Research Questions

Ri-on Kim, Manuelito Bagasina, Youngjun Jhee

January 24, 2025

1 Background

Stock market price prediction has been a longstanding challenge in finance due to the complex interplay of various factors influencing market movements. Traditional models often rely heavily on historical price trends and technical indicators, which may overlook crucial external factors such as macroeconomic changes, firm-specific financial performance, and sentiment-driven market dynamics.

Incorporating macroeconomic indicators like GDP growth and inflation, firm-specific metrics(microeconomic indicators) such as P/E ratios and EPS, and sentiment data from news articles and social media offers a holistic approach to understanding stock price dynamics. Natural Language Processing (NLP) techniques enable the quantification of market sentiment, providing insight into short-term and long-term investor behavior.

This project combines these diverse data sources to create a comprehensive stock price prediction model that surpasses the limitations of traditional approaches. Integrating data science and machine learning has the potential to enhance investment strategies and decision-making for both retail and institutional investors.

2 Research Question

How can we effectively predict the stock price or movement of a given company using a combination of historical stock data, macroeconomic indicators, microeconomic data, and sentiment analysis?

2.1 Sub-Question 1

Are certain industries or sectors (e.g., technology, healthcare, energy) more predictable than others?

Industries behave differently based on unique factors. Understanding this variability can help create sector-specific models.

2.2 Sub-Question 2

How does sentiment analysis improve stock price prediction when combined with traditional economic indicators?

Investor sentiment extracted from news articles and social media can provide real-time insights into market behavior, complementing traditional stock prediction methods. However, the effectiveness of sentiment data remains uncertain—does it enhance predictive accuracy when combined with historical stock data, macroeconomic indicators, and microeconomic metrics? Additionally, does sentiment analysis contribute more to short-term price fluctuations or long-term trends? Understanding the role of market sentiment in stock prediction can help optimize feature selection and improve model performance.

2.3 Sub-Question 3

Which features (historical stock data, macroeconomic indicators, macroeconomic indicators, sentiment scores) contributed the most to predicting stock price movement? / Which machine learning model performed the best depending on the features?

The features will be the ones to help predict the stock price movements and their use will help refine the models' accuracy. Historical stock data, including price and volume, provides the foundation for time-series analysis, with technical indicators derived from this data often being key predictors. Sentiment scores from social media and financial news articles are significant predictors, reflecting investor sentiment which can drive market movements. Identifying the most influential features allows researchers to focus on collecting and refining the most relevant data, thereby improving the accuracy of stock market predictions. It's also important to fine-tune future models, in the effort to help predict stock market price movement more accurately.

3 Data

3.1 Stock Data

Daily stock prices: Open, Close, High, Low, Volume.
API: Yahoo Finance [8].

3.2 Macroeconomic Indicators (Low-Frequency Data)

Interest rates, exchange rates, GDP growth, and the Consumer Price Index (CPI).
Frequency: Monthly or quarterly.
Sources: FRED [2], IMF [3], World Bank [7].

3.3 Microeconomic Indicators (Low-Frequency Data)

Firm-specific financial metrics: EPS (Earnings Per Share), P/E ratio, revenue growth.
Sources: Company financial reports, SEC filings, or stock analysis platforms like Bloomberg or Yahoo Finance [8].

3.4 Market Sentiment Data

News articles and social media data (Twitter, Reddit) [1, 6].
Extracted sentiment scores using Natural Language Processing (NLP).
** All data should be organized as a daily time-series table at the end of the data collection process.*

4 Methods

Our project will be conducted using Python and Jupyter Notebook environments (Google Colab could be a possible environment for the project). We will combine stock prices, macroeconomic indicators, microeconomic indicators, and quantified market sentiment data from news and social media using NLP techniques [1, 5] to create features for machine learning. The data will be organized as a table and used for EDA and Feature Engineering to extract the most relevant data for the prediction. Using the features, various machine learning algorithms [4] can be applied. It is important to handle the data carefully so that it is suitable for time series data.

References

[1] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[2] Federal Reserve Economic Data (FRED). Federal reserve bank of st. louis. Retrieved from `https://fred.stlouisfed.org`.

[3] International Monetary Fund (IMF). Data and statistics. Retrieved from `https://www.imf.org/en/Data`.

[4] Janak Patel, Sahil Shah, Parth Thakkar, and Krishna Kotecha. Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications*, 42(4):2162–2172, 2015.

[5] Robert P. Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfintext system. *ACM Transactions on Information Systems (TOIS)*, 27(2):1–19, 2009.

[6] Twitter. X api. Retrieved from `https://developer.x.com/en/docs/x-api`.

[7] World Bank. World development indicators. Retrieved from `https://data.worldbank.org`.

[8] Yahoo Finance API. Yahoo finance. Retrieved from `https://finance.yahoo.com`.