Twitter Sentiment Analysis

Paris Vu
Springboard
February 2023

Introduction

Stock has always been one of the main assets among institutions and it has grown even more popular among Millenials and Gen Z in recent times due to the redefinition of the money term and its price fluctuation.

Moreover, scientists and analysts have gradually recognized social media's predictive power for the financial market, especially Twitter where key opinion leaders feel comfortable expressing their opinions.

Correlation between tweets and stock prices

There is evidence to suggest that there can be a correlation between tweets and stock prices, particularly for publicly traded companies. The idea behind this correlation is that tweets can influence public perception, which in turn can impact stock prices.

Problem Identification

What opportunities exist for retail investors to develop a machine learning model that could measure sentiments of a stock using recently published Twitter's public opinion?

Historical Data: December 2014

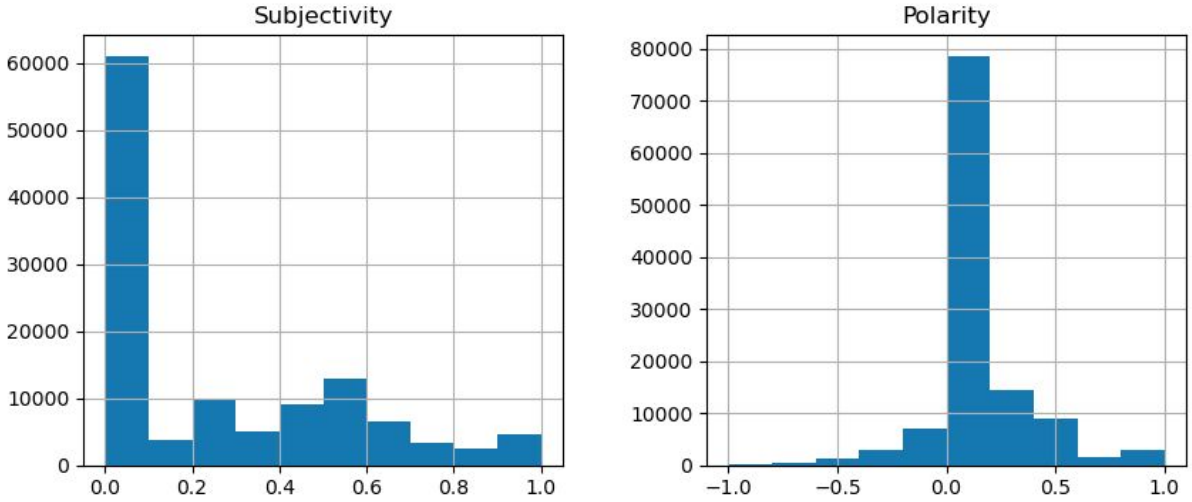
Before diving into live tweets scraping, I first explored two pre-scraped dataset (December 2014). Both dataset combined together produced a word cloud that suggests tweets are about 3 months ahead of real-life events.

One common 'big' word that appeared in both of these dataset is **SXSW**. After doing a simple google search, SXSW is a music festival that is usually held in mid-March. It is surprising to see that both the first dataset (dates in December 2014) and second dataset (unknown dates) mentioned this festival. This implies that tweets might be 3 months ahead of real life events!

Live tweets from December 2022

Scraped live Tweets with specific hashtags (#appl, #apple, #ipad, & #iphone) using the SNScrape method

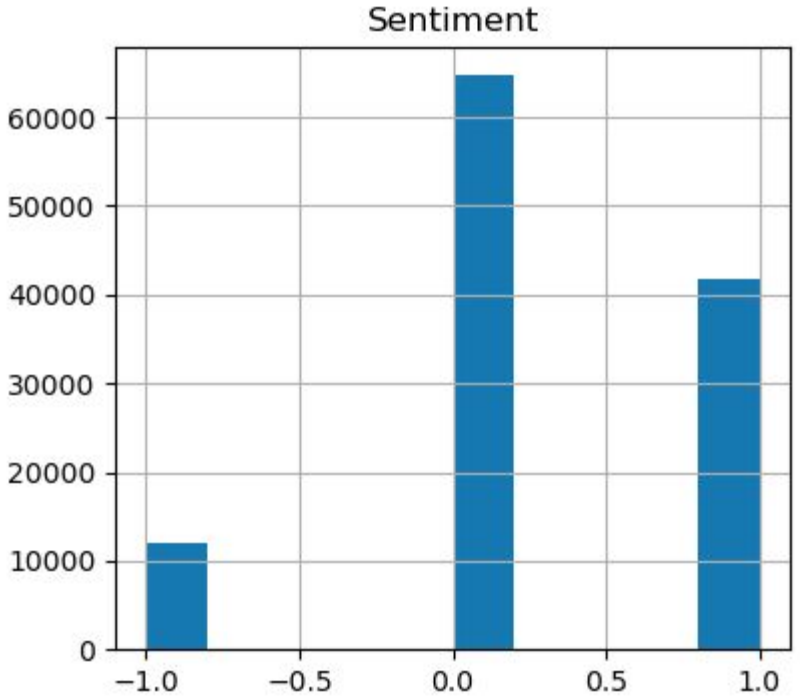
Word cloud did not return similar result to December 2014 word cloud



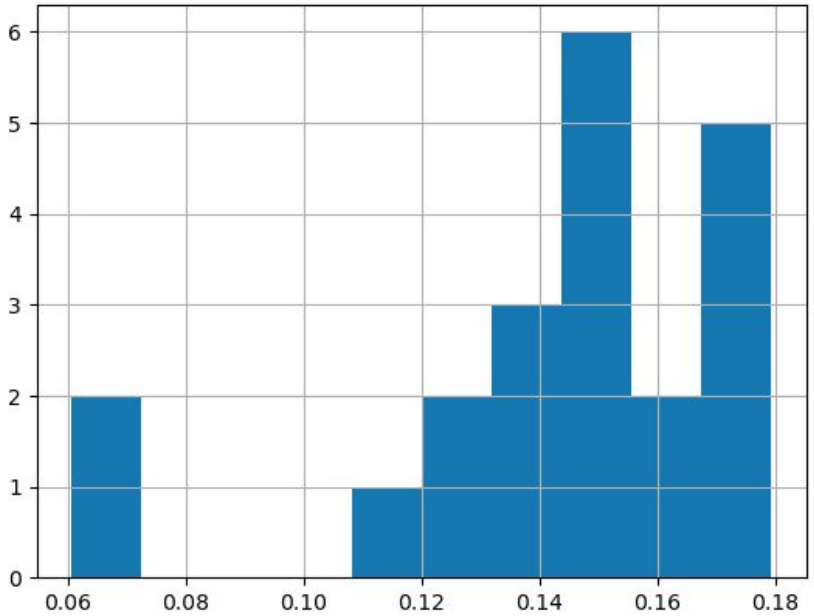
Subjectivity detection and polarity detection are subtasks under sentiment analysis. Subjectivity detection aims to remove 'factual' or 'neutral' content, i.e., objective text that does not contain any opinion. **Polarity** detection aims to differentiate the opinion into 'positive' and 'negative'

December 2022 was a positive month in terms of Twitter public opinion!

After having their polarity scores, I simplified it into numbers, called Sentiment numbers, representing the following: -1 = Negative (if polarity <0); 0 = Neutral (if polarity = 0); 1 = Positive (if polarity > 0)



After appointing the daily polarity scores to its corresponding label, we now have 11 positive days, and 10 negative days.



Modeling Overview

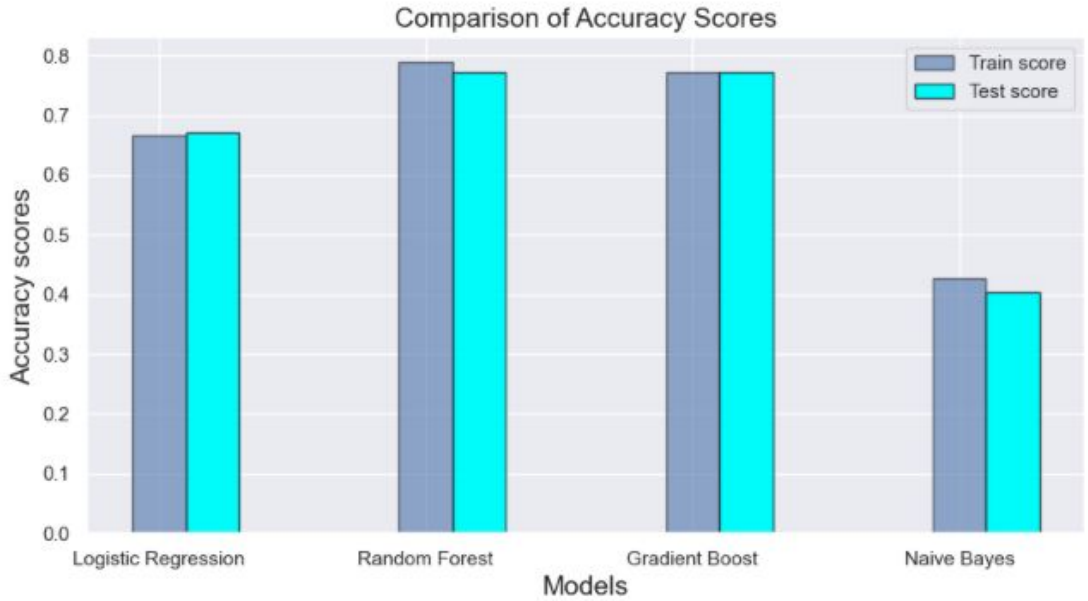
Multi-class Classification in Supervised Learning

Model Evaluation: Precision

In general, if the goal of the sentiment analysis model is to identify all instances of negative sentiment in a dataset, regardless of false positive predictions, then recall would be the priority. However, if the goal is to ensure a high level of accuracy in sentiment predictions, such that false positive predictions are minimized, then precision would be the priority.In our capstone, we value high level of accuracy, hence precision is chosen.

Precision Score Comparison

Accuracy Score Comparison



Bias/Variance Tradeoff

Due to our high precision scores in Random Forest and Gradient Boost models, our models are very prone to overfitting. Both models are nonlinear and systematically have low bias/high variance.

In order words, if we pass one training dataset into different models, they are more likely to yield different outputs. Results are not consistently and difficult to generalize.

Conclusion

In order to predict the sentiment of each tweet, we have vectorized the tweets and applied classification machine learning models. Here we have used the following classification models:

- Logistic Regression
- Random Forest
- Gradient Boost
- Naive Bayes

The model evaluation is based on splitting the dataset into train and validation set.

Cross-Validation procedure is used under the k-fold CV approach

We have evaluated each model in terms of model accuracy score, and 'precision' score for both the training and test data, and plotted them. The two best performing models are the Random forest and the Gradient boost. Both are the ensemble model, based on decision trees.

Conclusion (continued)

Next, we have carried out the Randomized Search CV for the hyperparameter tuning for both the models separately. This step was the most time consuming one in terms of computation. (The RF model took much longer time). I originally attempted the exhaustive Grid Search CV, but with the result of the optimized hyperparameters, we have again fitted the two models, and got the predictions separately. The model precision did not improve much with the optimized parameters, increasing from 0.800 to 0.801 and accuracy staying the same as 0.785, for RF and GB, respectively.

Future Research

There are possible alternatives to this project that could lead to potential improvements:

- Include engineered features from the stock data obtained from the Exploratory Data Analysis notebook as the values for X in our models
- Include text mining features from the tweets data obtained from the Preprocessing and Training Data notebook as the values for X in our models
- Use a whole year worth of tweets instead of just the month of December 2022
- Use different hashtags