NLP Stock Predictor

Team: Jesse Wang, Joseph Schmidt,
Alborz Ranjbar, Aoran Wu

Motivation and Overview

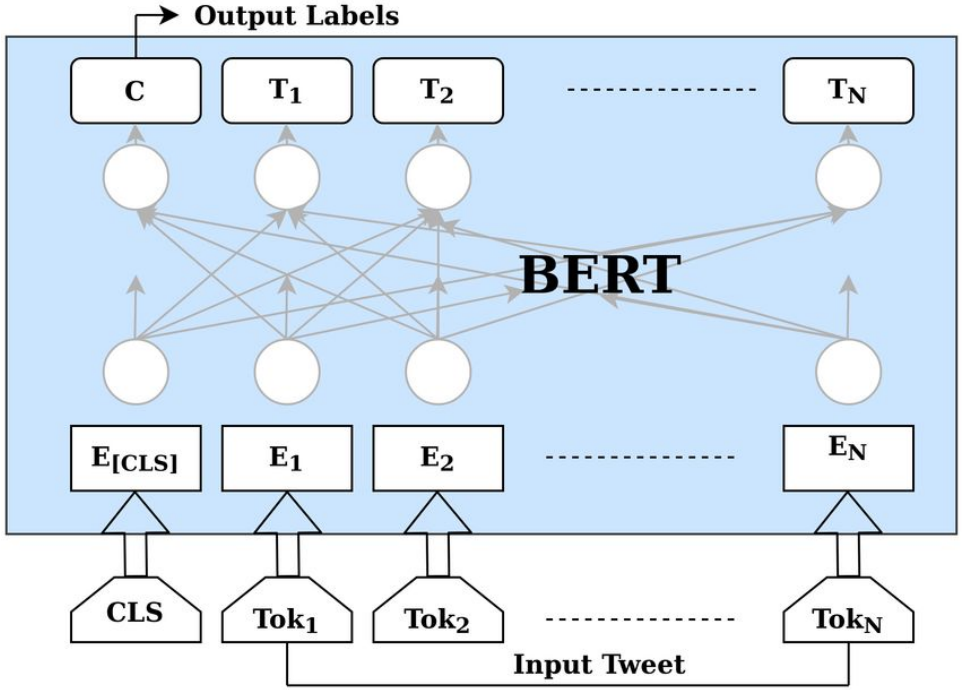
- Various tweets are generated continuously and unconstructively
- Media data can provide insights outside of traditional financial models

Overview of this project:

1. Use historical tweets data with corresponding sentiment to train a sentiment classification model
2. Collect independent new data with raw tweets data and stock data within the same time frame
3. Preprocess the data using pretrained NLP model and extract key factors aligned by time
4. Use the extracted features and apply various DS/ML models to establish bot trading algorithm

BERT & Sentiment analysis

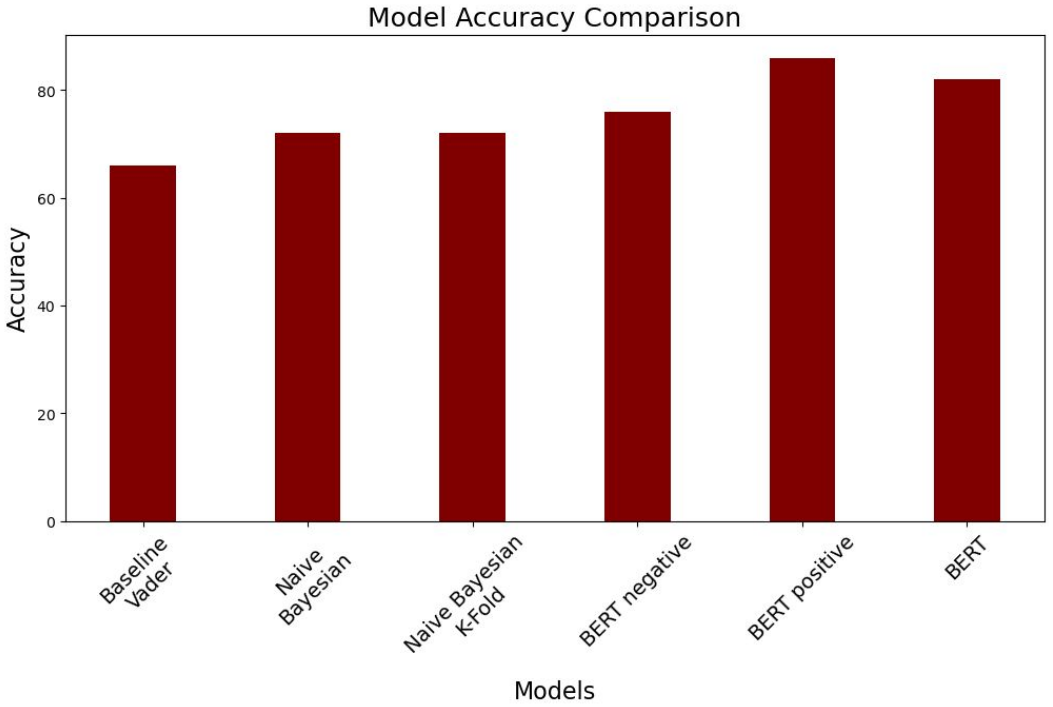
- BERT is a transformer based encoder trained to read in a full sentence to understand context for tasks like:
  - Question and answer tasks
  - Fill-in the blank
  - Sentiment analysis
- By using a pre-trained model to read in tweets we can train it for sentiment analysis



BERT architecture used for reading tweets (Transformer based automatic COVID-19 fake news detection system)

Training BERT for sentiment analysis

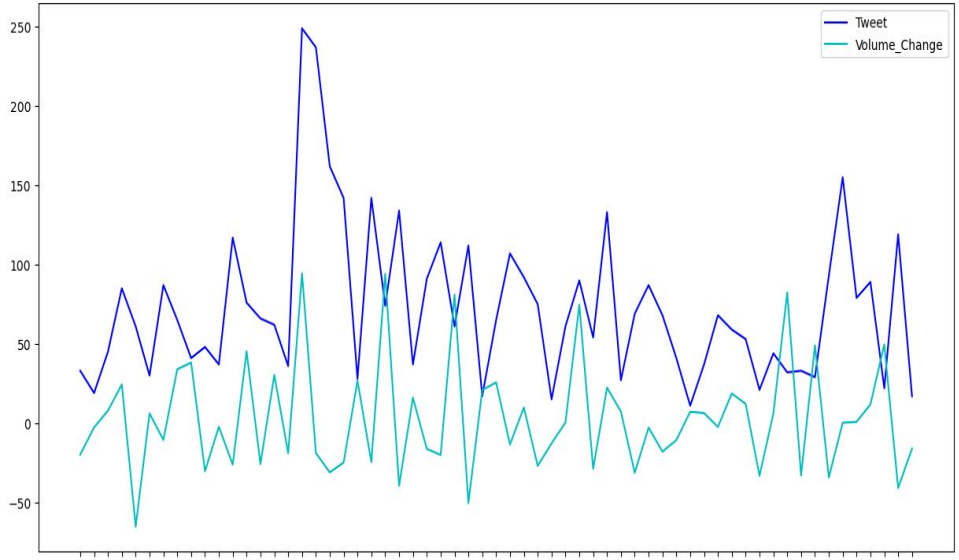
- As tweets often have abbreviations, acronyms, slangs or string numerics (“ten” vs “10”), we preprocess them to standardize the tweets
- We train several NLP models on ~6000 tweets that include a target sentiment:
  - 1 for positive
  - -1 for negative
- Overall, BERT performs better than other NLP models with improved accuracy for accurate tweets

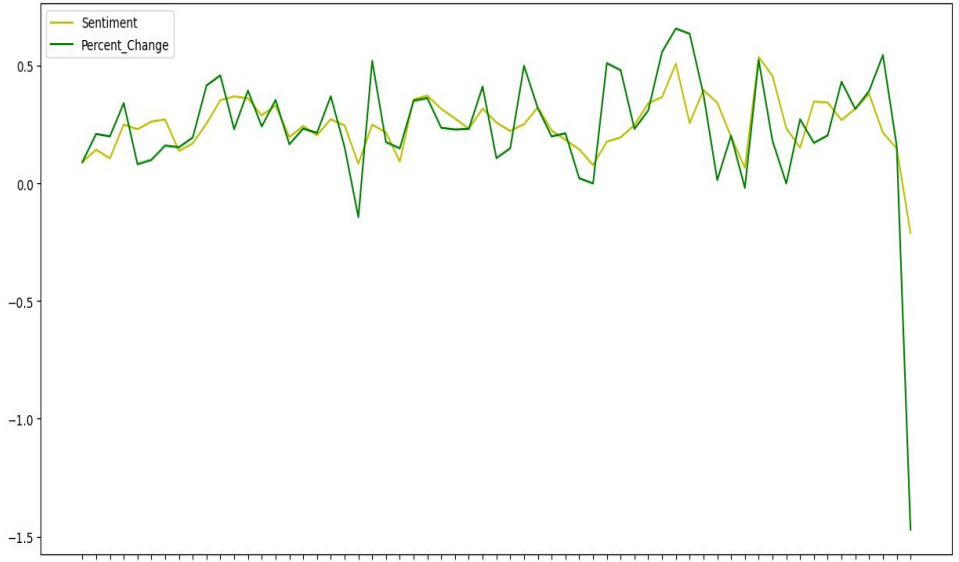


Data used for training

- We used stock data from Yahoo Finance and tweets data from data.world dataset
- We used the BERT model we trained in the previous section to perform sentiment analysis on the tweets data
- To simplify decisions, we classify the sentiment based on the attitude (+/- w.o neutral) with the most likelihood. We further specify the ++ sentiment based on distribution
- We add weight on sentiment based on the influence of the tweet
- We also do rolling average on the sentiment to equalize each time window
- We align the stock data together with the sentiment data by date and ticker
- We choose the percent change between two days adjusted close price as our target of prediction. We have both binary and continuous version available for different models.

Data Analysis

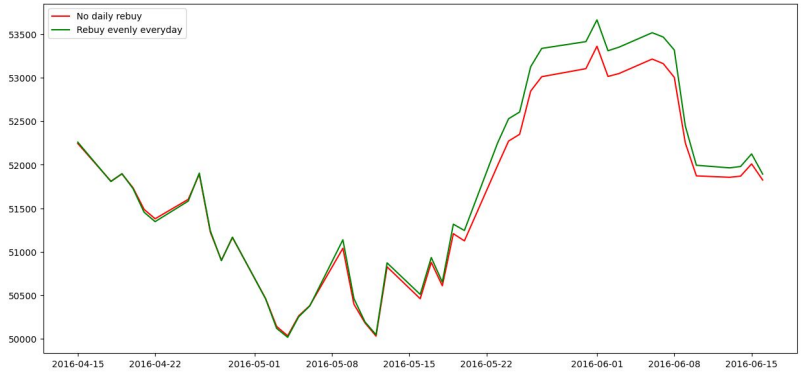


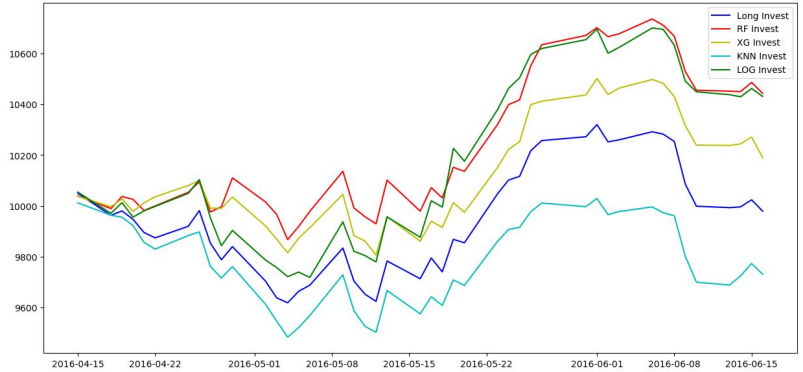


- The sentiment basically aligns with the percent change (after rescaling).
- The volume change basically aligns with the number of tweets

Trading Algorithm

- Two baseline models:
  - Rebuy each stock everyday evenly
  - Hold every stock without further operations
- The models we compare are:
  - Log Regression
  - **RandomForestClassifier** (Best performing)
  - XG Classifier
  - KNN Classifier (worse than baseline)
- Our classification model yields (0,1,2) corresponding to the following actions:
  - 0 - don't buy any stocks
  - 1 - buy regular
  - 2 - buy double





Trading Algorithms - Continued

To see that the sentiment prediction is valid, we need to compare with 2 other RF baseline models:

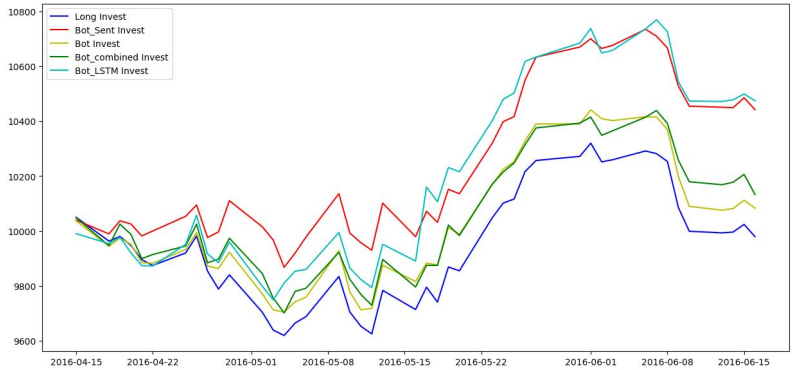
- Prediction based on stock data (High/Low/etc.)
- Prediction based on both stock data and sentiment data

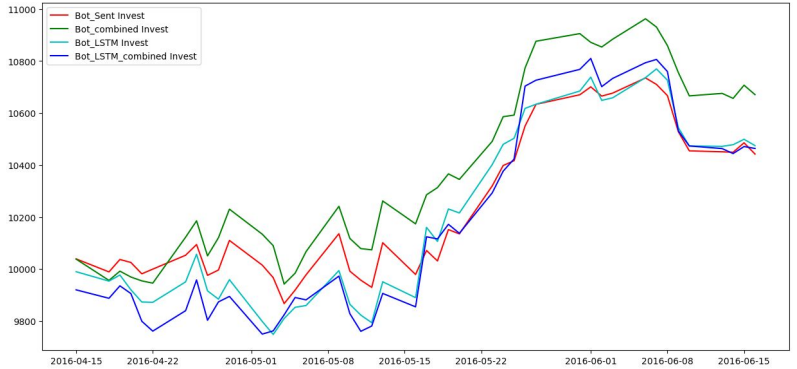
Outside of basic classification models, we use the LSTM neural network model. In order to appropriately apply the model, we did the following:

- Normalize every factor into similar range of values
- Increase the gap between +/- of our target value

We see that LSTM/RF works equally good, we propose a trick model: sometimes traders spread rumors on twitter as a “trap”, we consider the situation when sentiment is + while volume-change is - as a “trap”.

We see that the performance of RF vastly increased while the LSTM performance stay the same in the trick model. Our best model is 7% performance better than the baseline model, and the gap will increase following the trend shown in the graphs.





Summary and Future Steps

Advantages:

- Sentiment analysis is generally performing better than basic stock prediction models and baselines.
- Profitable trading strategy combining RF and LSTM

Future Steps:

- LSTM didn’t outperform RF due to the small amount of parameters
- Work better on enhancing profit when the market is good than preventing loss when the market is bad