Healthcare Sentiment Analysis for Market Trends

Johann Tin Tsin Fong
Simon Fraser University
`jtintsin@sfu.ca`

Emily Wong
Simon Fraser University
`ewa47@sfu.ca`

Abstract

This project explores the use of healthcare sentiment analysis for stock market predictions using natural language processing (NLP) techniques. The objective is to assess the impact of news on stock prices in the healthcare sector by extracting key entities and analyzing sentiment. The approach integrates Named Entity Recognition (NER) to identify company names, drug names, healthcare policies, and other relevant keywords from news articles, ensuring a focus on healthcare-specific articles for sentiment classification.

To achieve this, we use existing BERT model and fine-tune it to classify sentiment as positive, neutral, or negative and the score result is compared against FinBERT, an existing financial sentiment analysis model. The resulting sentiment scores are used as inputs for a Long Short-Term Memory (LSTM) model to predict stock price movements. The effectiveness of our fine-tuned model + LSTM is evaluated against FinBERT + LSTM, with predictions assessed based on their correlation with actual stock price trends in the healthcare sector.

By combining NER-driven sentiment analysis with deep learning-based stock predictions, this project aims to generate trend line predictions and offer clearer insights into how events impact stocks in the healthcare sector.

1 Introduction

This project explores healthcare sentiment analysis for stock market predictions using NLP techniques. The objective is to assess the impact of news on stock prices in the healthcare sector. Our approach integrates Named Entity Recognition (NER) to extract company names, drug names, healthcare policies and other keywords from news articles, filtering healthcare-specific articles for analysis.

We selected this project because news articles are just one of many important factors that influence market prices. For instance, UnitedHealth's stock experienced a sharp decline following the shocking news of its CEO's assassination, resulting in billions of dollars in lost market value(**?**). This event underscores how unexpected occurrences can significantly impact stock movements, a phenomenon also analyzed in studies such as *Impact of Major Health Events on Pharmaceutical Stocks*(**?**).

We fine-tune existing BERT-based model from HuggingFace (**?**) to classify sentiment as positive, neutral, or negative and compare its performance against FinBERT. FinBERT is a BERT-based NLP model trained on financial texts to classify sentiment (positive, negative, or neutral) for stock market analysis (**?**). These sentiment scores are then fed into Long Short-Term Memory (LSTM) models to predict stock price movements. Specifically, we compare the performance of our fine-tuned model with LSTM against FinBERT + LSTM to evaluate their effectiveness in generating buy/sell signals for healthcare stocks.

Our data sources include Yahoo Finance for news and stock prices. By combining NER-driven healthcare sentiment analysis with deep learning-based stock predictions, this project aims to generate trend line predictions and offer clearer insights into how events impact stocks in the healthcare sector.

2 Related Works

FinBERT is a BERT-based NLP model trained on financial texts to classify sentiment as positive, negative, or neutral, making it useful for stock market analysis(**?**). By analyzing financial news and reports, FinBERT helps assess market sentiment and predict stock price movements.

BioBERT, on the other hand, is designed for biomedical NLP tasks and trained on large-scale biomedical corpora like PubMed abstracts and clinical (**?**)notes. Unlike FinBERT, BioBERT does not perform sentiment analysis but excels at tasks

like Named Entity Recognition (NER) and relation extraction, making it useful for identifying healthcare-related entities such as drug names and diseases.

Since our model will be trained on financial news, we compare its performance in healthcare sentiment analysis against FinBERT. While FinBERT is designed for financial text, it is not specifically trained on healthcare-related financial news. By fine-tuning our model on healthcare-focused financial news, we aim to improve sentiment classification and better capture the impact of medical events on stock prices.

3 Methodology

We utilize the Yahoo Finance API to obtain news articles and stock prices for our dataset. The data will then be processed using our Named Entity Recognition (NER) model to extract healthcare-related news and relevant keywords, such as company names, drug names, healthcare policies and others. This step ensures that only news relevant to the healthcare sector is considered for sentiment analysis.

Once the healthcare-related news articles are extracted, we fine-tune an existing BERT-based model from Hugging Face to classify sentiment as positive, neutral, or negative. This fine-tuned model will be compared against FinBERT, to evaluate its effectiveness in analyzing healthcare-related financial news. The sentiment scores from both models will then be used as inputs for a Long Short-Term Memory (LSTM) model, which will predict stock price movements based on sentiment trends.

To assess the impact of sentiment on stock movements, we will compare the performance of two models: Our fine-tuned BERT model + LSTM versus FinBERT + LSTM. The predicted stock prices will be evaluated by analyzing their correlation with actual stock price movements in the healthcare sector. This comparison will help determine whether a sentiment model trained on financial news can provide more accurate stock price predictions than FinBERT.

To evaluate the performance of our fine-tuned BERT model, we will use classification metrics such as accuracy, precision, recall, F1-score, and the confusion matrix on the sentiment scores. These metrics will provide a comprehensive evaluation by measuring overall correctness, balance between false positives and false negatives, and the
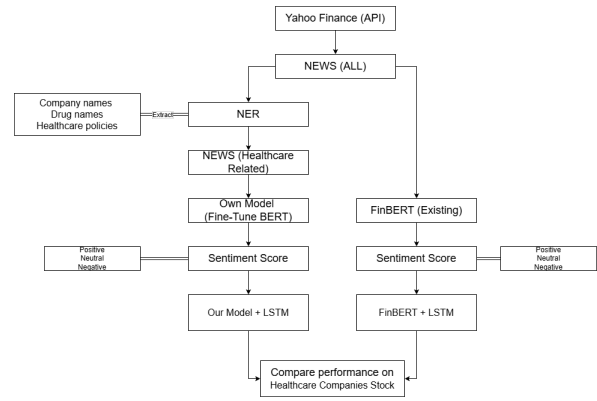

Figure 1: Architecture

model's ability to distinguish between sentiment classes effectively.

For stock price prediction, we will evaluate our models using Mean Absolute Error (MAE) and R-Squared (R^2) Score. MAE measures the average absolute difference between predicted and actual stock prices, providing an intuitive understanding of prediction accuracy. The R^2 score indicates how well the model explains the variance in stock prices, with values closer to 1 representing better predictive performance. These metrics will allow us to compare the effectiveness of our fine-tuned BERT model + LSTM against FinBERT + LSTM in forecasting healthcare stock movements.

4 Project Timeline

- **Feb 24 – Feb 26:** Set up Yahoo Finance API, collect initial dataset (news + stock prices).

- **Feb 27 – Mar 3:** Implement NER model to extract healthcare-related news and keywords.

- **Mar 4 – Mar 10:** Fine-tune BERT-based model for sentiment classification.

- **Mar 11 – Mar 17:** Compare fine-tuned model with FinBERT on sentiment classification tasks.

- **Mar 18 – Mar 24:** Implement LSTM model for stock price prediction using sentiment scores.

- **Mar 25 – Apr 2:** Evaluate model performance, analyze stock price correlations.

- **Apr 3 – Apr 10:** Finalize report, refine results, and prepare for submission.

All tasks will be split equally and done together.

5 References

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.