Tweets-based Regional Brand Analysis: Starbucks and Dunkin' Donuts

Fan Wu, Geqi Yan, Xi Yang
Electrical Engineering, Electrical Engineering, Earth and Environmental Engineering
Columbia University
fw2322@columbia.edu, gy2266@columbia.edu, xy2378@columbia.edu

Abstract—**In our project, we choose two giant coffee brands, Starbucks and Dunkin' Donuts, as our research objects. We focused on a brand analysis based on tweets from different states of United States. The first step is gathering the data, we collected our raw data using Twitter API and Python, this is known as 'data mining'. Spark machine learning modules support us to analyze these two brands. In detail, we first preprocess the raw data by using Regex Tokenizer, Stop Words Remover, Hashing TF-IDF, then we do sentiment analysis based on Native Bayes Model for each brand. In order to compare these two brands comprehensively, we visualize our results using D3.js to draw our conclusion on the popularity map and then put the results of these two brands together to get the conclusion.**

Keywords-Twitter API; Brand; Sentiment Analysis; Visualization; Popularity Map.

I. INTRODUCTION

Nowadays, the type and scale of data in human society is growing at an unprecedented speed with the emerging services such as cloud computing, the Internet of things and social network.[1] 'Big data' has become the most popular research fields in modern society, which shows that the era of big data has come.

Meanwhile, social media has completely changed the way people interact with information, and become the most essential part of our daily life. More and more people use social media to share their own perspectives and ideas. It is one of the best ways to view social media data, which refers to all the raw insights and information collected from individual's social media activity, as a source of raw data to do some relative analysis, such as sentiment analysis.[2]

Twitter is much like a data gold mine because it provides a free and public information platform for users. That is, unlike other social platforms, almost every Twitter user's tweets are completely open and accessible. So, if we want to try to collect a lot of data and analyze it, that's immensely helpful. Twitter API allows us to perform complex queries to get some very specific tweets, in other word, the specific data you want to analyze. For example, we can collect all the tweets that mention a specific topic. In addition, we can also collect tweets from target users who live in a certain location, which is known as spatial data. As you can see, Twitter data can be a very helpful source for analysis, and also can produce powerful results based on it.

Considering these advantages above, in our project, we select Twitter as the source of our dataset for data mining.[3] And we choose the text and location features as our analysis objects to do regional sentiment analysis based on spatial data. As for the popularity analysis of Starbucks and Dunkin' Donuts, we analyze people's different attitudes towards one specific brand using big data tool such as machine learning module of spark for sentiment analysis, python modules and D3.js for visualization.

II. RELATED WORKS

Sentiment analysis based on tweets data has been used for prediction or measurement in a variety of domains, such as stock market, politics and social movements.[4] For example, Hao Wang developed a system for real-time analysis of public sentiment toward the presidential candidates in the 2012 U.S. election as expressed on Twitter.[5]

Sentiment analysis of tweets data is much harder than that of traditional text like review documents, which is partly because of the short length of tweets, the use of informal and irregular words and symbols, the abbreviation of several words, etc. Also, there are various methods for training sentiment classifiers for datasets that comes from Twitter. For example, Naïve Bayes (NB), Maximum Entropy (MaxEnt) and Support Vector Machines (SVMs).[6] In particular, Naive Bayes is a simple model for classification. It is simple and works well on text categorization.

Our project focuses on the sentiment analysis of the tweet data. Previous researches regarding sentiment analysis place more importance on prediction and classification. Our work, however, use sentiment analysis to analyze the commercial value of two coffee brands: Starbucks and Dunkin' Donuts. Instead of simply concentrating on sentiment analysis, we also use special visualization method to draw a popularity map to study the spatial variance of brand popularity.

III. SYSTEM OVERVIEW

In the first part of system, we mainly prepare the raw Twitter data. Firstly, we collect data from Twitter Streaming API, then do data preprocessing for classification. As a result, we get the processed tweets for further analysis.

In the second part, we preprocess the training data csv file, then use it to train sentiment classifier and optimize it using PySpark.

Thirdly, we combine the above two parts together, and run classifier on collected tweets. We implement sentiment analysis of tweets data to predict the development prospect of two brands, and also calculate the total number of tweets.

In terms of the last part, we implement data visualization and do analysis based on the results.

IV. ALGORITHM

A. Data collecting

At the very beginning of our project, we utilize Tweepy, a python library for accessing the Twitter API, to collect tweets containing chosen brand name in their text from different states (shown below).

```
public_tweets = api.search(q="place:%s AND McDonalds" % place_id)
```

Figure 1. Code of Collecting Data using Tweepy

We use two filters (location, brand name) as parameters of the search function to search for the wanted tweets from Twitter API rather than employing Twitter Streaming API for following reasons. First, the Official Twitter Streaming API has a limitation of data acquisition which restricts us to get the oldest data within 7 days. Therefore, we use Twitter API as our main tool for data study and implement a python data crawler to form the dataset. Apart from that, the streaming API with two filters is rather slow when fetching tweets, thus making the whole collecting process too long for our project. Finally, we got 96 CSV files as our raw dataset, of which 48 are for Starbucks and the other 48 for Dunkin' Donuts. Each file contains the tweets we collect from one state in America except Hawaii and Alaska (sample shown below). To sum up, there are 295402 records (37.5M) related to Starbucks and 145571 records (19.1M) for Dunkin' Donuts. Overall, the whole dataset contains 440973 records (56.6M). We extract the following features of each tweet:

1. Location: The State the tweets come from.
2. Text: The content of the tweets.

```
index,location,text
0,New York,wanna take my new #daughter on #vacation and dunk her in the oce
https://t.co/IZAn0PSN0B
1,New York,@melissabethk @Kbratskeir @BreadsBakery Queens Pita in... Queens
Donuts more" 🐱
2,New York,TFW u definitely meant to say medium iced coffee @ Dunkin but ac
3,New York,Breakfast 🍳 stop 🛑 🥐🥓🍩☕ #usa #supersinghs🙏🕉us @ Dunki
4,New York,"Heroism is caring for the world, without expecting the same in
https://t.co/WjvbpydcSV"
5,New York,Dunkin on BCD in the morning .... horror
```

Figure 2. Collection Dataset Sample File for Starbucks in New York

B. Sentiment analysis

After going through some reading and studies on related work about sentiment analysis, we choose Naive Bayes classification model in ml.spark library to perform the Sentiment Analysis in our project. The whole process consists of two parts, the first of which is using open source dataset containing texts with labeled sentiment to train the classification model in spark, and the second is employing the model to predict the sentiments of our own dataset.

❖ **Sentiment Classifier Model Training**

We use the Twitter Sentiment Analysis Dataset for model training in Spark. The dataset is based on data from the following two sources:

1) University of Michigan Sentiment Analysis competition on Kaggle.
2) Twitter Sentiment Corpus by Niek Sanders.

The Twitter Sentiment Analysis Dataset contains 1,578,627 classified tweets, each row is marked as 1 for positive sentiment and 0 for negative sentiment. After preprocessing (which will be covered in the following part) the training data, we first split the processed dataset into 2 parts. The 0.85 of it are used to train the Naive Bayes classification model, and the other 0.15 are used to test the model's justness with the indicator Accuracy:

$$\boxed{Accuracy = \frac{TP}{TP + FP}}$$

----True Positive (TP) - label is positive and prediction is also positive

----False Positive (FP) - label is negative but prediction is positive

After the training and testing process, we finally get our model which reaches an Accuracy of 0.7381624187948265. Generally speaking (and particularly when it comes to social communication sentiment classification), 10% of sentiment classification by humans can be debated, so the maximum relative accuracy any algorithm analyzing over-all sentiment of a text can hope to achieve is 90%, thus this is not a bad starting point.

❖ **Sentiment Prediction**

1. Preprocessing

The preprocessing process is:

Data Cleaning → Regex Tokenizer → Stop Words Remover → Hashing TF-IDF

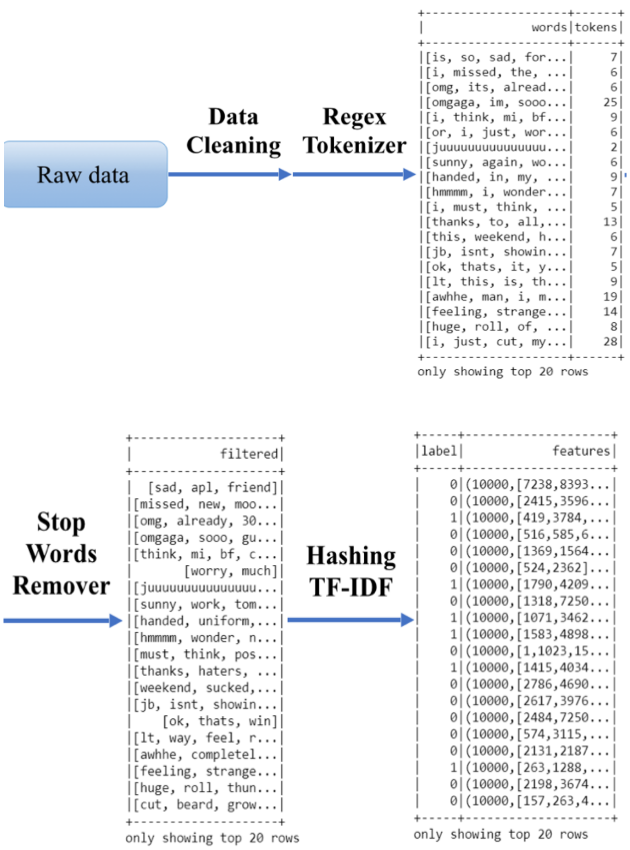

Figure 3. Preprocessing pipeline overview

Data Cleaning

Before applying Regex Tokenizer, at first, we need to do some basic data cleaning to remove components that affect final results. Data cleaning can be divided into several steps:

- Removing HTML Characters: HTML related components such as &gt, &amp are useless characters for sentimental analysis. We utilized html parser to remove these entities.
- Decoding Data: For sake of easy understanding, we need to decode complex symbols in tweets into standard, simple and understandable characters. UTF-8 encoding is chosen since it is the most widely accepted data decoding method.
- Finding Apostrophe: Apostrophe is ambiguous under certain circumstances. To get rid of ambiguation, we need to transferring apostrophe into a uniform and standard form.
- Finding Slangs: Slangs are understandable to human beings, but they can be hardly understood by computer. So, we need to standardize them.
- Finding created words: Tweet users sometimes create words by themselves. To help computer understand these words, we need to convert them into standard format.

Regex Tokenizer

Compared with original tokenizer which could split sentences into words. It is better to use regex tokenizer for tweets, which use regex to judge the split positions. Since tweet data are often less standard than other articles, including extra space and symbols which are hard to split, regex tokenizer proves to be more effective.

Remove stop words

Tweet data has many stop words to be removed. In sentiment analysis, high frequency words without sentimental meaning might cause great error in sentiment prediction.

Hashing TF & IDF

Hashing TF-IDF is a feature vectorization method widely used in text mining to reflect the importance of a term to a document in the corpus. Term frequency TF is the number of times that term t appears in document d. If a term appears very often across the corpus, it means it doesn't carry special information about a document. Inverse document frequency IDF is a numerical measure of how much information a term provides. In our project, we employ such methods for converting words into vectors, which would be directly used in the following sentimental classifying and prediction.

2. Prediction and Results

With featured vectors for each tweet generated before, we are able to classify them as positive or negative with our Bayes Classifier (Binary Classification). Up to now, all tweets' sentiments have been predicted and could be used for statistics and analysis.

After the prediction, we collect the results for each file and get following properties of it referring to each state for a specific brand name:

Positive: number of positive tweets in a file
Negative: number of negative tweets in a file
Count: total number of tweets in a file
Ratio: ratio $= \frac{positive}{count}$ ----to some extent reflecting the satisfaction score (0~1) of one state towards a chosen brand

Processed results of sentiment analysis are shown below:

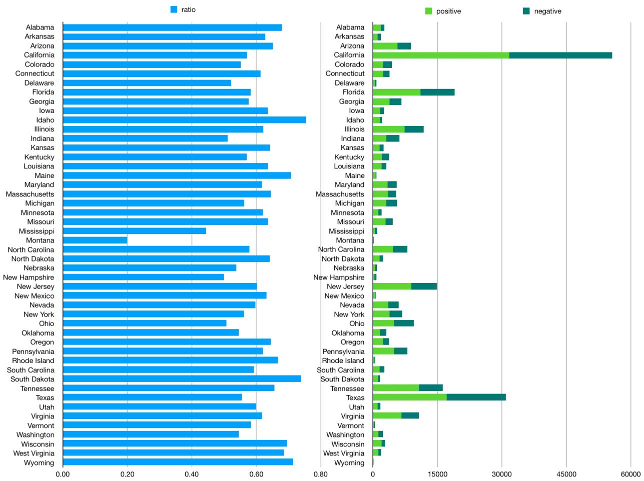

Figure 4. Collecting Results over 48 States for Starbucks

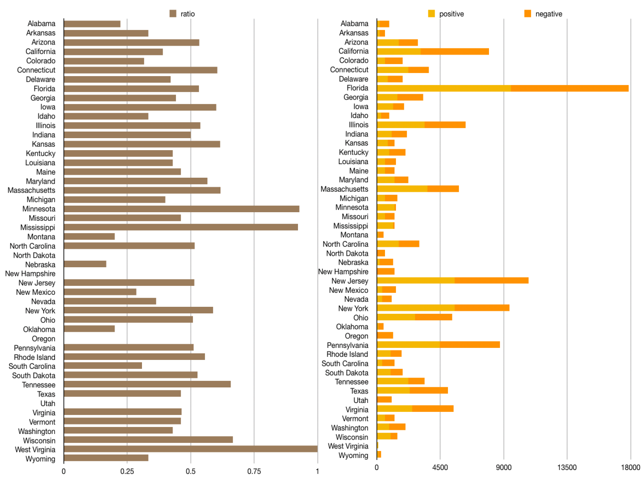

Figure 5. Collecting Results over 48 States for Dunkin' Donuts

V. VISUALIZATION

This project is aimed to analyze the popularity of two brands in different states in the US. The visualization therefore mainly focuses on manifesting the spatial variation of brand popularity. Instead of simply using charts, popularity maps are drawn to demonstrate the popularity of the brands in different regions.

1. Drawing underlay

To draw a popularity map, the first step is to draw an underlay. States are drawn based on the geographical information. The abbreviation of state name, the name and the geographical information of all the states are stored for reference. Since d3 does not support geo data, shapefiles which store the original geographical information should be converted to GeoJSON and then Topojson for further use.

```
(function(){
    var uStatePaths=[
 {id:"FL",n:"Florida",d:"M759.8167,439.1428L762.08236,446.4614L765.81206,456.20366L771.14685,465.57996L774.86504,471.88472L779.714
{id:"NH",n:"New Hampshire",d:"M880.79902,142.42476L881.66802,141.34826L882.75824,138.05724L880.21516,137.14377L879.73017,134.07221
{id:"MI",n:"Michigan",d:"M581.61931,82.059006L583.4483,80.001402L585.62022,79.201221L590.99286,75.314624L593.27908,74.743065L593.7
{id:"VT",n:"Vermont",d:"M844.48416,154.05791L844.80086,148.71228L841.91015,137.92811L841.26351,137.60479L838.35361,136.3115L839.16
{id:"ME",n:"Maine",d:"M922.83976,78.830719L924.77969,80.932305L927.04294,84.650496L927.04294,86.590422L924.94135,91.278575L923.001
{id:"RI",n:"Rhode Island",d:"M874.07001,178.89536L870.37422,163.93937L876.6435,162.09423L878.83463,164.02135L882.14112,168.342L884
{id:"NY",n:"New York",d:"M830.37944,188.7456L829.24781,187.77564L826.66123,187.61398L824.39799,185.67406L822.76738,179.54493L819.3
{id:"PA",n:"Pennsylvania",d:"M825.1237,224.69205L826.43212,224.42105L828.76165,223.1678L829.97353,220.68473L831.59014,218.42148L83
{id:"NJ",n:"New Jersey",d:"M829.67942,188.46016L827.35687,191.19443L827.35687,194.26599L825.41693,197.33754L825.25527,198.95416L82
{id:"DE",n:"Delaware",d:"M825.6261,228.2791L825.99441,226.13221L826.36948,224.44116L824.74648,224.83892L823.13102,225.30648L820.92
{id:"MD",n:"Maryland",d:"M839.79175,252.41476L833.7832,253.6186L828.6403,253.73606L826.79674,246.81373L824.87193,237.64441L822.299
{id:"VA",n:"Virginia",d:"M831.63885,266.06892L831.49494,264.12189L837.94837,261.57201L837.17796,264.78985L834.25801,268.56896L833.
{id:"WV",n:"West Virginia",d:"M761.18551,238.96731L762.29752,243.91184L763.38096,249.94317L765.51125,247.36283L767.77449,244.29127
{id:"OH",n:"Ohio",d:"M735.32497,193.32832L729.23143,197.38167L725.35158,199.64492L721.95671,203.36311L717.9152,207.24296L714.68199
{id:"IN",n:"Indiana",d:"M619.56954,299.97132L619.63482,297.11274L620.11981,292.58623L622.38305,289.67635L624.16133,285.79648L626.7
```

Figure 6. Geographical information of states in the US

2. Popularity map of a single brand

After drawing the underlay, states are filled with colors to represent the popularity. For the popularity map of a single brand, colors are of same hue but are different in saturation and brightness. Saturation and brightness are used to manifest the popularity level of the brand. Colors of the states depends on the results of sentiment analysis. CSV files which store the result of sentiment analysis are imported through d3.csv module. d3.interpolate module is then applied to interpolate the value of the color based on the ratios of positive tweets of each state. Generally, the darker the color, the higher the ratio and the more positive the attitude towards the brand in the state.

```
queue()
        .defer(d3.csv, "starbucks.csv")
        .await(ready);
function ready(error, starbucks){
    var sampleData ={}; /* Sample random data. */
    ["Alabama","Arkansas","Arizona","California","Colorado","Connecticut","Delaware","Florida","Georgia","Iowa","Idaho","
        Illinois","Indiana","Kansas","Kentucky","Louisiana","Maine","Maryland","Massachusetts","Michigan","Minnesota","
        Missouri","Mississippi","Montana","North Carolina","North Dakota","Nebraska","New Hampshire","New Jersey","New Mexico
        ","Nevada","New York","Ohio","Oklahoma","Oregon","Pennsylvania","Rhode Island","South Carolina","South Dakota","
        Tennessee","Texas","Utah","Virginia","Vermont","Washington","Wisconsin","West Virginia","Wyoming"]
        .forEach(function(d){
            for (var j = 0; j < 48; j++){
                if (starbucks[j].states == d){
                sampleData[starbucks[j].states]={negative:starbucks[j].negative, positive:starbucks[j].positive, ratio:
                    starbucks[j].ratio, color:d3.interpolate("#ffffcc", "#800026")(starbucks[j].ratio)
                    }
                }
            };
```

Figure 7. visualization of a single brand

3. Comparison map

Comparison map is drawn to show which brand is more welcomed in a specific state. Green and orange are used to represent Starbucks and Dunkin' Donuts respectively. Each state will be filled with color that represents the brand with higher ratio. In this case, two CSV files that store the results of sentiment analysis of different brands will be imported to make a comparison.

```
queue()
        .defer(d3.csv, "dunkin.csv")
        .defer(d3.csv, "stabucks.csv")
        .await(ready);
function ready(error, dunkin, stabucks){
    var sampleData ={}; /* Sample random data. */
    ["Alabama","Arkansas","Arizona","California","Colorado","Connecticut","Delaware","Florida","Georgia","Iowa","Idaho","
        Illinois","Indiana","Kansas","Kentucky","Louisiana","Maine","Maryland","Massachusetts","Michigan","Minnesota","
        Missouri","Mississippi","Montana","North Carolina","North Dakota","Nebraska","New Hampshire","New Jersey","New Mexico
        ","Nevada","New York","Ohio","Oklahoma","Oregon","Pennsylvania","Rhode Island","South Carolina","South Dakota","
        Tennessee","Texas","Utah","Virginia","Vermont","Washington","Wisconsin","West Virginia","Wyoming"]
        .forEach(function(d){
            for (var j = 0; j < 48; j++){
                if (dunkin[j].states == d && stabucks[j].states == d && stabucks[j].ratio<=dunkin[j].ratio){
                sampleData[dunkin[j].states]={winner:'Dunkin', dunkin_ratio:dunkin[j].ratio, starbucks_ratio:stabucks[j].
                    ratio, color:"#C57844"
                    }
                }else if (dunkin[j].states == d && stabucks[j].states == d && stabucks[j].ratio>dunkin[j].ratio){
                sampleData[stabucks[j].states]={winner:'Starbucks', dunkin_ratio:dunkin[j].ratio, starbucks_ratio:stabucks[j]
                    .ratio, color:"#2A7E42"
                    }
                }
```

Figure 8. visualization of comparison

4. Dynamic viewing

Colors manifest the popularity of brand in 48 states in the US. To help users better understand the results of sentiment analysis and view the original results, we enable dynamic viewing. Once users move mouse onto a specific state, they can check the detail result of sentiment analysis of the state.

For popularity map of a single brand, users can view the number of positive and negative tweets as well as the ratio of positive tweets. For the comparison map, they can examine the positive ratio of these two brands.

VI. SOFTWARE PACKAGE DESCRIPTION

1. Directory data collection contains source code for crawling tweets from internet, we have two filters: the location and the key words which refer to the brand names. We collect tweet data for 48 states in USA except Hawaii and Alaska. The file tweet+cleaning.ipynb is used to standardize the tweets.

2. Directory sentiment_analysis contains two .ipynb notebooks for sentiment analysis. In this part, we first use PySpark to train a sentiment classifier and then do sentiment prediction for tweets we collected.

3. Directory visualization contains one java script file and three html files as well as two .ipynb files. We draw three popularity maps to demonstrate the popularity of each brand in different states in United States and a final map which shows the more welcomed brand in each state. Also, we draw a pie chart to compare the overall sentiment analysis ratio of these two brands. Finally, we draw word clouds of these two brands.

VII. EXPERIMENT RESULTS

1. Accuracy

The accuracy of sentiment analysis is approximate 73.8%.

2. Sentimental ratio comparison

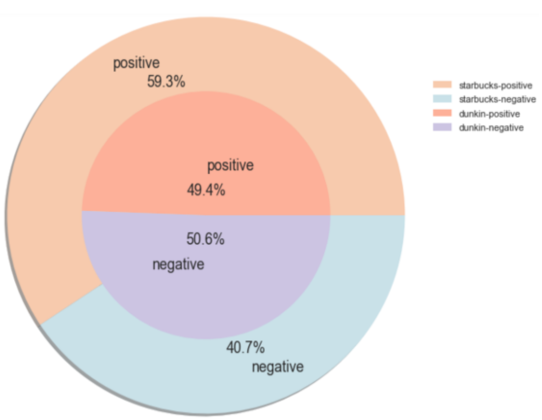

Figure 9. Ratio for Starbucks vs Dunkin' Donuts

***Dataset**:* The dataset contains the overall number of positive and negative tweets of these two coffee brands.

***Analysis**:*

- For Dunkin' Donuts, the negative ratio is over 50%, exceeding the ratio of positive tweets. From the perspective of tweet users, the overall expression of Dunkin' Donuts is negative.
- For Starbucks, the ratio of positive tweets is larger than that of negative tweets, which indicates the overall attitudes towards Starbucks is positive.
- In general, Starbucks outperforms Dunkin' Donuts, it has a lower negative ratio, compared with Dunkin' Donuts. Also, its ratio of positive tweets is higher than that of negative tweets.

3. Popularity map of a single brand

A. Starbucks

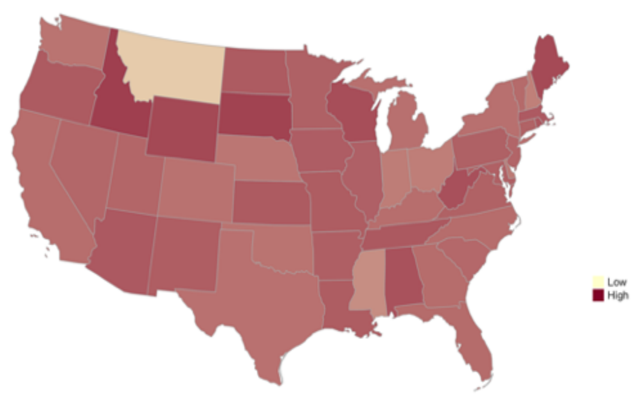

Figure 10. Popularity map of Starbucks

***Dataset**:* The dataset contains the ratio of positive tweets of Starbucks in 48 states.

***Analysis**:*

- The popularity of Starbucks varies between states.
- Starbucks is most welcomed in Idaho, Wyoming and South Dakota.
- Starbucks is least welcomed in Montana.
- The origin of Starbucks is in the state of Washington, states around its origin show a relatively positive attitude towards this brand, except Montana.
- Generally, Starbucks are more welcomed in western United State.

B. Dunkin' Donuts

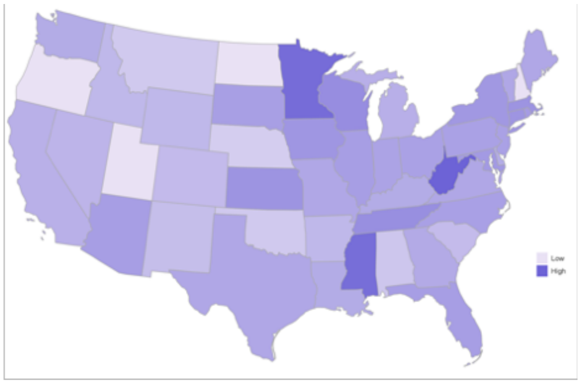

Figure 11. Popularity map of Dunkin' Donuts

Dataset: The dataset contains the ratio of positive tweets of Dunkin' Donuts in 48 states.

Analysis:

- The popularity of Dunkin' Donuts varies between states.
- Dunkin' Donuts is most welcomed in Minnesota, Mississippi and West Virginia.
- Dunkin' Donuts is least welcomed in North Dakota, Oregon and Utah.
- The origin of Dunkin' Donuts is in Massachusetts, but the states around its origin do not show significant high ratio of positive tweets.

General Analysis:

- West Virginia shows comparatively positive attitude towards both coffee brands.
- Montana shows comparatively negative attitude towards both coffee brands.
- Alabama and Georgia are two states that show opposite attitudes towards these two coffee brands.

4. Comparison map

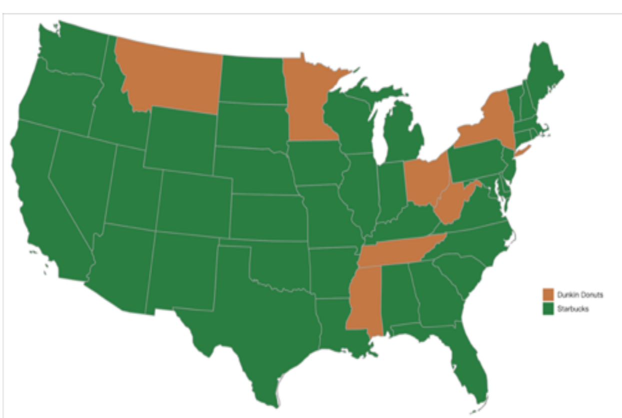

Figure 12. Comparison map

Dataset: The dataset contains the ratio of positive tweets of Starbucks and Dunkin' Donuts in 48 states.

Analysis:

- In most of the states in the US, Starbucks is dominant.
- Dunkin' Donuts is more welcomed in New York, Ohio, West Virginia, Tennessee, Mississippi, Minnesota and Montana.
- Dunkin' Donuts enjoys a higher reputation in the Eastern America.

5. Word Cloud Comparison

Word cloud is an intuitive word map built based on word counting. It demonstrates the most frequent word in the dataset. We first implement some data cleaning for the raw tweets data, then get the cleaned tweets and put them into word cloud generator to create word cloud logo for each brand.

A. Starbucks

Figure 13. Word Cloud of Starbucks

Dataset: The dataset contains all the tweets regarding Starbucks.

Analysis:

- The word cloud of Starbucks manifests that the most frequent word in the dataset is Starbucks, the name of the brand, which is understandable.
- Coffee is also a key word of this dataset. Since Starbucks is a famous coffee brand, it is also reasonable.
- Except the name and the key word 'coffee', dataset of Starbucks contains many words related to employment such as hiring, job, and supervisor. Since this project is target at analysis of users' attitude towards coffer brand, such information may interfere with our analysis.
- Words that can strongly manifest consumers' attitudes such as happy, love, best (positive), fuck and shit

(negative) are also in the dataset, which means the dataset collected is suitable for sentimental analysis.

B. Dunkin' Donuts

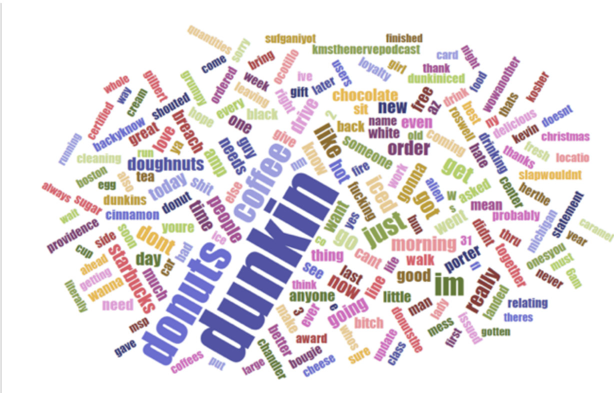

Figure 14. Word Cloud of Starbucks

Dataset: The dataset contains all the tweets regarding Dunkin' Donuts.

Analysis:

- Similarly, the most frequent words of Dunkin' Donuts dataset are the name of the brand and the key word 'Coffee'.
- The dataset of Dunkin' Donuts is neater, containing no tweets regarding hiring or advertisement.
- This dataset also contains many words that can be used for sentimental analysis such as like, better, love, good.

VIII. CONCLUSION

1. ***The overall accuracy of sentiment analysis is approximate 73.8%.***
2. ***Starbucks has more related tweets.***
   Considering the number of tweets collected, we conclude Starbucks has more related tweets on twitter. In most of the states, Starbucks is more frequently mentioned on twitter platform. However, as we have already discussed in the analysis of word cloud, many of these related tweets contain useless information such as hiring information and advertisement. For Dunkin' Donuts, although the dataset is smaller, it can strongly reflect users' attitudes towards the brand.
3. ***Starbucks wins in sentiment analysis.***
   As we have mentioned before, generally, Starbucks outperforms Dunkin' Donuts, it has a lower overall negative ratio, compared with Dunkin' Donuts. Also, its ratio of positive tweets is higher than that of negative tweets, which indicates that users hold a positive attitude towards this coffee brand.
4. ***Popularity of Starbucks varies between states.***
   Starbucks is most welcomed in Idaho, Wyoming and South Dakota, and is least welcomed in Montana. This brand is more welcomed near its origin: the state of Washington.
5. ***Popularity of Dunkin' Donuts varies between states.***
   Dunkin' Donuts is most welcomed in Minnesota, Mississippi and West Virginia, and is least welcomed in North Dakota, Oregon and Utah. Generally, it is more popular in the Eastern America.
6. ***In most of the States, Starbucks are more popular.***
   Except in New York, Ohio, West Virginia, Tennessee, Mississippi, Minnesota and Montana, people hold a more positive attitude towards Starbucks.

ACKNOWLEDGMENT

First, we would like to show our deepest gratitude to our professor, Dr. Ching-Yung Lin, a respectable, responsible and resourceful scholar. His keen and vigorous academic observation enlightens us not only in this project but also in our future study. We would also like to thank all the TAs who have helped us to overcome many difficulties. Last but not least, we would like to thank all our teammates for working hard together in this project, for the encouragement and support from each other.

APPENDIX

Fan Wu and Geqi Yan are mainly responsible for the sentiment prediction part. Xi Yang mainly worked on the data visualization part. All other works such as tweet collecting, further analysis, and report writing are done in group.

REFERENCES

[1] Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. *MIS quarterly*, 1165-1188.

[2] Zafarani, R., Abbasi, M. A., & Liu, H. (2014). *Social media mining: an introduction*. Cambridge University Press.

[3] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

[4] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, *2*(1), 1-8.

[5] Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012, July). A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations* (pp. 115-120). Association for Computational Linguistics.

[6] Saif, H., He, Y., & Alani, H. (2012, November). Semantic sentiment analysis of twitter. In *International semantic web conference* (pp. 508-524). Springer, Berlin, Heidelberg.

[7] Go, A., Huang, L., & Bhayani, R. (2009). Twitter sentiment analysis. Entropy, 17, 252.