Stock Market Price Prediction with a combination of NLP and Time Series

Sai Navyanth VOBBILISETTY

(55357700)

Department of Computer Science

City University of Hong Kong

Supervisor: Dr. SONG, Linqi

**Motivation and Background Information**

In recent years, the number of retail investors has significantly increased [1]. This rise can be attributed to the pandemic since retail investors became aware of and better informed about stock markets. Another factor that led to this is influential people like Elon Musk, etc.

Most retail investors invest based on the crowd/momentum and get losses. I have had to bear witness to the losses of many people who had invested their own money into the market as first-time retail investors with little knowledge about the intricate workings of the market. Moreover, it is very time-consuming to keep track of the various companies and industries in the stock market manually.

As a solution to this problem, my teammate (Varun) and I decided on this topic for our final year project, where we endeavor to assist retail investors to predict better whether they should invest in a particular company by checking the company's recent news and historical data to forecast the price of the stock.

The recent works in this area mainly were done only to predict the price of a stock using NLP on stocks’ news [2] or through time-series analysis of the historical prices of the stocks. [3,4,5] But prediction solely based on this information is insufficient. These don't attempt to explain how much money can be invested with their predictions or provide information about real life traders using their proposed solutions.

We use two different methods to predict the stock price/sentiment of a company.

1. NLP on news articles to predict the sentiment of a stock. (Done by my teammate)

2. Time Series Analysis to predict the closing price of a stock. (Done by me) In the end, we have a weighted combination of both these outputs to have a better prediction analysis.

# Problem Statement, Project Objective, and Scope

Historically, stock markets have primarily been influenced by the opinions of top political leaders or influencers. Any positive or negative comment could lead to a short-term change in the market accordingly. Most of the standard and rookie traders worldwide don't know how specific comments affect the stock market. Hence, they are misguided and ultimately lose money. This case is very valid for short-term traders and day traders who depend on short-term fluctuations in the market.

The majority of stock market research focuses on neither purchase nor sell, but it ignores the dimensions and expectations of a new investor. The general consensus in society is that the stock market is either too hazardous to invest in or too volatile to trade in, therefore most people are uninterested.

Our algorithm focuses on solving this problem. We use a combination of time series modeling and NLP-based sentiment analysis on news headlines and social media articles/tweets to predict the stock price in a particular time frame. The combination is meant to average out the errors in both models and provide a better picture of the stock price in the near future.

In this way, it is going to be simpler for retail investors to start investing in the share market.

(I will be working on the time-series part of the project).

## Scope:

1. Collect and preprocess closing prices of all companies present in the NASDAQ 100 index as of September 19, 2o21.
2. Use this cleaned data as an input to our LSTM model.
3. Explore other types of statistical methods (e.g., Moving Average, Exponential Moving Average, etc.).
4. Train different methods and finalize the best method by comparing the results from all the methods.
5. Containerize the code and create a web app with a server to make live predictions.

## Assumptions:

We are only considering the 100 companies present in the 'NASDAQ 100' index and train our model only with them. But, if given more time, we can add more companies to this list in the future. We only consider the 'closing price' on a particular day.

## Cases that will not be considered:

Companies that are not present in the 100 companies of 'NASDAQ 100' will not be considered.

From now on, the project plan would explain only the parts where I am responsible.

**Major Technical Components**

This project will be based on Statistical Analysis and Machine Learning models:

1. Statistical Analyses like Moving Average and Exponential Moving Average will be experimented with.

2. Machine Learning models like LSTM, RNN, and GRUs will be tested and experimented with.

In the end, we create a web app hosted on a server, where the users can see the live predictions.

So, the whole system will be implemented as follows:

**For preprocessing the training data:**

The 'yfinance' library of python will be used to get the data of all the companies present in NASDAQ 100. Later, we preprocess the data by handling missing values and then normalizing the data in the same range.

**Statistical Models:**

We find out the cleaned data's moving averages and exponential moving averages by implementing standard functions in python using the NumPy library. We will be using the window-scaling method to find out moving and exponential moving averages respective to windows.

**Machine Learning Models:**

Different Deep Learning models like LSTM, RNN, GRUs will be trained. The training data would be so that, to predict the price on a particular day, the prices of the stock in the previous 50 days would be the input to the model. So, we need to create the datasets so that input has 50 values, and output has one value of the next day. The value 50 here doesn't have to be a fixed value, and it can be changed and tried with different values to give the best output.

In this way, after we create the dataset, we train the LSTM model, multiple models, with different parameters to select the best model.

**Web app:**

After the model has been trained and evaluated, a web app will be created to display the results of our prediction of a selected stock by the user.

To make the web app's front-end, Vue.js, Node.js will be used. MySQL database will be used for storage.

For the backend, python APIs of the model will be created using the Flask framework.

**Evaluation:**

Evaluating the prediction results might be a bit difficult task. So, a paper trading account in any stockbrokers like Interactive Brokers, Robinhood, etc., will be opened to backrest the strategy using the model trained.

Various financial metrics like Sharpe ratio, Profit, etc., will be calculated to determine how the model has performed. If the ratios turn out to be bad, then retrain our model with different parameters again. In this way, the results of the system can be quantified.

**Expected Results and Deliverables**

As explained in the previous section, backtesting will be done to test the model's predictions. Various financial metrics like the Sharpe ratio, etc., will be calculated to validate the predictions.

A web app will be created with a single screen to display the results from the time series model, and the web app will also allow users to select a specific company to find out the prediction of it.

Hopefully, the model from my teammate (Varun) will also be added to find out a weighted prediction instead of only depending on the outputs of time series analysis.

When both the models are ready, a new screen will be added to the web app to showcase the prediction results of the weighted output of both outputs.

Finally, there would be three screens in the web app:

1. One main screen which displays result from both models.
2. One screen to display detailed information and results from only the time series model.
3. One screen to display detailed information and results from only the NLP model

**Project Schedule**

**Milestones:**

|  |  |
| --- | --- |
| **Task** | **Tentative Deadline** |
| Research:   1. Deep Learning models [7] 2. Web development stacks (React, Node) | October 15th, 2021 |
| Data Collection and Preprocessing:   1. Collect data from all 100 companies 2. Clean the data with respective cleaning methods | November 1st, 2021 |
| Model development and training   1. Develop statistical models like moving average. Exponential moving average, etc 2. Develop and train on deep learning models like RNN, LSTMs | December 15th, 2021 |
| Parameter tuning:   1. Improve the models by tuning its parameters | January 1st, 2022 |
| Evaluation and retraining:   1. Create a paper trading account in any of the available stock borkers. 2. Explore the API of the broker 3. Backtest the trained model using paper cash using the API 4. Calculate different financial metrics like the Sharpe ratio 5. Retrain the model if the metrics are unsatisfactory | February 28th, 2022 |
| Web app development   1. Research on the architecture. 2. Work on both the front and back-end to display the results | March 15th, 2022 |
| Combine results from the teammate   1. Create a weighted output from both the models. (NLP and time series) 2. Create a new screen explaining and displaying the results | April 1st, 2022 |

Gantt Chart:

Graphical user interface, application

Description automatically generated

**References**

1. Fitzgerald, M. (2021, August 18). *A large chunk of the retail investing crowd started during the pandemic, Schwab survey shows*. CNBC. <https://www.cnbc.com/2021/04/08/a-large-chunk-of-the-retail-investing-crowd-got-their-start-during-the-pandemic-schwab-survey-shows.html>
2. Mehtab, S. (2019, December 9). *A Robust Predictive Model for Stock Price Prediction Using Deep. . .* ArXiv.Org. https://arxiv.org/abs/1912.07700
3. Shen, J. (2020, August 28). *Short-term stock market price trend prediction using a comprehensive deep learning system*. Journal of Big Data. <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00333-6>
4. Hoare, T. (2019, July 11). *Stock price prediction using time series models*. Dublin Business School. <https://esource.dbs.ie/handle/10788/3830>
5. Mehtab, Sidra & Sen, Jaydip. (2020). A Time Series Analysis-Based Stock Price Prediction Framework Using Machine Learning and Deep Learning Models.
6. Mallikarjuna, M. (2019, December 2). *Evaluation of forecasting methods from selected stock market returns*. Financial Innovation. <https://jfin-swufe.springeropen.com/articles/10.1186/s40854-019-0157-x>
7. Staudemeyer, R. C. (2019, September 12). *Understanding LSTM -- a tutorial into Long Short-Term Memory. . .* ArXiv.Org. https://arxiv.org/abs/1909.09586