Predicting Stock Market Trends - Midway Report

Youssef Briki
youssef.briki@umontreal.ca

Yuan Li
yuan.li@umontreal.ca

Xintian Xu
xintian.xu@umontreal.ca

Abstract

This project aims to build a robust predictive model capable of forecasting stock prices and market trends over short-term periods by using sentiment analysis and multi-modal approach on historical stock data, social media posts, and news data. In this midway report, we will discuss our progress and challenges encountered, especially in data acquisition, features extraction, and multi-modal LLM fine-tuning. We will address our findings and their impacts on adjustments we made to the original proposal in order to better fit the project's objective. Specifically, we have changed our multi-modal LLM from Llama-3.2 to Deepseek-R1, and we will directly fine-tune this LLM on FinBERT dataset in order to get sentiment scores for textual data.

1 Introduction

The stock market is influenced by many aspects, such as government and economic policies, company and consumer behaviours, public opinions, and more. These sentiments about the future trend of the market are often reflected in general and financial news, company earning calls, and public discussions on social media. This project aims to compare different approaches to building a multi-modal deep learning model that can predict the short-term stock market behaviour. It focuses mainly on sentiment analysis on textual data such as news, earnings call transcript, and social media posts, and is supplemented by analysis on other types of data such as historical stock price and macroeconomic data. For this midway report, we will discuss our work in finding and acquiring suitable data sources, namely webscraping on Reddit and using Alpha Vantage API, and fine-tuning Deepseek-R1 in order to get sentiment scores on the textual data.

2 Related Work

2.1 FinBERT

One of the models that we will utilize is FinBERTAraci (2019), a version of BERT Devlin et al. (2019) pre-trained on financial corpus and fine-tuned for sentiment analysis for finance. It was found that FinBERT performed better than Vanilla BERT and other language models on finance tasks. The reason is although there exist models that can perform well in general sentiment analysis tasks, they may be more lacking in domain-specific tasks due to specialized vocabulary and data of that domain. In stock prediction, an understanding of finance language is crucial and one of the main challenges for sentiment analysis. Therefore, This paper also discusses the uniqueness of sentiment analysis in finance compared to general sentiment analysis tasks, namely that in finance the goal is to utilize sentiment scores to forecast the market. Therefore, this project draws inspiration from this paper, and will also utilize other pre-trained LLMs and fine-tune them for sentiment analysis stock market prediction.

2.2 Financial Sentiment Analysis: Techniques and Applications

In this paper by Du et al. (2024), a review of studies on techniques and applications of financial sentiment analysis is conducted. It discusses the interaction between financial textual sentiments, investor sentiments, and other market information with the market itself and stock prices. Textual sentiment information can be classified as subject or objective, in which the objective is often official information such as macroeconomic data to which investors later react, providing subjective data. Together, these two types of data provide sentiments that can influence stock market behavior. Various techniques used for financial sentiment analysis are reviewed, including machine learning, deep learn-

ing, and pre-trained models. Benchmarks such as PhraseBank, FiQA Task 1, SemEval 2017 Task 5 etc. are used for comparison. Although this paper can provide a more comprehensive comparison on how different techniques and models perform on the described benchmark datasets, it provides an overview and knowledge foundation for which we will leverage and consider in our project.

3 Method

Due to the sensitive nature of financial data, we attempted several ways of acquiring good quality data. The social media data was fetched using web-scraping. However, it was challenging to scrape financial news and earnings call transcripts as there are many restrictions and access barriers on more recent data. Therefore, we got our finance news and historical stock data from Alpha Vantage API.

For our experiments, we will use as a baseline sentiment scores from FinBERT and combine them with stock time-series prediction using LSTM. Then, we will finetune an LLM, namely Deepseek-r1, in order to fit it for finance data and acquire sentiment scores, and get the stock price prediction using LSTM.

4 Experiments

4.1 Dataset

- **Social Media:** Reddit
- **News outlets & Quarterly Earnings Call Transcripts:** Alpha Vantage API data (50 of the most traded stock tickers/companies), goes back at most 5 years for news, and 4 quarters in 2024 for earnings call transcripts. The API provides sentiment scores for textual data, however, we would compare them to FinBERT/Deepseek-r1 in order to align the scoring procedure with social media data.
- **Historic stock price:** Alpha Vantage API daily stock data (50 of the most traded tickers/companies), goes back 100 days, with daily open, high, low, close, and volumn data.
- **Macroeconomic:** World Bank.

4.2 Baseline

In the paper byJain and Agrawal. (2024), the authors employ a combination of FinBERT and GAN to predict stock prices, comparing their results with multiple models. Their experimental outcomes will serve as our baseline. Traditional Financial Indicators: Baseline predictions based purely on macroeconomic indicators (e.g., GDP, inflation, interest rates) without leveraging sentiment data.

4.3 Evaluation Methods

We will evaluate the performance of our models using the following metrics:

Stock Market Prediction Metrics:
Directional Accuracy: Measures how often the model correctly predicts the direction of the market movement (up / down).
Mean Absolute Error (MAE) , Root Mean square Error (RMSE): Evaluate prediction errors in stock price forecasting.
Correlation Coefficient (R): Measures the strength of correlation between sentiment-based predictions and actual stock market movements.

4.4 Experimental Details

The experiments are divided into following tasks.

4.5 Data Acquisition

1.Use AlphaVantage get news and stock data.
Data contains:

- For 50 of the most traded tickers/companies: news articles (earliest is 2019 if available, and onwards), earning call transcripts (Q1-4 2024), daily stock price (past 100 days)
- News articles in specific topics: `ipo`, `earnings`, `mergers-and-acquisitions`, `economy-macro`, `economy-fiscal`, `economy-monetary`, `technology`, `finance`
- Economic indicators : GDP, unemployment, inflation, interest rates

Data is in JSON format. (Finished work)
2. Use webscraper to get Reddit and Bluesky data. (Finished work)
3. Use API get world bank macro encomic data. (Finished work)
Data contains: GDP data, interest rates, unemployment data, inflation data. (Finished Work)

4.6 Calculate Sentiment Score

We are using API to get sentiment score for news data. For Social Media Data, we are using Reddit data collected from the following subreddits via the PRAW API (the official Reddit API):

- /r/stocks
- /r/StockMarket
- /r/StocksAndTrading
- /r/investing
- /r/wallstreetbets
- /r/pennystocks
- /r/options
- /r/valueinvesting
- /r/securityanalysis
- /r/Economics
- /r/finance
- /r/business

For each subreddit, we retrieve the 20 hottest (i.e., most upvoted recently) posts. For each post, we extract the title, description, number of upvotes and downvotes, and the top 5 comments. Currently, we do not consider the post date, which means we are retrieving the most recent posts from the past few days. However, incorporating older posts could be a valuable addition.
We are also working on making this analysis multimodal by including image attachments. These will be analyzed using multimodal LLMs such as Llama 3.2 or Qwen.
In addition to Reddit, we are collecting data from the emerging social media platform Bluesky, which gained significant traction in 2023. For each relevant post, we extract the author, content, date, and image URL. This enriches the dataset obtained from Reddit.
The following topics are used to filter Bluesky posts:

- stock market
- finance
- investing
- trading
- cryptocurrency
- Bitcoin
- Ethereum
- blockchain
- IPO
- Wall Street
- dividend stocks
- market volatility
- ETF
- options trading
- futures trading
- day trading
- financial crisis
- economic indicators
- global markets
- forex
- asset management
- market analysis
- risk management
- emerging markets
- tech stocks
- earnings reports
- mergers and acquisitions

All collected data from this section is stored in JSON format.

4.7 Combine all the data to time series data

1.Data preprocessing (normalization, etc.) (In progress)
2.Combine all the news data and sentiment score, social media data and sentiment score, stock history price data and macro encomic data to one time series data.(In progress)

4.8 Experiment 1: FinBERT & LSTM (In Progress)

This experiment aims to evaluate the performance of a hybrid model that combines FinBERT—a domain-specific BERT model fine-tuned for financial texts—with an LSTM for sequential data analysis.

1. **Model Architecture:**
   Use FinBERT to generate contextual embeddings from financial text.
   Add an LSTM layer to capture sequential dependencies in the embeddings.

2. **Training Process:**
   Apply FinBERT to the target financial dataset to extract embeddings.
   Train the LSTM on these embeddings to learn temporal patterns.

3. **Prediction:**
   Use the trained FinBERT+LSTM model to make predictions on held-out test data.

4. **Expected Outcome:**
   Enhanced accuracy in financial sentiment or text analysis compared to baseline models.

4.9 Experiment 2: Deepseek-r1 & LSTM

Assess the effectiveness of Deepseek-r1 (32B), a large language model, paired with LSTM for enhanced sequential prediction tasks.

1. **Model Architecture:**
   Use Deepseek-r1 (32B) for the sentiment analysis part.
   Combine with LSTM to model temporal patterns.

2. **Training Process:**
   Fine-tune Deepseek-r1 on the finbert dataset.

3. **Expected Outcome:**
   Potential improvements in tasks requiring both contextual understanding (Deepseek) and sequential modeling (LSTM).

5 Future Work (Next 4 Weeks)

1. Finish data combining

2. Implement experiment 1 FinBERT & LSTM (week 1)

3. Implement experiment 2 Deepseek-r1 & LSTM (week 1-2)

4. Evaluate result and identify improvements(week 3)

5. Develop final project report and presentation (week 2-4)

6 Task Division

1. Search and acquire data (social media - Youssef, news/transcripts/stock price - Xintian, macroeconomic data - Yuan)

2. Finish data combining. (Xintian, Youssef, Yuan)

3. experiment 1 FinBERT & LSTM (Xintian, Yuan)

4. experiment 2 Deepseek-r1 & LSTM (Youssef)

5. Develop project report (All team member)

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *Preprint*, arXiv:1908.10063.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186.

Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. 2024. Financial sentiment analysis: Techniques and applications. *ACM Comput. Surv.*, 56(9).

Abhimanyu Dubey, Abhinav Jauhri, and Abhinav et al. Pandey. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jainendra Kumar Jain and Ruchit Agrawal. 2024. Fb-gan: A novel neural sentiment-enhanced model for stock price prediction. *Preprint*, arXiv:2407.21783.