

 samlimish / Project5

★ 0 stars 🍴 0 forks

 Star Unwatch ▾[Code](#)[Issues](#)[Pull requests](#)[Actions](#)[Projects](#)[Wiki](#)[Security](#)[Insights](#)[Settings](#) main ▾

...



samlimish ...

3 minutes ago

[View code](#) README.md

Capstone Project: Compare Consumer Sentiments of Apple, Google, and Android from Past to Present

Instructor: James Irving

Date: 7/23/21

Contents



- [Business Problem](#)
- [Data Collection](#)
- [Data Cleaning](#)
- [Data Modeling](#)
- [Recommendation](#)
- [Conclusion](#)

Business Problem

For better or worse, people's perception of tech giants have changed over time. A company that consults these large companies' PR teams have hired me to find how the consumers' sentiments have changed. To gather the necessary information, I am going to go to Twitter, and I will compare the public's emotion towards these companies using vader sentiment analysis.

Data Collection

The old Twitter data were provided, but for the new Twitter data, I used tweepy and Twitter's developer API to collect 1,500 recent tweets for each of the three companies: Apple, Google, and Android.

 apple_gather
 rest_gather

I also decided that it would be best to exclude tweets that contained more than one company names, and for Apple, since there were multiple related words/phrases, more filtering had to be done.

Data Cleaning

First, I had to create a text cleaner to help get rid of non asc-ii characters and other pieces of strings that I did not need.

 cleaner


I noticed that many of the tweets did not contain any product or company names. Using regular expression, I looked for the company names in the text and gave values to the product or company column.

 companyname


For rows that already contained a company name or product name, I decided to reduce the number of options into three: Apple, Google, and Android.

VADER Sentiment Analysis


From the VADER sentiment analysis, I collected the 'compound' values as they represented the most accurate sentiment depiction. If the values were bigger than or equal to .05, it was considered positive; if less than or equal to -.05, negative; and the rest were considered neutral.

 vader The following shows the distribution of the sentiments.

Old Tweet

 old_senti

New Tweet

 new_senti

From these distributions I noticed that 0 was the most common sentiment value and that the number of negative tweets were much lower those of positive and neutral tweets.

Old Tweet

old_count

New Tweet

new_count

Tokenizing and Stop Words

For NLP analysis, it is important to have a token list. Token list is splitting a text file into words or characters.

token

Using this, I created a frequency distribution plot of the most commonly used words/characters

token_old

Then I created a list of common stopwords, added punctuations, as well as the company names and other common Twitter terms.

stopwords

First, I created a horizontal frequency distribution plot that used a token list from both the old and new tweets, then I created the same frequency distribution plot that used the stop words.

total_freqtotal_freq_stop

Data Modeling

Model Recommendation

For both of the tweets, the Random Forest models after the gridsearch performed the best. The random forest models before the pipeline and the models after performing the grid search performed similarly; however, because the low recall score for the negative tweets were improved to that of the original random forest models while maintaining the relatively high recall scores for the neutral and positive tweets, I recommend the **Random Forest models using gridsearch**. As mentioned before, the number of tweets with negative sentiments were only a third of the size of either the positive or neutral neutral tweets which led to the poor recall rate of negative tweets for all models.

Conclusion

Future Work

There are many things that I would like to incorporate into this project that I did not have the necessary skills or time to perform, but given the chance I would like to do these for the future:

1. Collect similar amounts of tweets for all three sentiments
 - This would help improve the recall scores for the models and maybe give an insight as to why the naive bayesian models did not perform as well before gridsearch.
2. Create a time line of sentiments and events.
 - By collecting tweets from the past throughout to the present, I would be able to show how the sentiment trends gradually changes.
 - Incorporating events from the past could also help understand the sentiment trends and highlight specific actions that the company had taken that negatively or positively impacted the public's sentiment.

Final thoughts

This project was an adventure from the start as I had to learn how to collect my own data. Another aspect that I had not experienced before was the lengthy text cleaning that came with using tweets. And while the old tweet data were provided, I had to use the vader sentiment analysis as I had noticed much of the tweets did not correctly represent the tweet's sentiment. The aspect that most intrigued me about the Twitter sentiment towards these large companies is the obviously smaller number of negative emotion towards the companies and the overwhelming number of neutral sentiments towards them as well. I had initially expected to see a large increase in negative sentiment towards companies such as Apple as they have become less consumer friendly in their product designs. And as mentioned above in the future works section, I believe that creating a time line of the sentiment changes and events could help paint a clearer picture of how much these companies have changed in their policies and in the eyes of the consumers.

Releases

No releases published

[Create a new release](#)

Packages

No packages published

[Publish your first package](#)

Languages

● Jupyter Notebook 99.9% ● Python 0.1%