Twitter Sentiment Analysis

Dean Stirrat (SUID - 72806-8008)
Anthony Verdone (SUID - 67162-1730)
Hossain Delwar (SUID - 59161-4692)
Jason Kirk (SUID - 47960-1331)
Zhiyuan Zhang(SUID - 74619-6880)
Erxi Liu(SUID - 65164-6358)

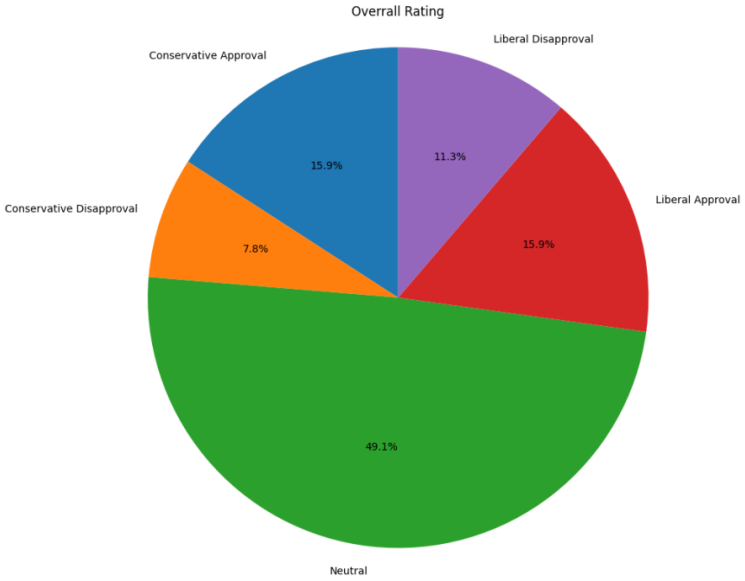
Social Media and Data Mining (CIS - 400)
Syracuse University
New York
May 2022

Table Of Contents

1. Introduction

As we young adults divulge ourselves into the real world without filters, we tend to see many remarkable actions throughout this small globe. Some of these actions may be regarded as negative while others may be categorized as positive but most of this power to determine something as right or wrong lays in the hands of elected officials who represent us as a group to a larger body to allow us as a society to have rules or laws govern over us. Although simply stating this may be a simple outlook of the government, there is always a division in parliaments and congress as different representatives have different outlooks on serious situations and how the laws made may affect us later on. Most of this division can be split between two different parties based on their ideologies, one that is left-wing while the other is right. In the current political climate of the US, most right-wing citizens classify themselves as Republicans while left-wing citizens identify as Democrats. The current US congress is mostly half and a half between the groups but the Democrats currently hold a majority. However, although 50% of the US congress is democratic, what percent of the US body of population identifies as such?

This can be analyzed based on a social media analysis. Since most adults now have some form of social media such as Twitter or Facebook, a lot of political information and opinions circulate on the web and it is a great form of analyzing what the US thinks. The division between different parties may be borderline equal but those who sway neutral and have to choose sides determine a big factor on the US political ideology.



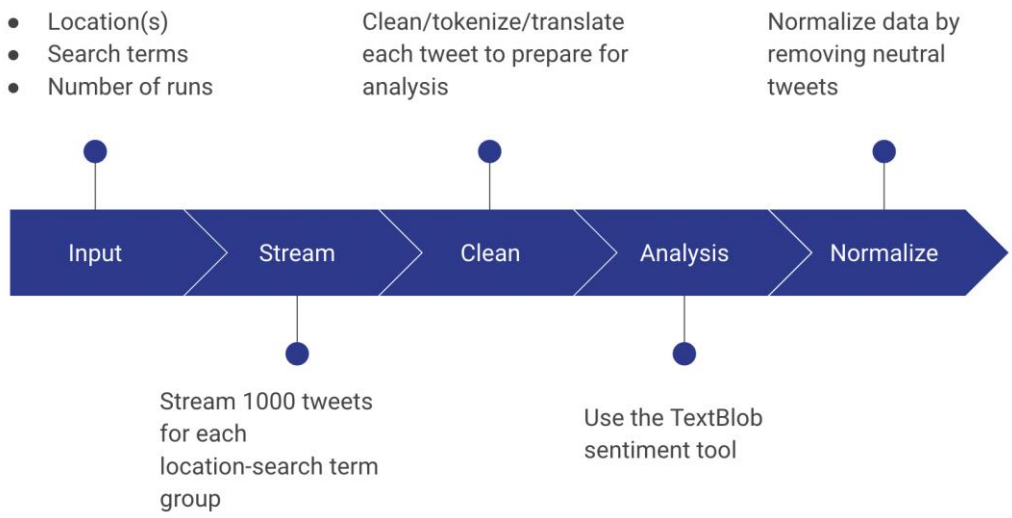
As we can see in the above chart created using twitter API and sentiment analysis on 20 various cities, those who remain neutral determine how the US is shaped and their choices is what matters the most. This example can be seen in our presidential election where swing states determine who the next president will be.

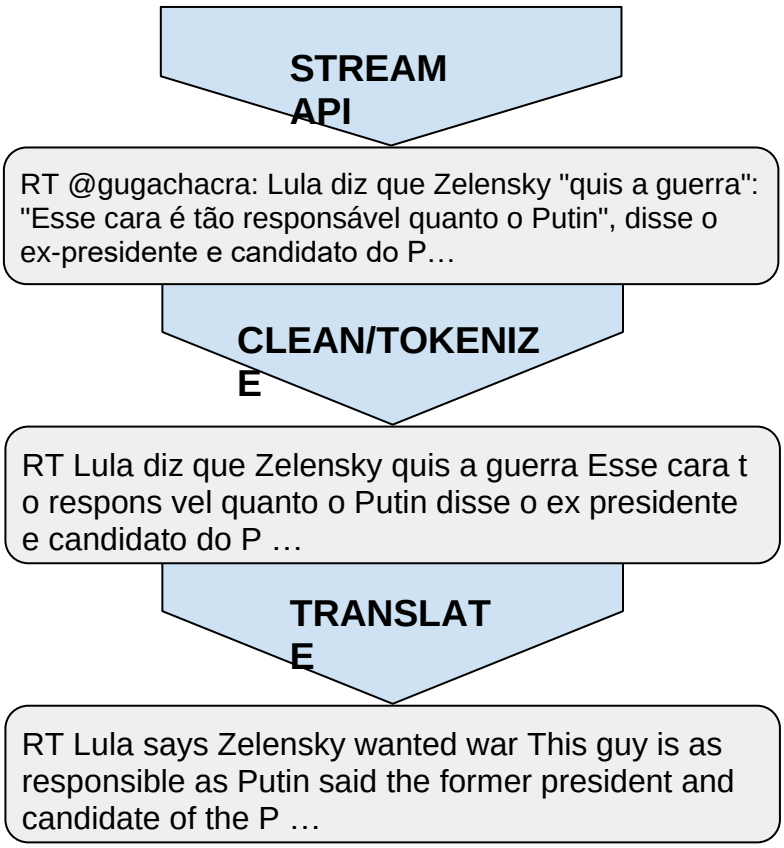
2. Prior Work

Twitter remains a dominant social media interface since 2007. With 21% of twitter users being American, it can determine the many different background Americans have. Since the use of streaming will be used, much of tweets users tweet will be recent and completely real time. Based on our previous attempt on understanding political ideologies in the US on social media, such as the previous graph, we came to an understanding that this is not enough. The main reasoning behind our scope of work is understanding the social beliefs people have using streaming. Our previous research was focused on labeling someone as left or right wing but that would mean analyzing previous tweets from a person which can take hours. Therefore, since there will always be those who are neutral, we suggested on seeing approval ratings of political leaders. There is always some form of news about each elected leader circulating throughout the nation and they each have either a positive or negative influence on users. We streamed at various locations for certain relevant leaders such as Joe Biden, Kamala Harris, Donald Trump, John Cox, Gavin Newsom, etc. Each stream were done based on certain timings as news outbreak determine if twitter users will tweet about this person.

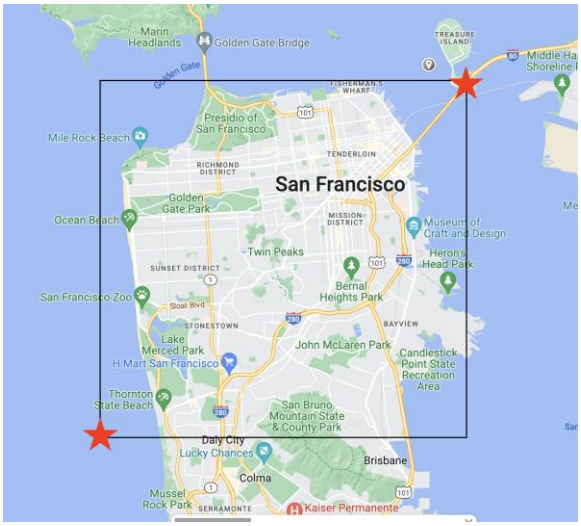
3. Process Flow

In order for our program to achieve precise sentiment analysis results, we must first process the data we retrieve from twitter. Once we get the result from the sentiment analysis tool, we must use the result to produce a useful measurement. The following is the structure for the flow of events which our program uses to get from input to results.





Firstly, our program requires a number of parameters to determine which tweets to fetch and from where. By using twitter's filtered stream API, we can provide search



terms which will be used to find tweets that they appear in. Terms are separated by commas, creating an "or" relationship that tells the stream API to search for tweets that include any combination of 1 or more terms in the list. The search terms are represented as values in a python dictionary where the key is a string describing the group.

The other input is a dictionary containing location information. The stream API takes in a location which is represented by 2 (longitude, latitude) pairs. The 2 pairs are the opposite corners of a square which when overlaid on a world map represent the area polled. We again use a python dictionary to map location names to location boundary pairs.

Our program will take this input and loop over each location and then, for each location, loop over the search group. This process will give us all location-search group pairings possible. For each pairing we call the twitter API to generate a stream object which we will use to retrieve 1000 tweets (this number may be changed if needed). For each tweet produced by the stream our program will clean the tweet. The cleaning process has 3 major steps; clean, tokenize, translate.

In the clean step, we remove all numbers and symbols from the tweet. This is a necessary step in order to get optimal results from our sentiment analysis tool. This step includes removing websites, equations, and unique punctuation. While these elements may hold some value for sentiment analysis, the tools we use do not do well with comprehending these kinds of text.

Next, we tokenize the tweet. This step works in concert with the previous step to add a predictable spacing between each word, or token, in the tweet. The result of the first two steps are an even spaced string that is ready for input into the sentiment analysis.

The third step takes the output of the first 2 steps and translates the tweet using the google translate package to translate the tweet. This step is unnecessary for a majority of tweets but is essential for generating good results from foreign locations and US cities with large populations of non-english speaking citizens. The translate string is what is returned from the clean_tweet function and can now be used for our sentiment tool

Now that we have a cleaned tweet, we can call our analysis tool to get an estimate of the tweets sentiment. The returned number is either positive, negative or 0 which represents how positive the given tweet is.

We noticed early in our work that a vast majority of tweets are neutral, especially when looking at very large data sets. This led us to results where the percent difference between the positive and negative portion of tweets was quite small and made the data seem insignificant. To remedy this problem we added a final step to our process in order to remove neutral data and normalize the results.

In the normalization process we use a simple mathematical technique to remove the percentage of neutral tweets and then scale up the positive/negative percentages proportionally. This gives us easily distinguishable results that more effectively show sentiment data.

4. Data

Now that we have an organized flow and organization of data, it became time to put our application to the test. The program is completely designed to better understand the political ideologies of a specific location. The Twitter streaming API is very useful because locations can be as big or little as the program needs. In our case, our analysis can be as big as entire countries or as small as local towns. Further, we can perform analysis on large-scale elections in the United States or the United Kingdom or small local elections for the election of a state governor or small-town mayor.

For the general scope of the project, we broke down multiple different subjects so we can use our tool for. We initially created a simple subject to analyze: Biden 2024 and Trump 2024. For this analysis, we decided to start big and abate by region. This meant doing an analysis on the entire United States, West Coast of the United States, East Coast of the United States, then the Southern United States. This made sense conceptually since these regions differ drastically from

each other in terms of political values, and it offered a decent way to test our program. The results are shown below in Figure 4.1.

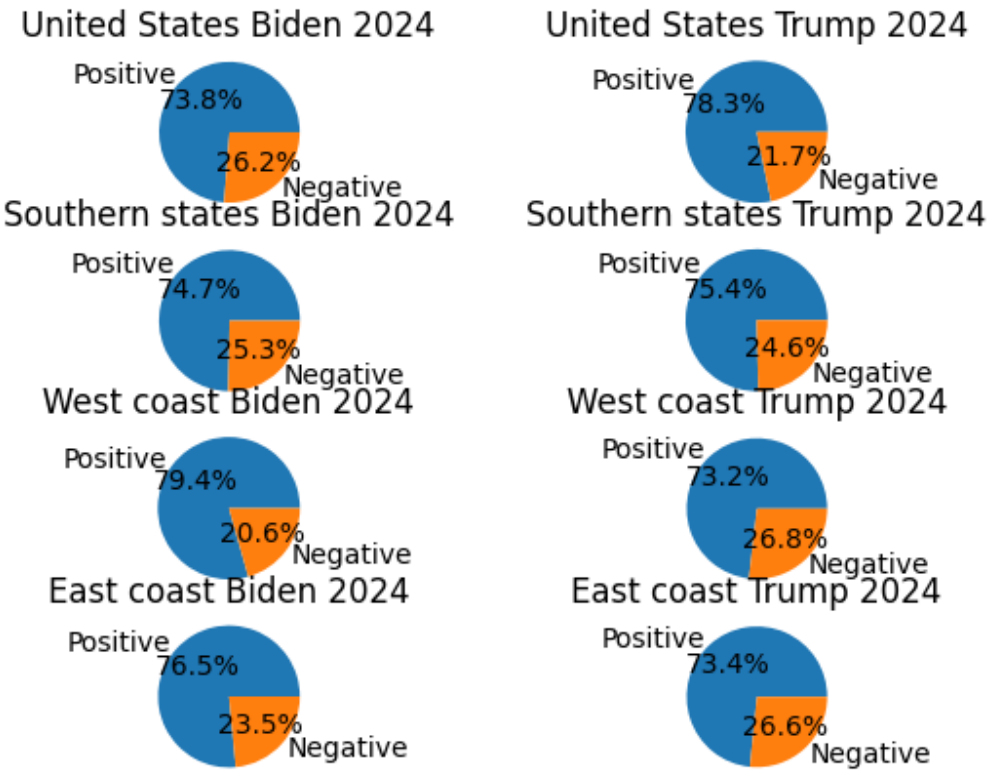


Figure 4.1: Analysis of Joe Biden and Donald Trump in the United States

The figure above represents the first analysis we performed with our application. The results appeared to be interesting because there doesn't seem to be an indisputable polarization, which is what most people would expect. Given the rise of online tension in political discourse, it's surprising to see positive feedback on Joe Biden and Donald Trump across the nation and within its divided regions. Nonetheless, this was our first analysis and there was much more room for deeper examination. The next step was to run the same program on a different selection of notable political figures across the same locations (United States, West Coast of the United States, East Coast of the United States, then the Southern United States). We decided to include Vice President Kamala Harris, Secretary Pete Buttigieg, Senator Bernie Sanders, and Governor Gavin Newsom. The results of the analysis can be seen below in figure 4.2

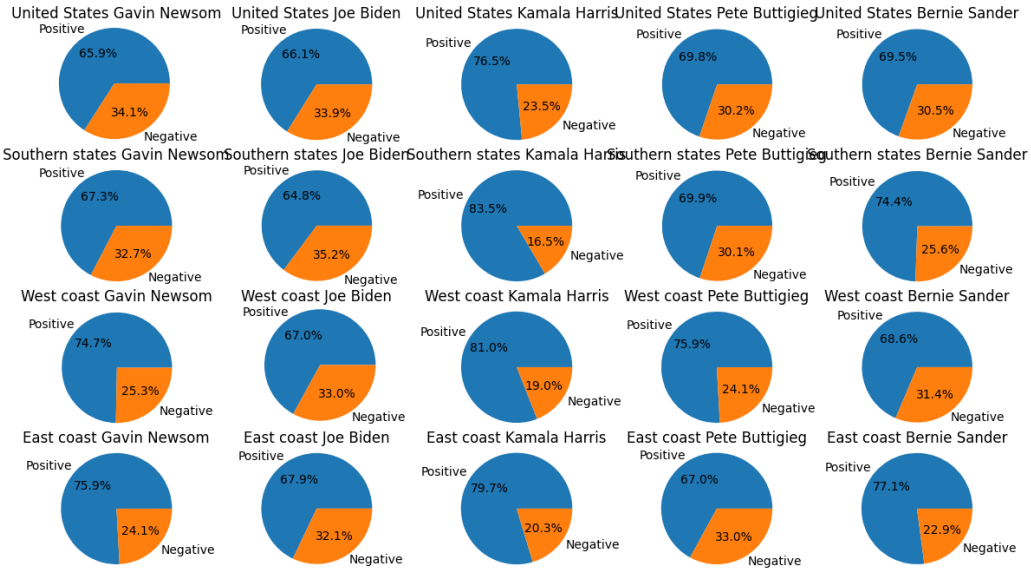


Figure 4.2: Analysis of Notable Political Figures in the United States

Looking at the results of our second analysis, we can see that our applications are producing a similar output which is shown in figure 4.1. However, when we repeated our analysis on Joe Biden, there was an increase in negative results from the Southern United States. This still left us pondering though. Is our application showing accurate data? In better words, can our application's output be compared to real-time political data? Figure 4.3 shows the approval rating as of the morning of April 28th, 2022. One possible factor of our positive-heavy results is due to the locations being too broad. Due to this, we have decided to focus on smaller areas in the country. Below, in Figure 4.4, you can see the analysis of Joe Biden in smaller cities in the U.S. This result shows that our tool can represent accurate political data on a general subject. It is also important to note that results are more likely to be accurate when the location parameter is contained in a smaller bounding box.

Now that the application has been tweaked to show more accurate results, we wanted to do further analysis on the political figures we had investigated in Figure 4.2. In this instance, because of the relevance, we chose to perform a twitter analysis on California Governor Gavin Newsom. This analysis also served as another opportunity to see if our application is outputting data that can be reflected in real-life political data. We hypothesized that if we used our analysis tool on different regions of California, we can see similar results to the recent recall

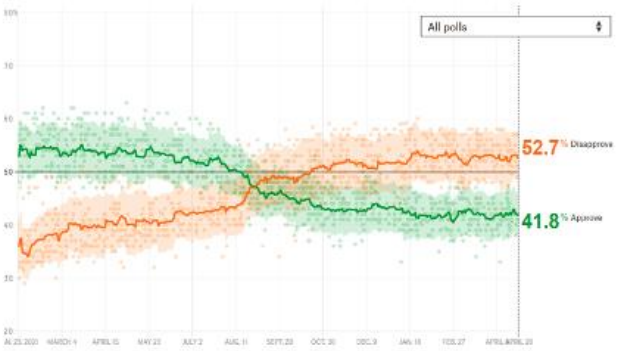


Figure 4.3: Joe Biden Approval Rating from FiveThirtyEight

that happened last summer. The figure below shows the results of the analysis of California's most populous areas: Los Angeles, San Fransisco, Lassen County, and Central Valley. The recall election resulted in Governor Newsom remaining in office, and as you can see from our analysis results, our program is outputting data that accurately reflects the political beliefs of Californians. Now that we have two examples of our tool working effectively and accurately, we wanted to use it for potential political predictions. To keep things simple, we decided to remain in the

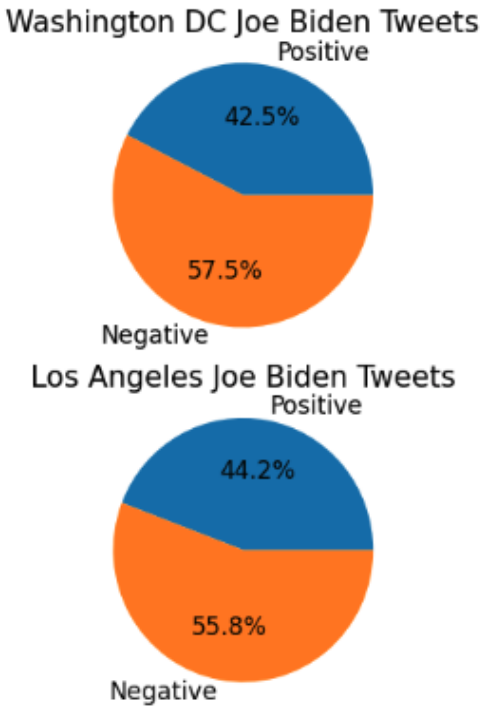


Figure 4.4: Joe Biden Analysis in smaller regions

state of California with some of the same political figures. We wanted to see if we can create an estimated electoral map of California regarding the upcoming governor election this November. As of May 2022, the top candidates consist of Gavin Newsom (D) and Jon Cox (R). Since it's already

indicated that our tool works best with smaller geographical bounding boxes, we decided to divide California into its three larger cities (Los Angeles, San Fransisco, and San Diego) and their surrounding suburban neighborhoods. The figure below represents the results of the analysis.

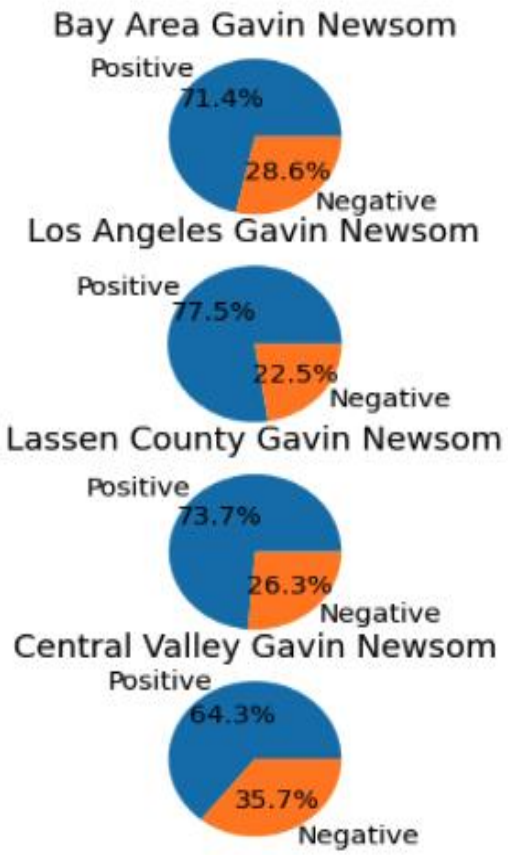


Figure 4.5: Gavin Newsom Analysis in California

Figure 4.6: Analysis of California Governor Candidates

With dark blue representing negative results and light blue representing positive results, the findings of these results show that there is a higher chance of Gavin Newsom winning the election in the fall. We came to this conclusion because Southern California suburbs was the only region with a negative result for Gavin Newsom. Additionally, this region only had a slightly higher positive result for Jon Cox. The rest of the regions leaned towards the current governor.

This political analysis tool is highly useful. When applied to topics on a local level, the results can be very insightful and offer potential predictions for results on future elections and other global events. With the evidence shown, this tool can be used to analyze different subjects and discourse online.

5. Sentiment Analysis

Sentiment Analysis is a crucial part for the project given that it provides sentiment score to collected tweets, which then yields data for analysis. For our project, TextBlob is used as the main sentiment analysis tool.

TextBlob is a python library for Natural Language Processing (NLP). It uses Natural Language ToolKit (NLTK) to achieve its tasks. TextBlob employs a lexicon-based approach, meaning that "it involves calculating the sentiment from the semantic orientation of word or phrases that occur in a text." (Jurek, A., Mulvenna, M.D. & Bi, Y.) The sentiment property of TextBlob returns a namedtuple of the form Sentiment(polarity, subjectivity). Our project only uses the polarity score, which is a float within the range [-1.0, 1.0], with -1.0 being most negative and 1.0 being most positive.

Our sentiment analysis setup categorizes each cleaned tweet into three main categories, positive, negative, and neutral. Then, the numbers of positive and negative tweets of each location are compared. If more positive tweets are collected, we conclude that the sentiment polarity of this location is more supportive toward the topic, and vice versa.

In addition to TextBlob, we incorporated another sentiment analysis tool named vaderSentiment. Similar to TextBlob, vaderSentiment is also lexicon-based. The goal here is to compare the results yielded by both tools, which, hopefully, improve accuracy of our conclusion. Figure 5.2 demonstrates the setup of vaderSentiment. Figure 5.1 demonstrates the setup of both sentiment analysis tools.

```python
def get_tweet_sentiment(tweet):
    # create TextBlob object of passed tweet text
    analysis = TextBlob(clean_tweet(tweet))
    # set sentiment
    if analysis.sentiment.polarity > 0:
        return 'positive'
    elif analysis.sentiment.polarity == 0:
        return 'neutral'
    else:
        return 'negative'
```

```python
def sentiment_scores(sentence):
    cleaned_tweet = clean_tweet(sentence)
    sid_obj = SentimentIntensityAnalyzer()
    sentiment_dict = sid_obj.polarity_scores(cleaned_tweet)

    if sentiment_dict['compound'] >= 0.05:
        return "positive"
    elif sentiment_dict['compound'] <= - 0.05:
        return "negative"
    else:
        return "neutral"
```

Figure 5.1 Sentiment Analysis Tools Setup

Despite the fact that both TextBlob and vaderSentiment are lexicon-based, different results are produced by simply replacing the sentiment analysis tool in our code. Figure 5.3 and Figure 5.4 are the results from TextBlob and vaderSentiment respectively on the topic of the war between Russia and Ukraine. From the result, it seems like the one generated using vaderSentiment makes

more sense. For example, in United States, the vaderSentiment graph says that tweets are more supportive of Ukraine over Russia, whereas the TextBlob graph suggests that both nations are supported, but Ukraine is more favorable.

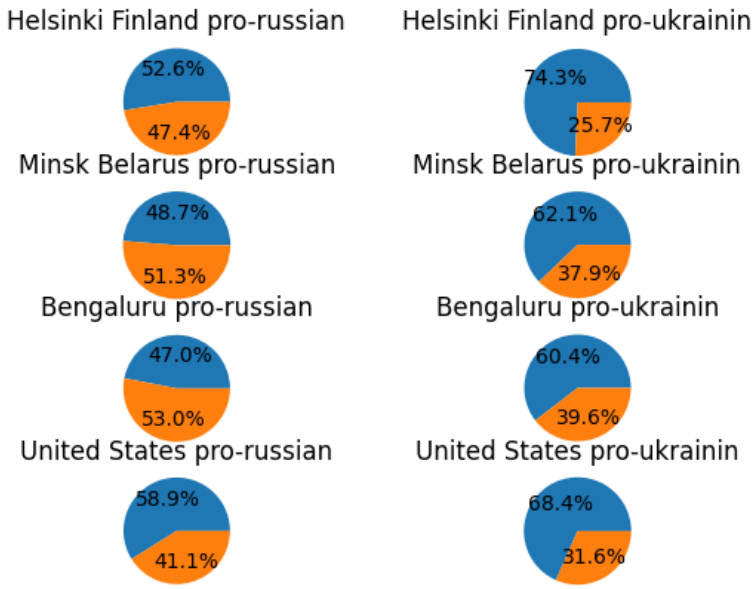


Figure 5.3 TextBlob Graph

Please keep in mind that multiple reasons could lead to this result. First, in these trials, the number of tweets collected for each location is 250. Knowing that so much more tweets are sent each second at these places, it is likely that the small example size could introduce a margin of error.

Also, notice that in the other three groups of comparison besides United States, it seems like TextBlob has a higher tendency of producing positive results, while vaderSentiment tends to produce negative results. Such different does not necessarily suggest that one sentiment analysis tool is more accurate than the other. We suggest to compare the results generated by different sentiment analysis tools in order to improve the accuracy of the conclusion.

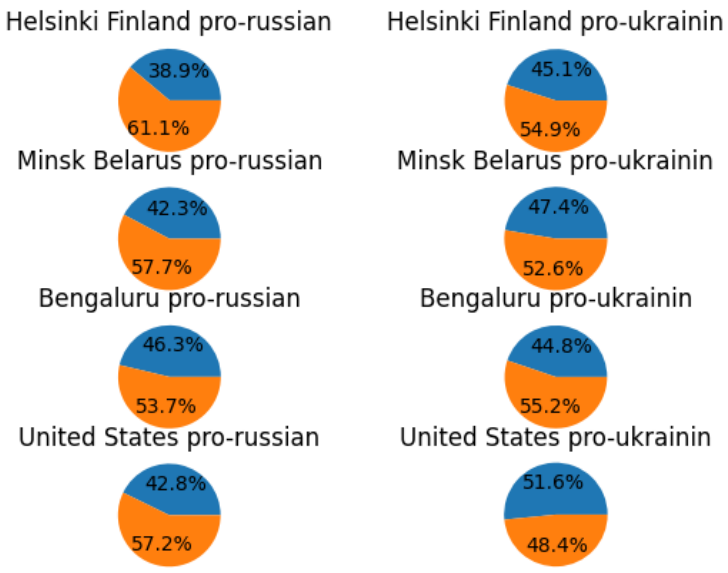


Figure 5.4 vaderSentiment Graph

6. Additional Analysis

Some additional analysis we decided to test our tool with was the Elon Musk prospective purchase of Twitter. Given that Twitter is one of the most popular apps in the world, and that the topic of running one of the largest media platforms in the world can definitely be considered political, we decided to research it.

As the current richest man in the world, Elon Musk has turned into a household name. While being incredibly successful and influential, he has quickly become one of the most polarizing names in today's media. Our goal was to simply use our tool to measure sentiment about his prospective purchase of twitter to see how the world is talking about the billionaire. Figure 6.1 below shows our initial results using VaderSentiment. Figure 6.2 Shows our results using TextBlob.

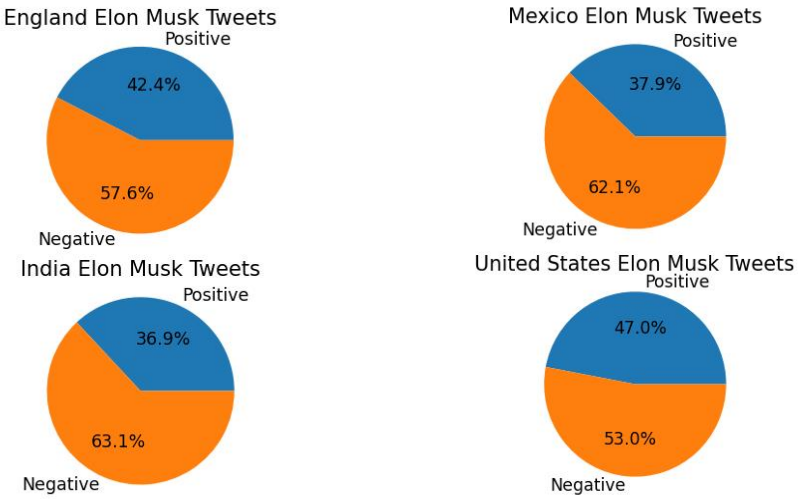


Figure 6.1 10,000 Tweet Analysis using VaderSentiment

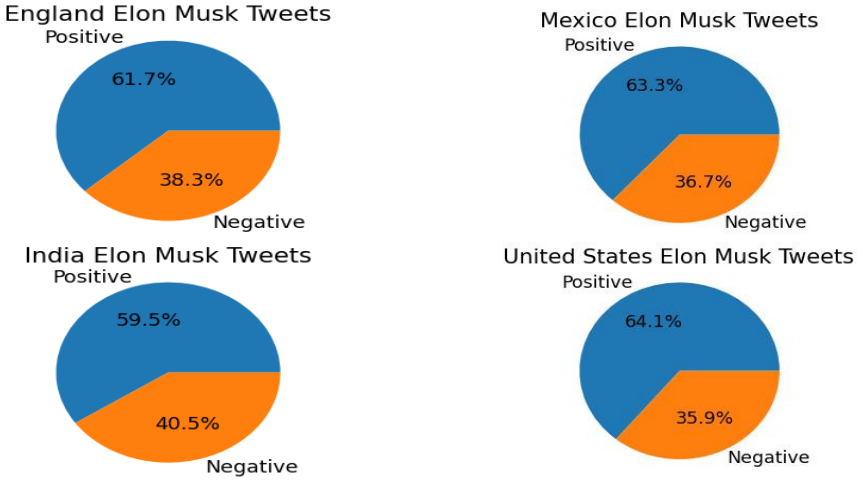
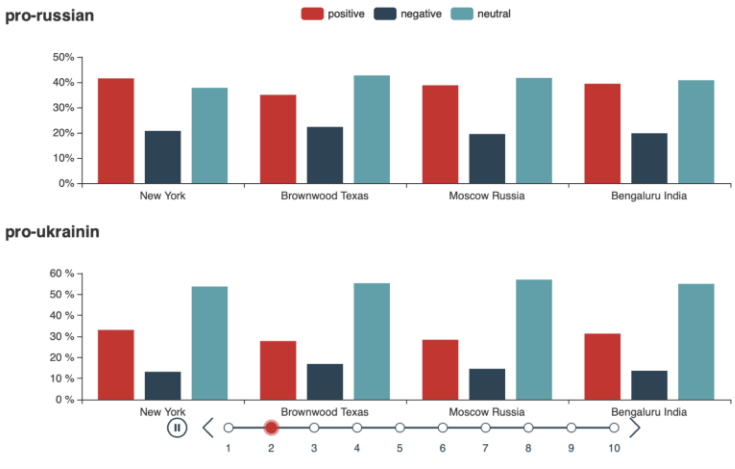


Figure 6.2 10,000 Tweet Analysis using TextBlob

These results again show the positive bias of the TextBlob sentiment tool and the negative bias of the TextBlob sentiment tool. We decided to compare these results to an existing Sentiment Analysis experiment on the Elon Musk buyout conducted by Tufts University. In their experiment, which consisted of analyzing over four million tweets from March 1 to April 27, they concluded that most tweets related to the Elon Musk Twitter takeover were mostly negative. Because of this, we determined that VaderSentiment is most likely the more accurate tool for our experiments.

Additionally, we tested our tool with the situation between Ukraine and Russia, and we used data visualization to present the data. The content of this visualization is the changing situation of the attitude in different areas between Ukraine and Russia. Ten groups of data are included in the

image below, and we used a timeline to display those ten groups of data dynamically. Four areas and three kinds of attitudes are included in the display, four areas are New York, Brownwood Texas, Moscow Russia and Bengaluru India, three attitudes are positive, negative and neutral.




```
paths=[os.path.join(dir_path,i) for i in os.listdir(dir_path)]
t=(
    Timeline()
    .add_schema(is_auto_play=True)
)
index=0
for i in paths:
    index+=1
    f=open(i,'r',encoding='utf-8')
    data=f.read()
    f.close()
    city=re.findall(r'Location: (.*?)\n',data)
    positive_tweet_percentage=re.findall(r'Positive tweet percentage: (.*?)\n',data)
    negative_tweet_percentage=re.findall(r'Negative tweet percentage: (.*?)\n',data)
    neutral_tweet_percentage=re.findall(r'Neutral tweet percentage: (.*?)\n',data)
```

For the realization of this display, we used pyecharts as our tool, the elements included timeline component, bar and grid (combination of chart). We used two bars to represent pro-Russia and pro-Ukraine, and used grid module to combine the two bars, and for the outside, we used a timeline component, walked through all of the data files, each file corresponde a time spot, then formed the visualization finally.

```
bar1 = (
    Bar()
    .add_xaxis(city)
    .add_yaxis("positive", [round(float(i),2) for i in positive_tweet_percentage[::2]])
    .add_yaxis("negative", [round(float(i),2) for i in negative_tweet_percentage[::2]])
    .add_yaxis("neutral", [round(float(i),2) for i in neutral_tweet_percentage[::2]])
    .set_series_opts(label_opts=opts.LabelOpts(is_show=False))
    .set_global_opts(
        title_opts=opts.TitleOpts(title="pro-russian"),
        yaxis_opts=opts.AxisOpts(
            axislabel_opts=opts.LabelOpts(formatter="{value}%")
        )
    )
)
bar2 = (
    Bar()
    .add_xaxis(city)
    .add_yaxis("positive", [round(float(i),2) for i in positive_tweet_percentage[1::2]])
    .add_yaxis("negative", [round(float(i),2) for i in negative_tweet_percentage[1::2]])
    .add_yaxis("neutral",[round(float(i),2) for i in neutral_tweet_percentage[1::2]])
    .set_series_opts(label_opts=opts.LabelOpts(is_show=False))
    .set_global_opts(
        title_opts=opts.TitleOpts(title="pro-ukrainin", pos_top="48%"),
        yaxis_opts=opts.AxisOpts(
            axislabel_opts=opts.LabelOpts(formatter="{value} %")
        )
    )
)
grid = (
    Grid()
    .add(bar1, grid_opts=opts.GridOpts(pos_bottom="60%"))
    .add(bar2, grid_opts=opts.GridOpts(pos_top="60%"))
)
t.add(grid,str(index)).render("tweet.html")  #the path and file name of the saved pages
```

7. Conclusion

Our goal for this project was to be able to create a tool that can accurately analyze the sentiment of tweets and use it to come up with some sound conclusions.

We mainly used political topics to filter tweets for our project. However, we are confident that our project will generate useful information on topics of other fields by modifying the filter words of tweets streaming section, i.e. product satisfaction, market preferences, etc.

In these regards, we feel that we have accomplished what we set out to do.

8. Future Scope

In this project we were focusing on creating a simple tool that was firstly robust and reusable. This means that our program is a perfect base for adding additional features. We agreed that, after lots of testing, additional features for data visualization as well as better tweet cleaning features are needed.

In our project we used a simple pie graph visualization to show our data. This gets the jobs done but is far from ideal when surveying a large number of search groups. Since our project searches

based on location, we think adding a data visualization tool to map results to the location used would be very useful. This would simplify the data and make differentiating much easier for non-technical users.

Such a map overlay could make our tool useful for showing how opinions in the country differ between state/county/city, specifically for political issues. The tool could also create maps based on electoral districts for the purpose of predicting outcomes.'

Regarding to the map plotting tool, we are currently looking at Geopands and Plotly. These two python libraries will allow us to plot map based on our results. We are still having problems getting it to work, as some dependencies required by Geopandas are unable to install. Below are some graphs that demonstrate how it potentially works. Figure 8.1 is a population density map Florida state based on data of each county. Figure 8.2 is shows the employment rates of United States.

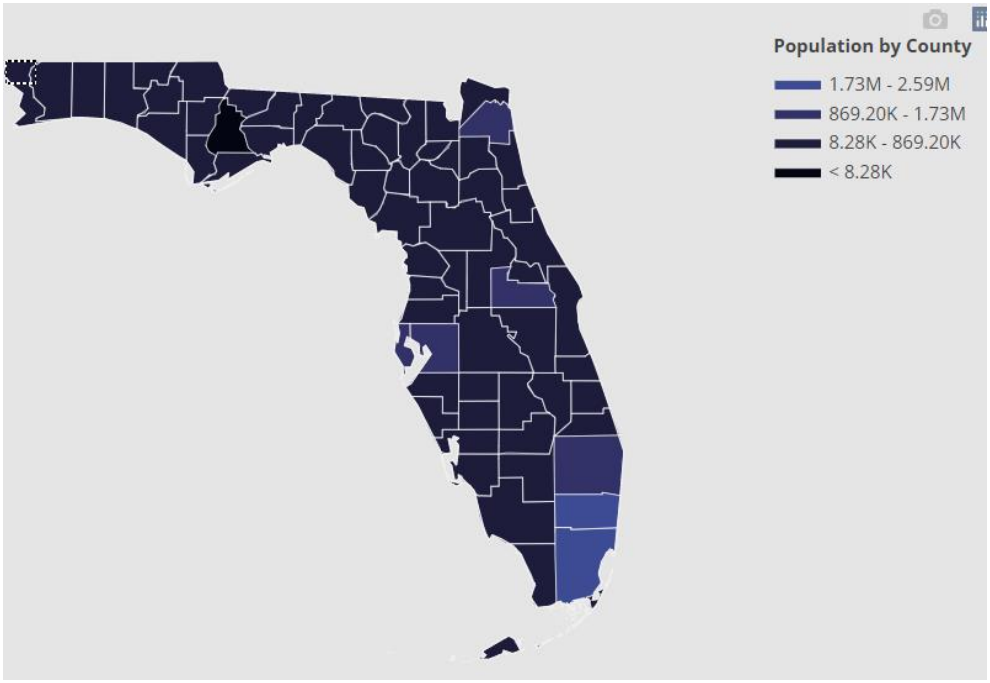


Figure 8.1 Florida Population Density

With the python libraries working, we need to convert collected data into an RSV file, which then can be easily plotted using tools provided by Geopandas.

Another improvement could come with the addition of greater tweet cleansing. Our current process is necessary for removing excess characters that could influence the sentiment analysis tool. While our process is sufficient, there are areas where it could be improved.

Some users write tweets that are longer then the 240 character limit and split them up into multiple tweets called a "thread". The filtered stream API will fetch the first tweet but the second may not be fetched if it doesn't have the search term in it. The second part of the tweet may have important context for sentiment analysis so without the second, the analysis of the first may

return an incorrect result. In a future version, we could implement a method to retrieve an entire thread if it is suspected that the tweet fetched is incomplete.

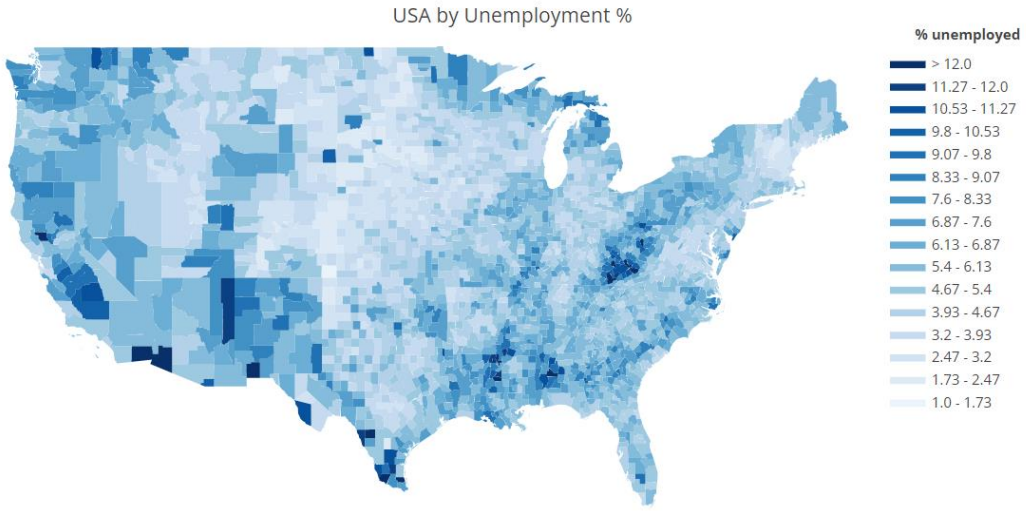


Figure 8.2 United States Unemployment Rate

Bibliography

Jurek, A., Mulvenna, M.D. & Bi, Y. Improved lexicon-based sentiment analysis for social media analytics. Secur Inform 4, 9 (2015). https://doi.org/10.1186/s13388-015-0024-x

The elon musk-twitter takeover saga: A multi-country sentiment analysis of Twitter users. Digital Planet. (2022, May 9). Retrieved May 10, 2022, from https://sites.tufts.edu/digitalplanet/the-elon-musk-twitter-takeover-saga-a-multi-country-sentiment-analysis-of-twitter-users/