Tweet Based Brand Analysis — Coca-Cola vs Pepsi

Haoran Guo, Yamei Zhu, Hao Fang
Electrical Engineering
Columbia University

hg2461@columbia.edu, yz3167@columbia.edu, hf2345@columbia.edu

Abstract—**Twitter provides a free and open platform for its users. Acting as a public social media, much information about commercial brands can be excavated from tweets and has relatively great value to be explored. In this project, we choose two beverage giants—Coca cola and Pepsi as study objects. Millions of relevant tweets are collected using python data crawler. We use spark machine learning techniques to process and analyze these tweets. In order to get an overall perspective, we implement multi-dimensional analysis including time-based comparison, sentiment-based analysis, keywords extraction, data visualization. We also make comprehensive comparison between Coca cola and Pepsi along with some brand commercial strategy analysis.**

Keywords- Twitter; Data Cleaning; Sentiment Analysis; Brand; TF-IDF; Naïve Bayes; Visualization

I. INTRODUCTION

We are now living in the information era. The Internet has created a virtual world that is closely joint with the real world like never before in history. Traditional industries are undertaking tremendous changes confronted with challenges. Every aspect of people's life has already been deeply reshaped by this unprecedented revolution. Among them, Social media does represent a characteristic industry that originated from the revolutionary Internet tide.[1]

Social media are computer-mediated technologies that facilitate the creation and sharing of information, ideas, career interests and other forms of expression via virtual communities and networks by definition. It has completely change the way people interact with information. For traditional media, there is usually a stable and one-way relation between users and media. No matter towards newspaper, television or broadcast, people as individuals are more tend to be in a way of receiving. Polling may be the only few ways to get personal views but gains features of time-consuming and delay. However, social media offers us a specialized way to learn about public views. It breaks the transmit-centralization pattern and obscures the bound of information senders between receivers to form a flat structure way of message transmission. Each entity becomes the subject of the network. They are both consumers and producers towards information.

By 2017, more than 2.4 billion people uses social media around the world. In U.S, over 81 percent of people have at least one social profile according to survey. Every minute people send almost 30 million messages on Facebook and 350,000 tweets. The scale of the users and information make social media like Facebook, Twitter, Wechat not only a communication platform but a huge social ecosystem. On account of the above factors, more and more enterprises start to realize the huge potential of commerce behind social media. They set up social accounts as online platform connected with customers, launch various events to advertise their brand and even target different consumption group by probing personal behaviors. Therefore, doing analysis related to commercial brands based on social media data is of great significance for both consumers and companies.

Regarding to our project, we select Twitter as the social media platform for data mining since Twitter is open and free for users to express opinions or ideas, share emotional feelings, follow interesting people and many other interactive behaviors. The short size of each tweet also enables it to be faster and more responsive in reflecting fluctuation. As for commercial brands, we choose two rival giants in beverage industry: Coca cola & Pepsi. We try to analyze their brand image strategy themselves and people's opinion or attitudes towards similar competitive brands using big data tools and general computer technologies from spark machine learning, python natural packages and data visualization. Multi-dimension analysis from quantitative statistics, sentiment analysis and brand comparison based on tweets for both are implemented and we get relevant conclusions in various levels.[8]

II. RELATED WORKS

There have been various kinds of sentiment analytics based on twitter data previously. A considerable part of them stops with principle parts on retrieving tweets and doing sentiment analysis by directly importing tweets into models. Analysis closely related with reality life or industry mostly focuses on prediction and classification. Predictions on presidential election based on user overall mood tendency or
Stock market price trend using users related comments take up a large part of the relevant research. Analysis or classification focuses largely focuses on single object. However, there is not enough works towards brand analysis based on twitter data. For the current consumption market, there is a saying that people don't buy products, they buy

brand. There is necessity to do relevant analysis on brand in a business domain.

Our works focuses on doing analysis on commercial brands which involves business fields content. We choose two beverage brands Coca cola and Pepsi in this project but the analysis flow can be applied to other brands such as Nike and Adidas. The comparison between two competitive counterparts is another new attempt. We also involve different methods and perspectives in analysis like quantitative trends, word count and visualization besides sentimental analysis to get a multi-dimension view.[9]

III. System Overview

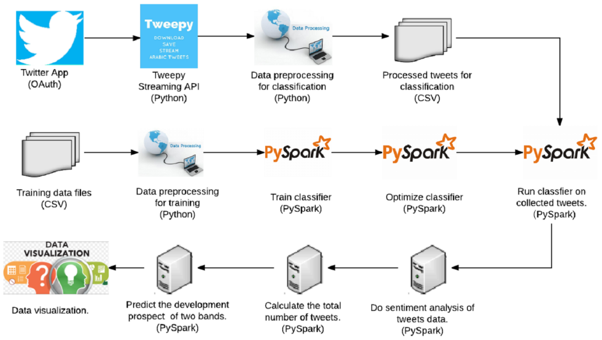

Tweepy
DOWNLOAD
SAVE
STREAM
ARABIC TWEETS
Twitter App (OAuth)
Tweepy Streaming API (Python)
Data Processing
Data preprocessing for classification (Python)
Processed tweets for classification (CSV)
Training data files (CSV)
Data Processing
Data preprocessing for training (Python)
PySpark
Train classifier (PySpark)
PySpark
Optimize classifier (PySpark)
PySpark
Run classfier on collected tweets. (PySpark)
DATA VISUALIZATION
Data visualization.
Predict the development prospect of two bands. (PySpark)
Calculate the total number of tweets. (PySpark)
Do sentiment analysis of tweets data. (PySpark)

Figure 1: System Overview

The first stage is crawling data from twitter. The second stage is carrying out data cleaning and training sentiment classifier. Then we do a sentiment prediction of all the tweets we have got. At last, we use seaborn to do data visualization.

IV. Algorithm

Detailed implementation and tools used for different tasks are documented in this section.

A. Dataset

We use the Twitter Sentiment Analysis Dataset for model training in Spark. The corpus is based on two official open source data: *1. Sentiment Classification by University of Michigan. 2. Niek Sanders Twitter Sentiment Corpus.* It contains 1,578,627 classified tweets of which each one is labelled 1 as positive 0 as negative.

Our own twitter dataset has a requirement to track tweets in a long-term time span for historical data since analysis of variation tendency towards two brands in a multi-dimension perspective is included as part of the data analysis. The Official Twitter Streaming API has a limitation of data acquisition which restricts us to get the oldest data within 7 days ago. Therefore, we use Twitter API as a supplementary tool for data study and implement a python data crawler to form the dataset based on parsing html using packages as pyquery, urllib, urllib2 and re.

The final dataset contains 1,237,349 tweet records for brand keywords 'Coca-cola' and 'Pepsi'. The time period ranges from January 1st 2017 to November 30th 2017. We extract the following features of each tweet:

1. Tweet ID: Unique number for each tweet
2. Username: Name for the tweet sender.
3. Text: Tweet main content
4. Date: Time for tweets with accuracy to hour
5. Hashtag: Tag labelled to tweets.

B. Data Cleaning

1. Escaping HTML Characters

Tweets crawled from internet might contain html entities such as *&lt; &gt; &amp;*, which might cause error in future analysis. Removal of these entities is essential for sentimental analyzing. In this project, html parser of python is used.

2. Decoding Data

Raw tweet data often contain complex symbols, which would influence the analyze process. Therefore, we are supposed to transfer them into simple and understandable characters. Moreover, for better analysis, it is necessary to keep the complete data in standard encoding format. Generally, UTF-8 encoding is a good choice, for it is widely used and easy accepted.

3. Apostrophe Lookup

In tweets, apostrophes sometimes lead to disambiguation and mistakes. For example, when specifying *'s*, it is hard to tell whether it means *has* or *is*. Therefore, to avoid such conditions, words with apostrophe should be converted into standard format.

4. Split Attached Words

Tweet data are often lack of language standardizing. One situation is that we may create some attached words. These words are usually a combination of two or more different words. Attached words are easily understood by human, but hard to tell by computers. We should split them into the original formats.

5. Slangs lookup

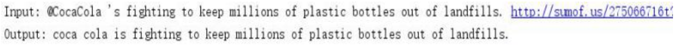
For tweet data, it is common to use slang words. Therefore, to minimize the influence of slang words, we should transform them into standard words.

6. Standardizing Words

Words like *love* might be wrote as *loooveee* or other formats. Therefore, before doing further analysis, it is crucial to transform them back to the original style. [2]

7. Removal of URLs

URLs is an important part in tweets. But for us, urls would cause unexpected misleading. In many tweets, urls is used for advertising or just be forwarded by users. Remove all urls in twitter data would remarkably increase the accuracy of analyzing.



```
Input: @CocaCola 's fighting to keep millions of plastic bottles out of landfills. http://sumof.us/275066716t
Output: coca cola is fighting to keep millions of plastic bottles out of landfills.
```

Figure 2:Example of data cleaning (part1-part7)

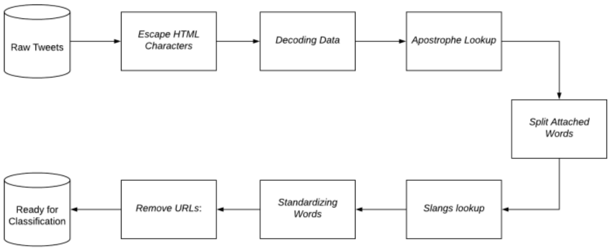
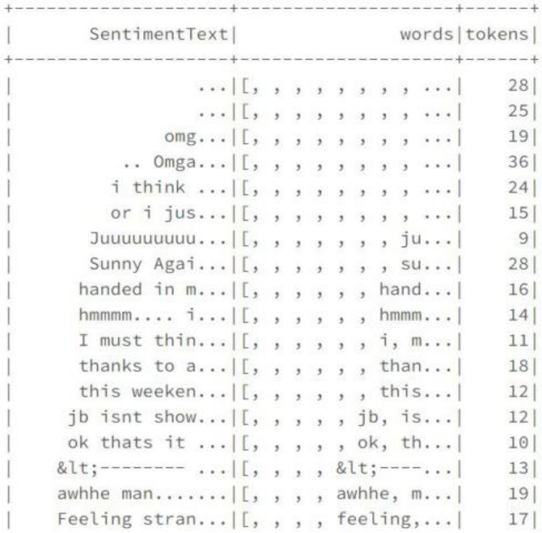


Figure 3: Data cleaning overview (part1-part7)

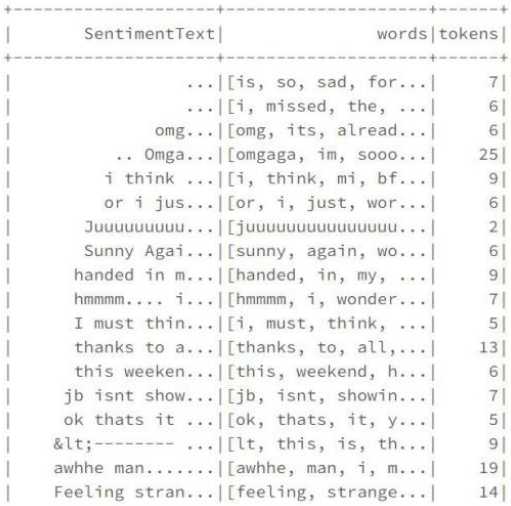
8. Regex Tokenizer

Original tokenizer could split sentences into words. For tweets, it is better to use regex tokenizer, which use regex to judge the split positions. Because tweet data are often less standard than other articles, including extra space and symbols which are hard to split.



```
+--------------------+--------------------+------+
|       SentimentText|               words|tokens|
+--------------------+--------------------+------+
|                 ...|[, , , , , , , , ...|    28|
|                 ...|[, , , , , , , , ...|    25|
|              omg...|[, , , , , , , , ...|    19|
|          .. Omga...|[, , , , , , , , ...|    36|
|         i think ...|[, , , , , , , , ...|    24|
|         or i jus...|[, , , , , , , , ...|    15|
|       Juuuuuuuuu...|[, , , , , , , ju...|     9|
|       Sunny Agai...|[, , , , , , , su...|    28|
|      handed in m...|[, , , , , , hand...|    16|
|      hmmmm.... i...|[, , , , , , hmmm...|    14|
|      I must thin...|[, , , , , , i, m...|    11|
|      thanks to a...|[, , , , , , than...|    18|
|      this weeken...|[, , , , , , this...|    12|
|     jb isnt show...|[, , , , , jb, is...|    12|
|     ok thats it ...|[, , , , , ok, th...|    10|
|    &lt;-------- ...|[, , , , &lt;----...|    13|
|    awhhe man.......|[, , , , awhhe, m...|    19|
|    Feeling stran...|[, , , , feeling,...|    17|
```

Figure 4: Tokenizer



```
+--------------------+--------------------+------+
|       SentimentText|               words|tokens|
+--------------------+--------------------+------+
|                 ...|[is, so, sad, for...|     7|
|                 ...|[i, missed, the, ...|     6|
|              omg...|[omg, its, alread...|     6|
|          .. Omga...|[omgaga, im, sooo...|    25|
|         i think ...|[i, think, mi, bf...|     9|
|         or i jus...|[or, i, just, wor...|     6|
|       Juuuuuuuuu...|[juuuuuuuuuuuuuuu...|     2|
|       Sunny Agai...|[sunny, again, wo...|     6|
|      handed in m...|[handed, in, my, ...|     9|
|      hmmmm.... i...|[hmmmm, i, wonder...|     7|
|      I must thin...|[i, must, think, ...|     5|
|      thanks to a...|[thanks, to, all,...|    13|
|      this weeken...|[this, weekend, h...|     6|
|     jb isnt show...|[jb, isnt, showin...|     7|
|     ok thats it ...|[ok, thats, it, y...|     5|
|    &lt;-------- ...|[lt, this, is, th...|     9|
|    awhhe man.......|[awhhe, man, i, m...|    19|
|    Feeling stran...|[feeling, strange...|    14|
```

Figure 5: Regex Tokenizer

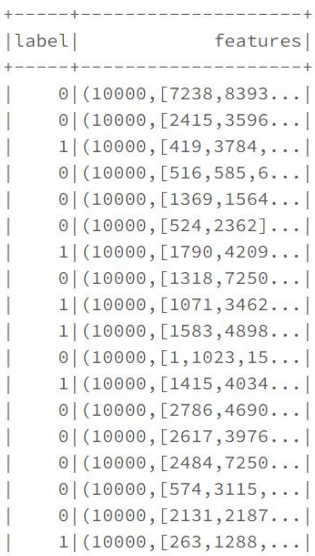
9. Remove Stop Words

Many stop words in tweet data must be removed. For sentimental analysis, high frequency words without sentimental meaning might cause great error in sentimental prediction. In this project, we remove stop words twice. The first removal for high-frequency words in English language. The second removal for high-frequency words in certain tweets about a brand.

C. Sentimental Classification

1. Hashing TF & IDF

Hashing TF means mapping a sequence of terms to their term frequencies using the hashing trick. IDF is inverse document frequency. The main idea of IDF is that if any word appears less frequently than others, the more important it actually is in classifying the texts. Hashing TF and IDF is widely used in natural language processing. In this project, such methods are used for converting words into vectors, which would be used in sentimental classifying and prediction.



```
+-----+--------------------+
|label|            features|
+-----+--------------------+
|    0|(10000,[7238,8393...|
|    0|(10000,[2415,3596...|
|    1|(10000,[419,3784,...|
|    0|(10000,[516,585,6...|
|    0|(10000,[1369,1564...|
|    0|(10000,[524,2362]...|
|    1|(10000,[1790,4209...|
|    0|(10000,[1318,7250...|
|    1|(10000,[1071,3462...|
|    1|(10000,[1583,4898...|
|    0|(10000,[1,1023,15...|
|    1|(10000,[1415,4034...|
|    0|(10000,[2786,4690...|
|    0|(10000,[2617,3976...|
|    0|(10000,[2484,7250...|
|    0|(10000,[574,3115,...|
|    0|(10000,[2131,2187...|
|    1|(10000,[263,1288,...|
```

Figure 6: Hashing TF-IDF

2. Naive Bayes Classifier Training

For sentimental classifying, we train a naive bayes classifier. Training data are tweets labeled as positive or negative. We implement all above algorithms on the training data, converting text to vectors. With such vectors and labels, a naive bayes classifier could be trained. In this project, we finally trained an classifier with an accuracy of about 73.83%.

3. Sentimental Prediction

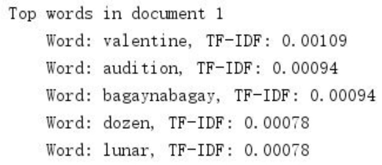
Tweets scrawled from internet are cleaned and vectorized by the above methods. With vectors for each tweet, we are able to classify them as positive or negative with our byes classifier. Up to now, all tweets' sentiment have been predicted and could be used for statistics and analysis.[5]

D. Finding Important Words

1. Using TF-IDF

TF-IDF means *Term Frequency, Inverse Document Frequency*. It is used to judge the importance of words in a document based on how frequently they appear in the whole document set. It is said that if a word frequently appears in a document but is not so frequent in the whole data set, then it must be an important word for that certain document. But if a word, such as *go, hello, pepsi,* appears frequently in all documents, then it should not be counted as an key word.[3]

In this project, TF-IDF is used as a tool for extracting important words in sentiment variation part.



```
Top words in document 1
    Word: valentine, TF-IDF: 0.00109
    Word: audition, TF-IDF: 0.00094
    Word: bagaynabagay, TF-IDF: 0.00094
    Word: dozen, TF-IDF: 0.00078
    Word: lunar, TF-IDF: 0.00078
```

Figure 6: Finding important words with TF-IDF

2. *Word Cloud*

Word cloud is based on word counting. It is a simple and effective tool when finding key words for a data set. By building a word cloud, we could build up an intuitive word map for a brand.

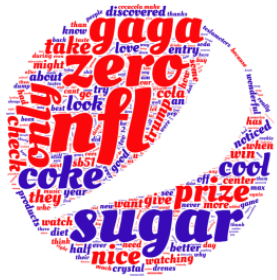


Figure 7: Word cloud for Pepsi in Feb.2017

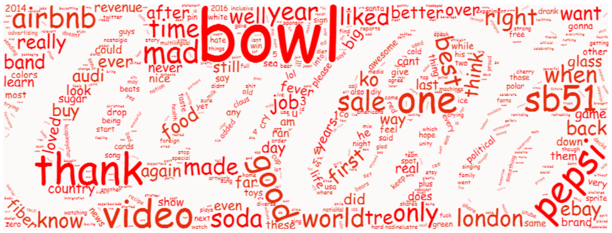


Figure 8: Word cloud for Coca-cola in Feb.2017

V. Software Package Description

1. File tweet_crawl contains source code for crawling tweets from internet.
2. Directory data_cleaning_tf_idf contains two python files. tweet_cleaning.py is used for cleaning raw twitter data and make them standarized and easy for future use. tfidf_count.py is used for calculating tf-idf in different files. It would import tweet_cleaning.py for tweet data.
3. Directory sentiment_analysis contains a ipython notebook. This part is implemented in databricks, connected with AWS. Using pyspark to train an sentiment classifier and do dentiment prediction for tweets we scrawled.
4. Directory visualization_code cotains several ipython notebooks. These python codes do visualization on tweets.

6. EXPERIMENT RESULTS

1. Heatmap

A heat map (or heatmap) is a graphical representation of data where the individual values contained in a matrix are represented as colors. Seaborn is a Python visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics.We utilized Seaborn to present a heatmap of the number of tweets in 11 months of two brands - pepsi and coca cola. The coloring on the map is based on the number of the tweets tweeted by the users. The lighter color in the heatmap represents the larger number of tweets in a time slot, which implies more active users post tweets in that time range. The darker color in the heatmap otherwise reflects users are less active in that time slot in post tweets.

The overall process flow involved is detailed in the following steps[4]:

- Count On The Overall Tweets: After extracting all the tweets, we use Spark to count the overall number of tweets in each month for pepsi and coca cola.
- Categorization of tweets: We Categorize the tweets by timestamp in a day and count the average tweets in the same timestamp in 11 months.
- Integrate with Seaborn: Transfer the imported data into pivot_table and generate the heatmap.
- Detect Word Frequency: When abnormal data are found, we detect the keyword frequency in that time slot to figure out what's happened.

Dataset: the dataset contains the tweets of coca-cola and pepsi in 24 timestamps in 11 months.

Analysis: The heatmap of coca cola and pepsi are showed as below:

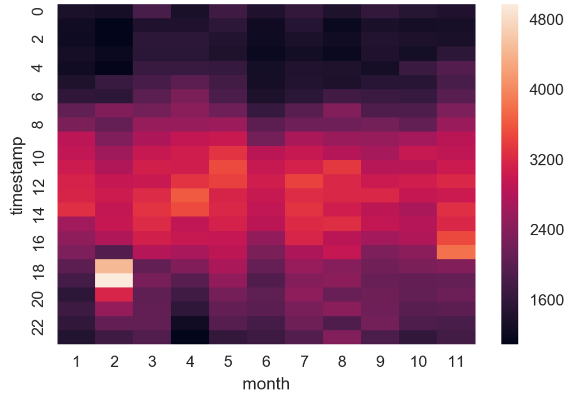


Figure 9: Heatmap Plot for coca cola

The heatmap of coca cola reveals two interesting facts:

1. The time slot between 9:00 o'clock and 15:00 o'clock is the time users are active in tweeting content about coca cola.
2. Abnormally high data takes place in the time slot between 18:00 o'clock and 20:00 o'clock in Feburary.

After we extracted the word frequency of tweets in February, we found the keyword Super Bowl Ads occurs most frequently. We can figure out the reason of the abnormity in the Feburary arises from a commercial show - Super Bowl Commercial which is sponsored by coca cola.

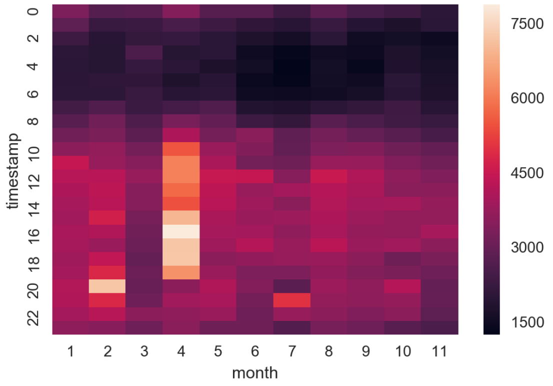


Figure 10: Heatmap Plot for pepsi

The heatmap of pepsi reveals three interesting facts:

- The time slot after 11:00 o'clock is the time users are active in tweeting content about pepsi.
- Abnormally high data takes place in the time slot around 20:00 o'clock in Feburary and between 10:00 o' clock and 19:00 o'clock in April.
- Compared with the heatmap of coca cola, we can figure out that the active time range of pepsi is a little larger than that of coca cola. Through extracting the data and analysing the word frequency, we can find that the reason of the abnormity in the April arises because of Kendall Jenner's commercial for Pepsi - "Live for Now" . The abnormal data in Feburary is caused by Super Bowl Half Time Show.

2. Sentiment Ratio Comparison

The ratio comparison not only makes our sentiment analysis more visualized but also gives the company runners a more intuitive display of the statistics of twitter uses' attitudes towards them.

Dataset: The dataset contains the overall number of positive and negative tweets of coca cola and pepsi.

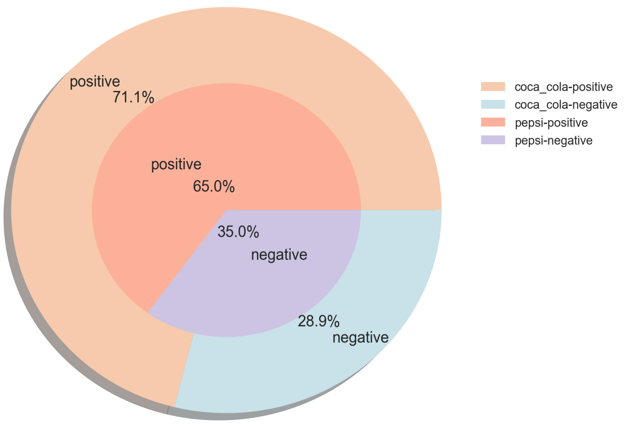


Figure 11: Ratio for coca cola vs pepsi.

Analysis: According to the chart, we can easily found the negative rate of pepsi is larger than that of coca cola. From the perspective of the branding officer of pepsi, it's time to invest more in improving public opinions in order to make itself not at a disadvantage in the competition with coca cola.

3. Month Trend on Tweet Amount

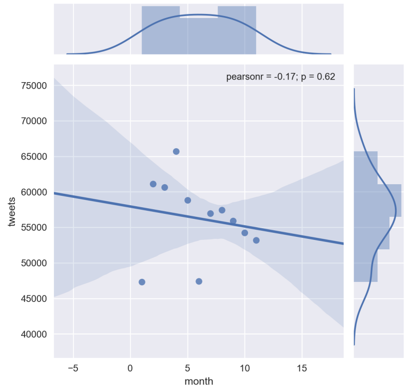


Figure 12: Joinplot for coca cola.

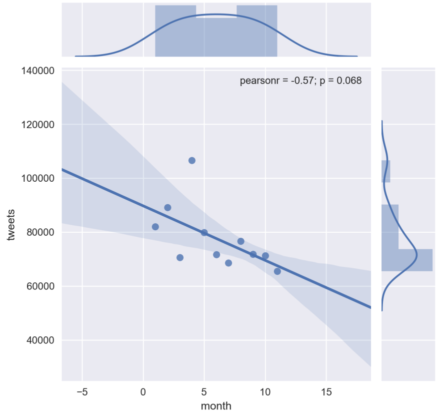


Figure 13: Joinplot for pepsi.

Dataset: the dataset contains the tweets of coca cola and pepsi for 11 months.

Analysis: The trends of both two brands is that with the time passing by, the tweets about them is decreasing.

4. Sentiment Variation Analysis

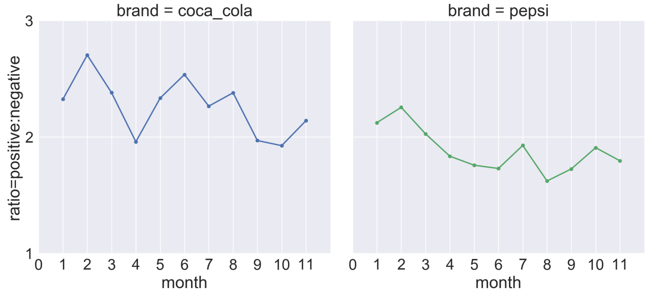


Figure 14: line chart for coca cola and pepsi.

Dataset: The dataset contains the number of positive and negative tweets of coca cola and pepsi.

Analysis:

1. In the line chart of coca cola, we find there's a peek in February and a sharp decrease from February to April. As for the peek, after detecting the word frequency, we can infer the peek arises from super bowl advertisement. As for the valley, we have found three reasons. Firstly, we've found the word frequency of two keywords: Jobs, Cut is incredibly high. A large-scale job cutting took place and affected people's attitudes towards coca cola. Secondly, an iditarod dog doping involved in abusing dogs sponsored by coca cola made the public angry, leading to the ratio of positive and negative getting lower. At last, a tweet of committing to keeping your plastic out of our oceans caused many forwards. Many people contribute to the events and keep a negative attitude towards plastic producer including coca cola. [6]
2. Speaking of the line chart of pepsi, there are two turning points we need to put attention on. The first one is in February. That is the time of Super Bowl Halftime Show. In July, there's a talk show called conanco which holds an activity online to give out free pepsi. So far we have just found this two information and more information needs to be explored in the future.

7. Conclusion

Pepsi got more related tweets.

- Analysis: pepsi got more related tweets according to the overall number of tweets posted by users but the conclusion is not fair because many people call coca cola just coca but we just use the keyword coca cola to search for related tweets. It’s the same with pepsi.

Coca cola beats Pepsi in sentimental analysis.

- Analysis: As we can see, people holds more positive public opinons towards coca cola.

Most of the commercial show containing no controversial parts will promote positive public attitudes towards the brand.

- Analysis: We can see obvious improvement of the positive : negative ratio of both brands in the super bowl advertisement.

Some controversial marketing campaign advertisement is counterproductive in the aspect of improving the public opinions.

- Analysis: We can see a commercial show called Kendall Jenner’s Pepsi Commercial – “Live for Now” really triggered a heated discussion among people. However, when looking at sentiment analysis, we can see people hold more negative attitudes towards pepsi in April than last month. It’s apparent that the commercial show leads to a negative effect in promoting the brand images.

Acknowledgment

We would like to express our tremendous gratitude to Professor Ching-Yung Lin. He provided us with lots of creative ideas and great inspirations in our project. Also, we appreciate the help from the TAs, especially mao sun and eric lee. Their patience and devotions help us overcome dozens of difficulties and finish this project.

Appendix

Haoran Guo and Hao Fang mainly worked on the sentimental prediction part. Yamei Zhu mainly contributed to the data visualization part. All other works such as tweet crawling, data cleaning and further analysis are done in group.

References

[1] Chen, Hsinchun, Roger HL Chiang, and Veda C. Storey. "Business intelligence and analytics: From big data to big impact." MIS quarterly 36.4 (2012).

[2] Rahm, Erhard, and Hong Hai Do. "Data cleaning: Problems and current approaches." IEEE Data Eng. Bull. 23.4 (2000): 3-13.

[3] Ramos, Juan. "Using tf-idf to determine word relevance in document queries." Proceedings of the first instructional conference on machine learning. Vol. 242. 2003.

[4] Thorvaldsdóttir, Helga, James T. Robinson, and Jill P. Mesirov. "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration." Briefings in bioinformatics 14.2 (2013): 178-192.R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[5] Pak, Alexander, and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining." LREc. Vol. 10. No. 2010. 2010.M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[6] Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore. "Twitter sentiment analysis: The good the bad and the omg!." Icwsm 11.538-541 (2011): 164.

[7] K. Kaviya, C. Roshini, V. Vaidhehi, J. Dhalia Sweetlin, "Sentiment analysis for restaurant rating", Smart Technologies and Management for Computing Communication Controls Energy and Materials (ICSTM) 2017 IEEE International Conference on, pp. 140-145, 2017.

[8] A. Deshwal and S. K. Sharma, "Twitter sentiment analysis using various classification algorithms," 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, 2016, pp. 251-257.