak, D., Kim, D., Liu, R., Iftode, L. and Nath, B., 2014, November. Tweeting traffic image reports on the road. In 6th International Conference on Mobile Computing, Applications and Services (pp. 40-48). IEEE.