Topic:
Sentiment Analysis of Three Chosen Automobile Companies

Ziming Ping
N19463405; zp584@nyu.edu

Tingyu Yang
N12235013; ty2069@nyu.edu

Xinyu Zhao
N12924775; xz2671@nyu.edu

Date: 12/09/2020

Table of Contents

Part 1. Introduction

Basic idea: Sentiment Analysis Based on the Social Media

This task was basically an NLP (Natural Language Processing) project. During this project, our team will be responsible for performing some sentiment analysis tests for the public attitudes towards several famous automobile companies in the USA in the number of 3. The choices of those companies should be well-known and have some individual differences. We decided to perform this analysis test on Twitter as the social media platform because Twitter provides the developers with the legit web crawler interface called Twitter API to collect the useful information behind those public tweets but only for the purpose of legal research or study. In other words, the Twitter API provides the tools for us to contribute to, engage with, and analyze the real-time conversation happening on Twitter. We plan to collect between 250- 400 public real-time tweets for each of those 3 automobile companies by using the API.

The Necessary Steps that We Have Designed

First, we must build a pipeline in order to clean up the raw data such as removing useless symbols and unmeaningful words by compiling a function named Word_Clean in Python.

Then, we need to introduce the concept of the NLP model to calculate the polarity and subjectivity of each meaningful word within the sentences. We have made some research and we all think TextBlob could probably be a good choice for this sentiment test. However, we need to figure out a proper method to improve its accuracy such as providing it with those efficient training data sets. After that, the program should show us the polarity of each sentence by judging whether each of the tweets is positive, neutral, or negative. Moreover, we will define a kind of metric to calculate the sentiment scores of each selected company based on the sum of all the total tweets' scores. For example, every Positive tweet +2 points; every Neutral tweet +0.5 point; every Negative tweet -2 points. Based on the final sentiment scores, there will be a ranking among those companies for us to understand the real-time public attitude towards each of them.

Finally, we think we can also create a word cloud for each of those automobile companies based on the keywords within all the tweets after the cleaning process, which reveals the company's true figure behind their brands suggested by their clients. As for this task, the greatest challenge for us should be the training process and the proper use of the NLP technique.

Part 2. Our Choice & Task Performance Process

Selection of the top 3 automobile companies

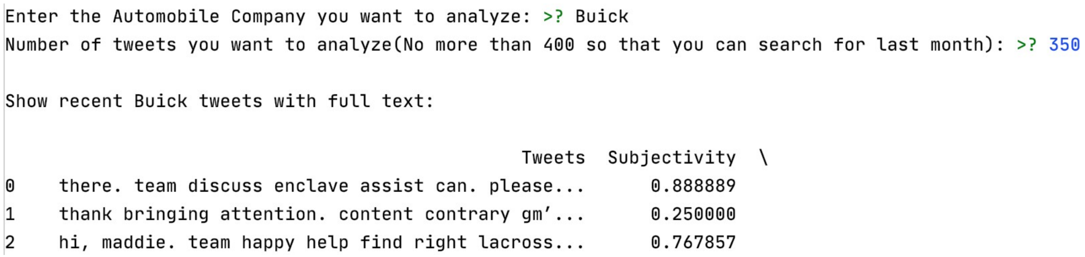
1. Mazda USA
2. Buick
3. Ford

Reason: The choice of these three famous automobile companies are mainly based on the cars.com—a well-known App regarding new and pre-owned cars' sale. These three automobile companies rank the top 3, recommended by cars.com in the year of 2020. We have a very good user experience with cars.com. Therefore, the result from cars.com looked reliable to us, and as a result, we decided to choose these three companies for our sentiment analysis. However, our program still can calculate the scores and so the relevant analysis for any other official account on Twitter.

Language Use and Logical Concept

We used the python (3.6) for the analysis tool for this project.

1) It will ask you about the company you want to make the research and the number of tweets you want to catch. (Must be the official name of its Twitter account)
2) It can then catch at most 400 tweets from the Twitter (no more than 400 tweets at one time because our analyses were mainly based on the tweets from last month, probably around 350 tweets)



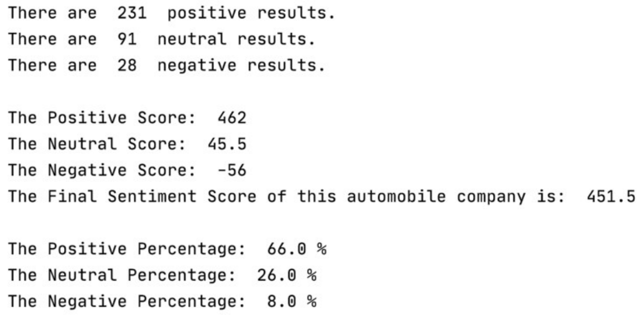
```
Enter the Automobile Company you want to analyze: >? Buick
Number of tweets you want to analyze(No more than 400 so that you can search for last month): >? 350

Show recent Buick tweets with full text:

                                              Tweets  Subjectivity  \
0    there. team discuss enclave assist can. please...      0.888889
1    thank bringing attention. content contrary gm'...      0.250000
2    hi, maddie. team happy help find right lacross...      0.767857
```

3) Then it will clean the data with our defined data_clean function and create a DataFrame to display the cleaned data
4) It proceeded the sentiment analysis by importing TextBlob package—a famous natural language processing package which can be used for sentiment analysis in Machine Learning. Our program defined the cleaned tweets' polarity as positive, neutral and negative, also along with its subjectivity. we used the existed lexicon supported by TextBlob for the scoring.
5) Then, add the corresponding subjectivity and polarity into the previous DataFrame and display the Scatter diagram.
6) We set the scoring rule that Positive—gain 2 points, Neutral-gain 0.5 point and Negative lose 2 points in order to calculate the total scores of the specific company.
7) After that, the program will generate the relevant graph regarding to the score results. we used the same amount of the data for these 3 companies—350 tweets.
8) It can calculate the scores and shows the details of each positive, neutral and negative points. Then, it will generate the graph of scores of this company.
9) After the score part, it will finally generate a word cloud in the background of black related to the company by using the cleaned data and display the word cloud.

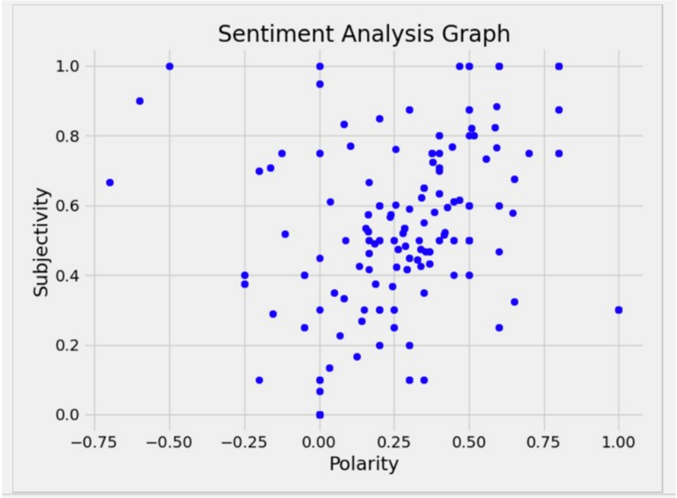
We removed the unnecessary components of the sentences such as special characters, repeat blank, https link, "an, a…" and those redundant stop words in the sentences so that we can increase the accuracy of the program.

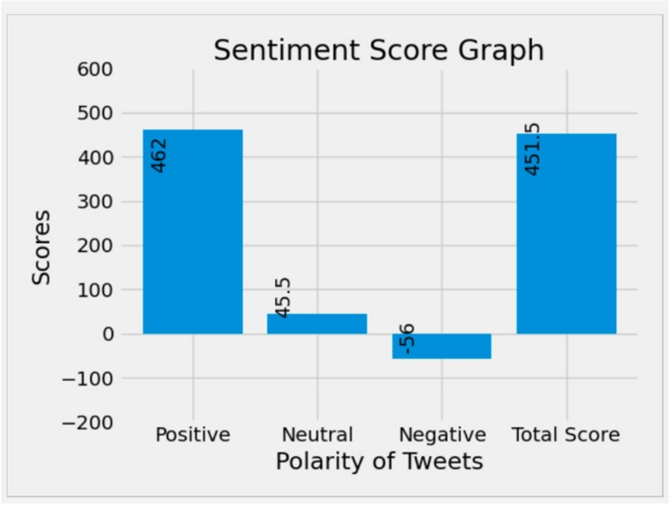
Part 3-Scores, Graphs Word Cloud.

1. MazdaUSA—451.5 points

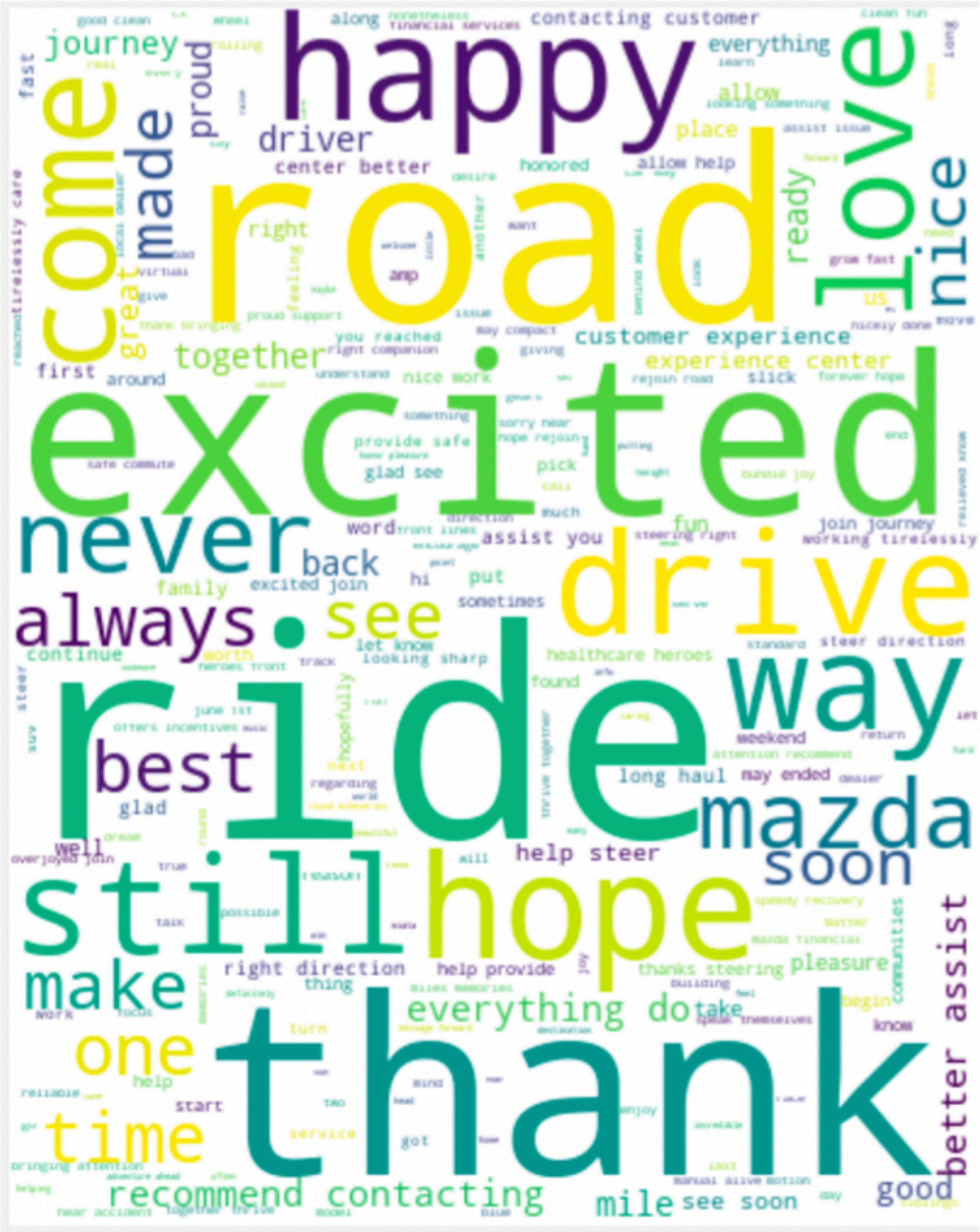


```
There are  231  positive results.
There are  91  neutral results.
There are  28  negative results.

The Positive Score:  462
The Neutral Score:  45.5
The Negative Score:  -56
The Final Sentiment Score of this automobile company is:  451.5

The Positive Percentage:  66.0 %
The Neutral Percentage:  26.0 %
The Negative Percentage:  8.0 %
```

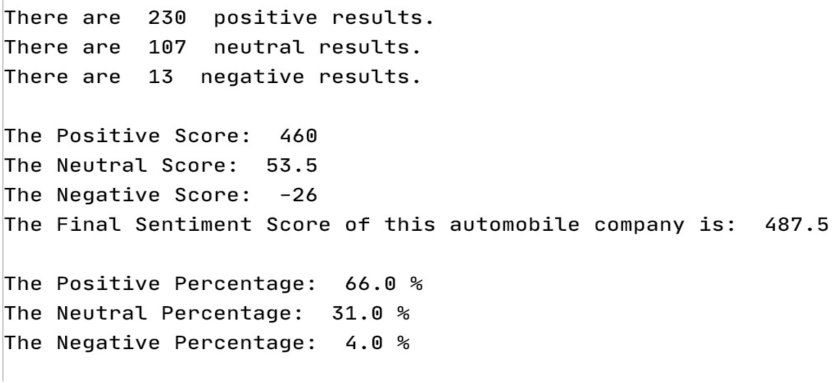




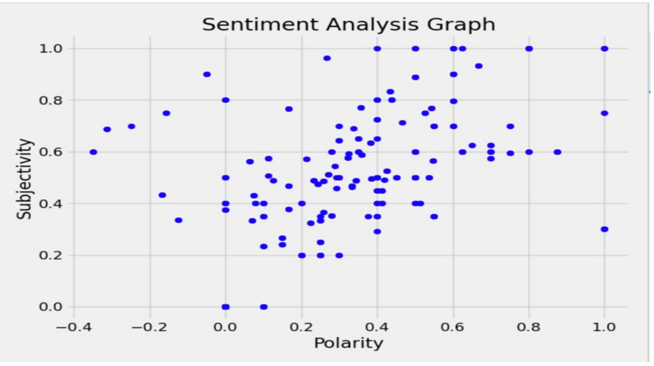
Mazda USA’s Word Cloud:

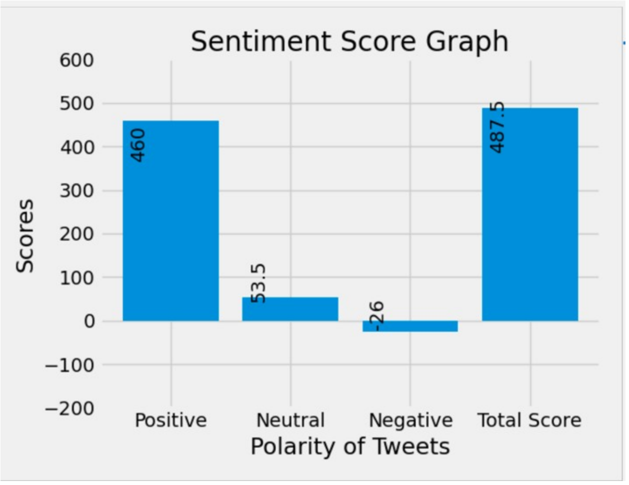


2. Buick—487.5 points

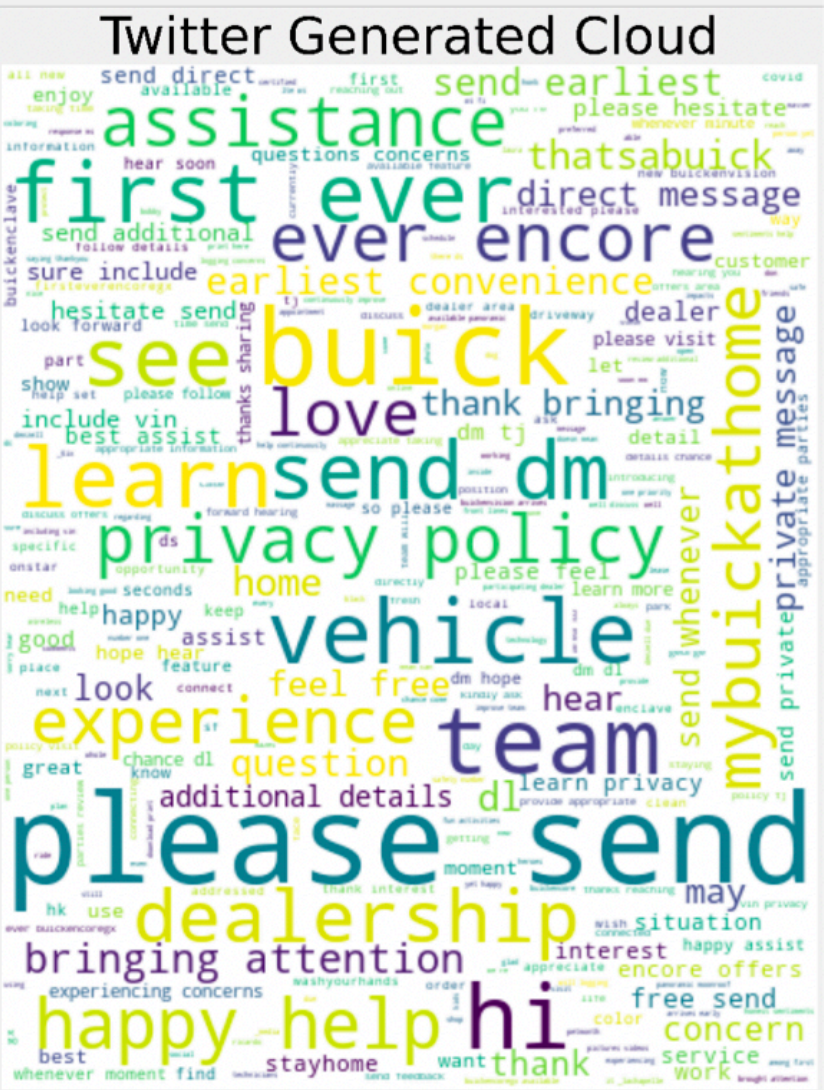


```
There are  230  positive results.
There are  107  neutral results.
There are  13  negative results.

The Positive Score:  460
The Neutral Score:  53.5
The Negative Score:  -26
The Final Sentiment Score of this automobile company is:  487.5

The Positive Percentage:  66.0 %
The Neutral Percentage:  31.0 %
The Negative Percentage:  4.0 %
```

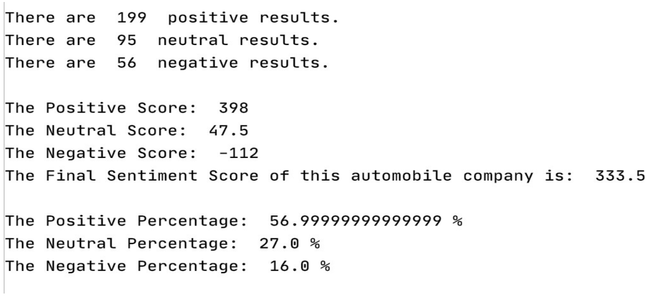




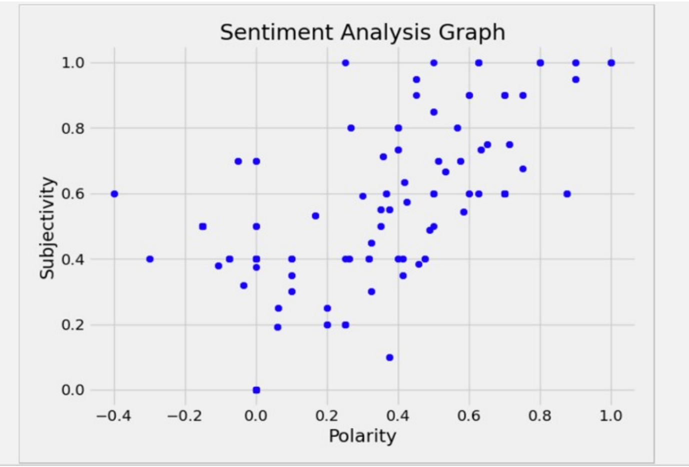
Buick’s Word Cloud:

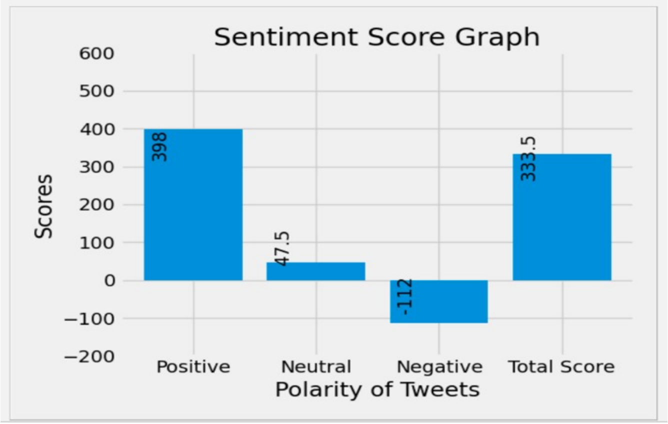

Twitter Generated Cloud

3. Ford—333.5 points

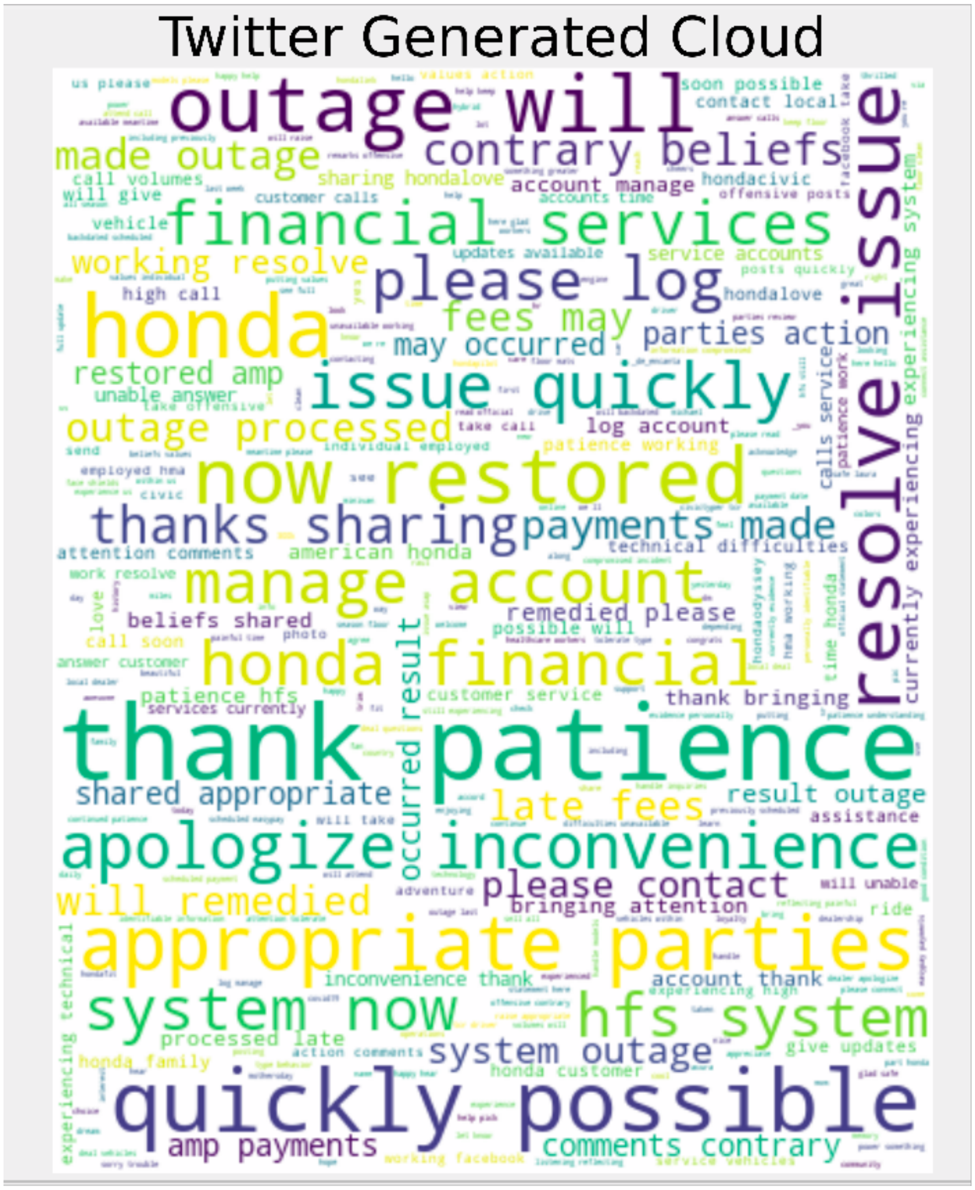


```
There are  199  positive results.
There are  95  neutral results.
There are  56  negative results.

The Positive Score:  398
The Neutral Score:  47.5
The Negative Score:  -112
The Final Sentiment Score of this automobile company is:  333.5

The Positive Percentage:  56.99999999999999 %
The Neutral Percentage:  27.0 %
The Negative Percentage:  16.0 %
```





Ford’s Word Cloud:



Part 4: Analysis & Conclusion:

In calculation of the final sentiment scores of these three companies based on 350 tweets, Mazda and USA Buick achieve the pretty high scores (451.5 & 487.5 points). However, Ford's score (333.5 points) is the lowest one compared to the other two companies based on 350 tweets.

Sensitivity & Polarity:

The program used the **TextBlob NLP** model to capture global syntactic dependencies and semantic information, based on which the weight of each sentiment word together with a sentence-level sentiment bias score are predicted.

Sensitivity: Range = [0,1] and No negative values.

It is also called the true positive rate, and also represents the probability of detection in those tweets that measures the proportion of actual positive rates that are correctly identified among all. It is calculated with the help of NLP models in English language trained by TextBlob.

Therefore, when it comes to this case, sensitivity refers to the ability to designate each tweet in the samples with as positive. A higher sensitive result means that this tweet tends to become more positive compared to other samples showing a lower sensitive number.

Polarity: Range [-1,1]

Polarity is defined as having two opposite tendencies or opposite attitudes such as positivity (+1) or negativity (-1) or neutrality (0) of that keywords. When it goes up to +1, it means the expression of the sentence detected by NLP model tends to be positive. On the other side, the express tends to be negative if it goes up to -1. The value of 0 means the middle value, which represents the neutrality. Our scoring rules basically depends on the results of polarity. If the company comes out more positive results and less negative results and it can gain more scores. As for these three selected companies, the Ford has the most ratio of negative results (16%) compared to the other two. Therefore, it has the lowest score, but Buick has the highest one.

Company’s Strategy Analysis:

After calculating the scores and generating the word cloud of the 3 companies related to their published official tweets and their daily replies. It is clearly that Buick and Mazda show a much better sentiment figure than Ford, at least on Twitter. In other words, they have built a good reputation in the social media environment and set up a very positive figure to the public. According to its word cloud, the positive key words are mostly “happy”, “help”, “dealership”, “protect privacy”, “team”, “good services”. As for the negative word, there is almost no obvious negative words showing in the word cloud. Therefore, we can conclude that they have successfully build up their reputation by offering good services. When it comes to Ford, its word cloud contains fewer positive words. Instead, most of the key words are some neutral or negative words such as “inconvenience”, “restored”, “issue”, “late fees” “outage”. It is reasonable to infer that Ford usually brings more issues or bad experience to its customers. It is important and immediate for them to improve their customer service and build a much better population to the public.

Reference

Twitter API Documentation. (2010). Docs | Twitter Developer. https://developer.twitter.com/en/docs/twitter-api

Jain, S. (2020, December 8). *Natural Language Processing for Beginners: Using TextBlob*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/

Hsu, S. (2018, September 11). *Introduction to Data Science: Custom Twitter Word Clouds*. Medium. https://medium.com/@shsu14/introduction-to-data-science-custom-twitter-word-clouds-704ec5538f46