NLP Stock Predictor Executive Summary

Team Members: Joseph Schmidt, Alborz Ranjbar, Aoran Wu, Jingheng Wang
Github: https://github.com/HiggsP10/Stock_pred

Goal:
Use sentiment analysis from user tweets to determine how much one should invest in a particular stock. Derive a profiting trading algorithm based on this.

Data:
- Sentiment analysis CSV data from ~5000 tweets (with sentiment score 0/1)
- Tweets containing a stock ticker/names with BERT appended sentiment
- 2016/03-2016/06 Stock data from stock tickers mentioned in tweets
- Every tweets data is preprocessed to better fit into the model

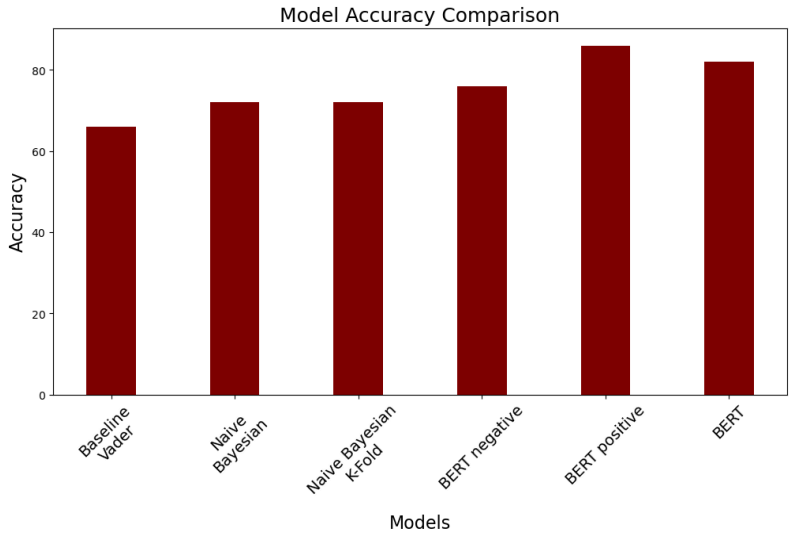
Evaluation Methods:
- Sentiment analysis
    - Generally, all models were compared to each other using **accuracy** (correct guesses/total data)
    - Specifically
        - Naive Bayes used Cross-Validation AUC (scoring="roc_auc") to find the best performing model amongst several
        - For the best model, BERT, negative and positive sentiment was also computed to see any bias in the algorithm
- Stock prediction
    - All models are compared to each other using the outcome of trading $10000 during the same timeframe under each model.
    - Baseline model uses:
        - Long-term holding each stock without additional operations
        - Long-term holding but rebuy each stock evenly everyday
    - In order to show that Sentiment analysis works, we also use ML-model without sentiment analysis as a comparison

Models:
- Sentiment analysis
    - **BERT** - Model selected for final sentiment analysis
    - Vader
    - Naive-Bayesian
- Stock prediction
    - Classification (XGBoost/RandomForest/KNN/Log-regression)

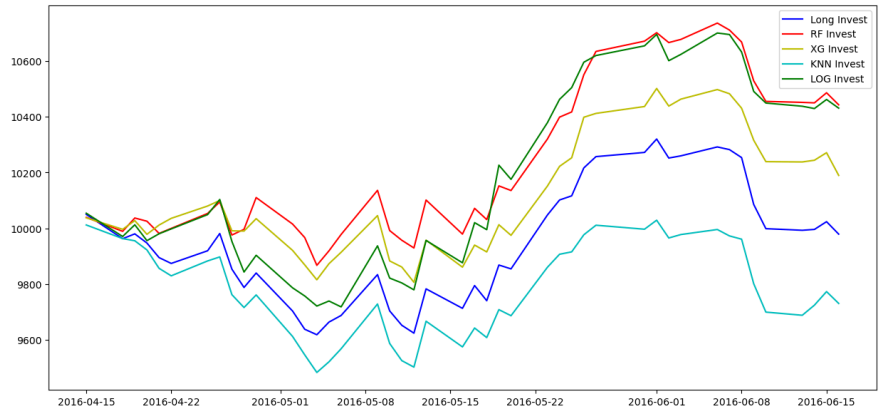
- Shallow Neural Network (LSTM)
- Bonus model that avoid media trap (when stock volume decrease while sentiment is positive)

Comparison of Models:

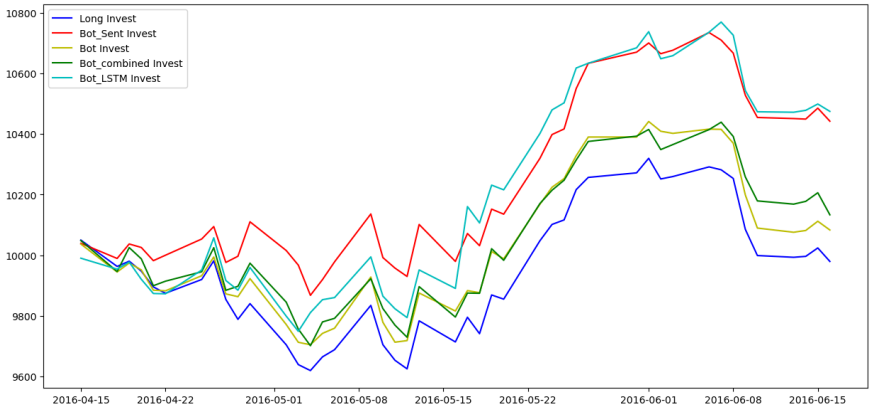
- Sentiment analysis
  - Amongst all models compared, BERT performed better than the Vader and Naive Bayes models in overall accuracy
  - We see that BERT's positive and negative sentiment accuracy performs above all models as well but there is a bias toward better prediction for positive sentiment guesses



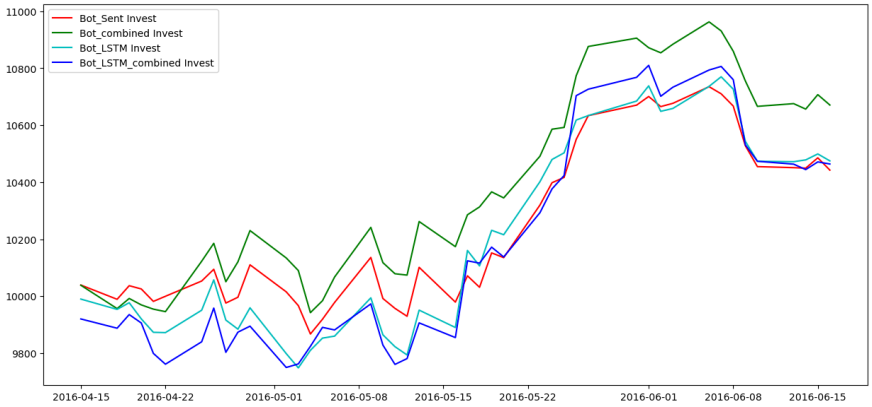
- Stock prediction
  - Among all classification models, RandomForest works the best



  - LSTM works slightly better than Random Forest
  - Every ML-model works better than the baseline, while the sentiment analysis model works the best



- Avoiding media trap successfully increase the profit of the model



Future Directions:

- LSTM didn't outperform RF due to the small amount of parameters. In order to optimize the prediction, we need to creative more parameters and add weight on each depending on data analysis
- Work better on enhancing profit when the market is good than preventing loss when the market is bad. Consider a better preprocessing method or better pretrained NLP model.