**CSCE 5290: Natural Language Processing**

**Project Increment 1**

**Project Proposal**

This project is my personal attempt to predict the stock market. It is a challenging task to predict the stock market with a high degree of accuracy because there are so many factors affecting the prediction. Fundamentals of the market, human behaviors, and physical and psychological factors are among factors which can add noise to the model. That being said, correctly identify the market movement will result in lucrative rewards.

This project will try to predict stock price/movement of APPL and the S&P 500 using sentiment analysis of social media.

**Project Proposal Description:**

**1. Project Title and Team Members:**

Project Title: Sentiment Analysis of Social Media for Predicting Stock Market Movements

Team Members: I, Truc Nguyen, will be a sole member of the project

Source code and all related documentations will be uploaded to: <https://github.com/trucntx0550/NLP>

**2. Goals and Objectives:**

• Motivation: The popularity of social media and investment platforms has attracted more and more retail investors to participate in the stock market. It has, for the time being, changed the way the markets behave. Investing is no longer just the game of WallStreet and Hedge [fund’s](doi:%2010.1109/BigData.2018.8621884) managers. Retail investors, with the enabling of social platforms, can gather together to manipulate a company’s stock. The recent example of WallStreetBets has proved the influence of social media to the market movement. Correctly analyze the posts/tweets can lead to an improvement in predicting the market.

• Significance: This project aims to expand the sentiment analysis dictionary for the financial sector introduced by D. Shah et al. ([2018](https://ieeexplore.ieee.org/document/8621884)).

• Objectives: Find the correlation between the sentiment on social media platforms and a particular stock movement. This can be later expanded to a group of stocks.

• Features:

- Stock price will be pulled from Yahoo Finance for stock price prediction

- Data from Twitter, Stockwits, Reddit, etc. will be used for sentiment analysis.

**Increment 1**

1. Related Work (Background)

This project is inspired by Khedr and Yaseen ([2017](https://doi.org/10.5815/ijisa)). They developed a model to predict stock market future trends with small error ratio and improve the accuracy of prediction using KNN and naïve Bayes algorithm.

Das et al. ([2018](https://doi.org/10.1016/j.procs.2018.05.111)) work study the correlation between stock market and twitter data. Their classifying model takes in historical data to improve the accuracy of the predictions with the assistance of Twitter data.

The impact of COVID-19 has fundamentally changed the way investors view the market. Lee ([2020](https://doi.org/10.3390/su12166648)) explores the initial impact of COVID-19 sentiment on US stock market using big data notedly Daily News Sentiment Index (DNSI) and Google Trends data on coronavirus-related searches. This study investigated the correlation between 11 indices and investors’ sentiment during COVID. They have found a significant connection between COVID-19 sentiment and various industries and classified them into different correlation groups.

1. Dataset

The dataset was obtained from Kaggle. It combined stock news and the related stock prices from 2008 to 2016.

Dataset description:

* 4101 rows x 27 columns:
  + Top daily news headlines – 25 columns
  + Date
  + Label (target variable): Dow Jones Industrial Average stock index. Taking 2 values:
    - 1 – The stock price increased when
    - 0 – The stock price stayed the same or decreased.

Table

Description automatically generated with medium confidence

1. Detail design of Features

I have decided to keep all the features in the original dataset since the headlines may have equally effectiveness on the target value.

The dataset is divided into training and testing using TimeSeriesSplit from sklearn.model\_selection

Stop words and special characters are remove from the dataset using regular expressions

Text

Description automatically generated with low confidence

Column names are removed since they are irrelevant to the sentiment analysis

Graphical user interface, text, application

Description automatically generated

Headlines of a same date are combined into a paragraph to create BagofWords to use for sentiment extraction

Text

Description automatically generated

1. Implementation

headlines is used to train the model

Text

Description automatically generated

For the time being, I used RandomForest Classifier. If time permits, other models will be implemented.

Text

Description automatically generated

1. Preliminary Results

A screenshot of a computer

Description automatically generated with medium confidence

1. Project Management

Implementation status report:

* Work completed:
  + Cleaned the dataset
  + Feature engineering to select appropriate features for the model
  + Implemented RandomForest Classification to predict whether the index will increase (1) or decrease (0) based on headlines
* Work to be completed
  + applied other approaches of sentiment analysis, such as the lexicon-based approach of SentiWordNet
  + Implement other models, such as RNN, DNN and Naive Bayes classifiers

**References**

<https://github.com/you915/Sentiment-Analysis-of-Twitter-Data-for-predicting-Apple-stock-price>

Pagolu, Challa, Panda, Majhi, Sentiment Analysis of Twitter Data for Predicting Stock Market Movements. International conference on Signal Processing, Communication, Power and Embedded System (SCOPES)-2016

A. Pak and P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, in Proceedings of the Seventh International Conference on Language Resources and Evaluation, 2010, pp. 13201326

D. Shah, H. Isah and F. Zulkernine, "Predicting the Effects of News Sentiments on the Stock Market," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 4705-4708, doi: 10.1109/BigData.2018.8621884.

László Nemes & Attila Kiss (2021) Prediction of stock values changes using sentiment analysis of stock news headlines, Journal of Information and Telecommunication, 5:3, 375-394, DOI: 10.1080/24751839.2021.1874252

Khedr, A. E., & Yaseen, N. (2017). Predicting stock market behavior using data mining technique and news sentiment analysis. International Journal of Intelligent Systems and Applications, 9(7), 22. https://doi.org/10.5815/ijisa

Das, S., Behera, R. K., & Rath, S. K. (2018). Real-time sentiment analysis of twitter streaming data for stock prediction. Procedia Computer Science, 132, 956–964. <https://doi.org/10.1016/j.procs.2018.05.111>

Lee, H. S. (2020). Exploring the initial impact of COVID-19 sentiment on US stock market using big data. Sustainability, 12(16), 6648. https://doi.org/10.3390/su12166648