FINANCIALBERT: PREDICTING STOCK PRICE WITH BERT AND SEMANTIC ANALYSIS

INDEPENDENT RESEARCH PAPER

Hokyung Lee
Department of Computer Science
University of Waterloo
Waterloo, ON
a362lee@uwaterloo.ca

December 9, 2023

ABSTRACT

Traditionally, stock price models have been built using time-series analysis, where models are trained on past stock price data to predict future prices. Recently, however, the advent of Transformer-based language models has made it feasible to integrate a critical component into stock price prediction models: the analysis of financial news, which consists largely of text information. In this context, I developed FinancialBERT, a model that utilizes the text-encoding capabilities of the BERT model to interpret the semantic meanings of financial texts. This model is then trained on a regression model to predict the upcoming month's stock prices for thousands of companies.

1 Introduction

The recent surge in advanced language models, capable of encoding and generating text with high contextual awareness, has been noteworthy. BERT, a model in this category, encodes text information into embeddings, representing the semantic meaning of the text [Devlin et al., 2018].

Stock price prediction models traditionally fall within the time-series analysis framework, using various approaches suitable for financial data. Classical statistical models, such as ARIMA and GARCH, are adept at handling numeric market patterns but lack the ability to incorporate textual information like news headlines or financial reports.

This gap highlights the need for new models that analyze textual information to predict stock market trends. The recent progress in language models presents an opportunity to experiment with these models in analyzing financial news data for stock price prediction using semantic analysis. The models were trained as follows:

Datasets A dataset from Kaggle with extensive financial news headlines from numerous companies was compiled. Additionally, stock price data for each company was sourced using the Polygon.io API service.

Models The FinBERT model from HuggingFace, already fine-tuned for financial data, was chosen for its ability to capture semantic meanings [Arac, 2019]. For prediction, regression models were employed, using FinBERT's output to forecast the stock price change rate from one month to the next.

Evaluations The dataset was split into training and testing segments, with the latter used to assess model accuracy by comparing predicted and actual stock prices from Polygon.io.

2 Training regression models

Three models were trained for stock price prediction:

Model v1 The initial model's training data encompassed five distinct financial news headlines per month for each company. These headlines were concatenated and inputted into the FinBERT model. In cases where the training data

exceeded five headlines for a given month, we grouped them into sets of five and processed each set separately through the FinBERT model, thereby maximizing the use of available data. Subsequently, the regression model utilized the raw output embedding from the FinBERT model to produce a single numerical value, indicative of the stock price rate change from the current month to the following month.

Model v2 The second model underwent a training process akin to the first model, with a notable deviation: instead of clustering five headlines together monthly, each headline was transformed into an embedding. These embeddings were then compared against a search query to identify the top-5 headlines most pertinent to stock prices and the financial market. This approach ensured the integration of only the most relevant training data.

Model Top 5k The third model utilized a dataset selection approach akin to the second model, where the top five headlines were chosen for each month. In contrast to the second model, where headlines were concatenated into a single string before feeding into the FinBERT model, this model processed each headline individually through the FinBERT model, which included a layer for positive/negative/neutral semantic classification. The outputs comprised five vectors, each of length three, representing the proportion of positive, negative, or neutral sentiment in the text. These vectors were then concatenated to train the regression model. By receiving a more concise input compared to the raw output embedding, the regression model was notably smaller.

Each model generated a forecast for the upcoming month's stock price change rate, which, when combined with the current stock price, provided a prediction for the subsequent month's price. This predicted price was then compared with the actual price to calculate the loss value. During the training phase, these loss values were used in the backpropagation process, and in the testing phase, they functioned as key metrics for evaluating model accuracy. The design of the training loop is outlined as follows:

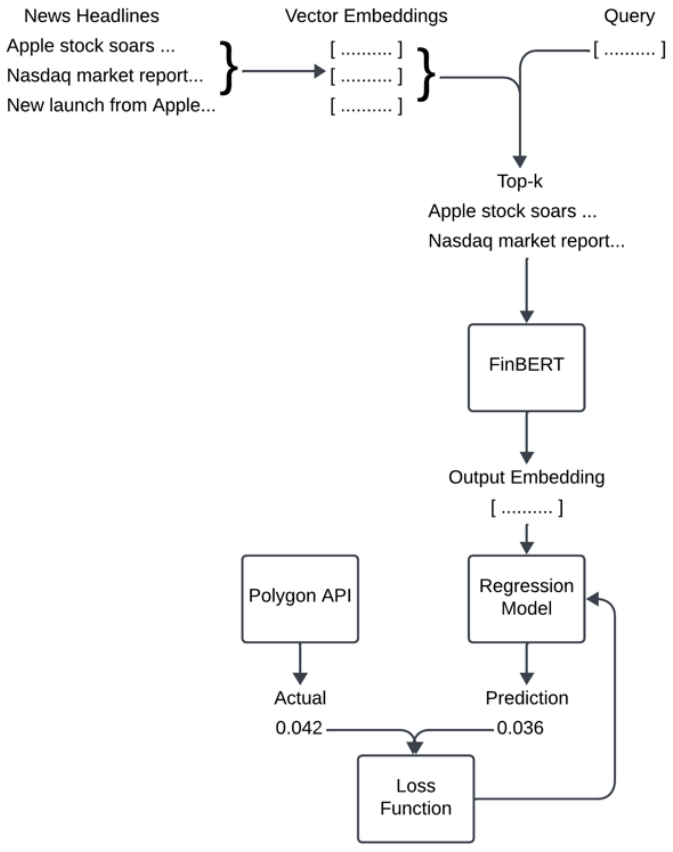

Figure 1: The training loop for the models.

3 Analyzing model accuracies

The performance of these models was compared and graphed:

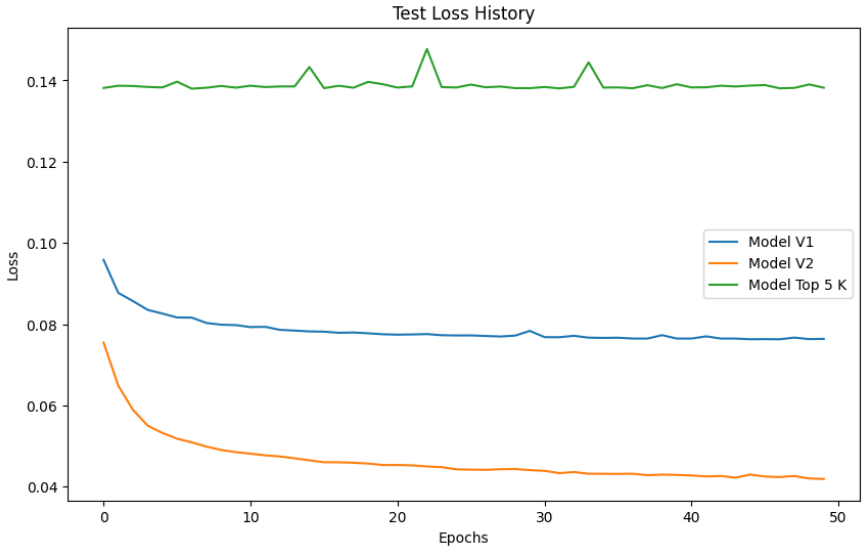

Figure 2: The comparison between the loss history of the models.

Model v2 consistently outperformed Model v1 across all training epochs. Model Top 5k struggled to converge, likely due to limited contextual information from the final probability predictions of the FinBERT model. Ultimately, Model v2 achieved an impressive ±4% accuracy compared to actual stock prices, while Model v1 achieved ±8% accuracy.

4 Designing RAG system for stock price prediction

A RAG (Retrieval-Augmented Generation) system, which retrieves relevant information before generating a response, was also developed for predicting future stock prices using real-world financial data [Lewis et al., 2020]. The architecture of the system is as follows:

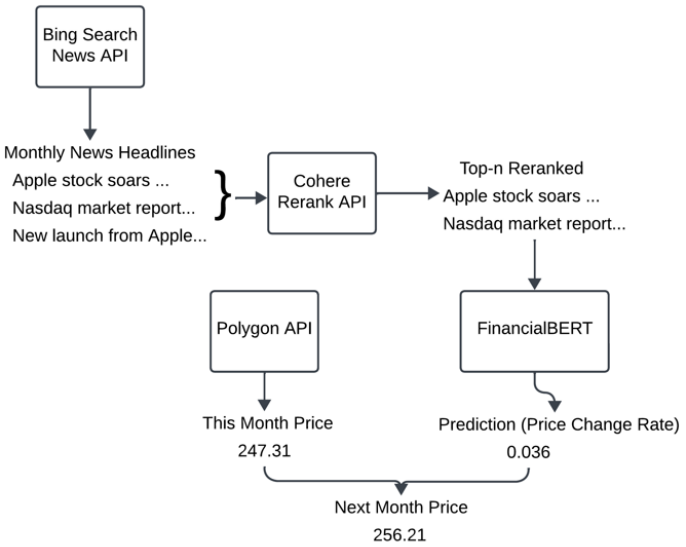

Figure 3: The design for stock prediction system.

Recent financial news was sourced using the Bing Search News API and processed through the Cohere Rerank API to select the top 5 relevant headlines. Model v2, with its superior performance, was employed. It analyzed the concatenated headlines via the FinBERT model to predict the monthly stock price change rate. This predicted rate, applied to the current month's average stock price from Polygon.io, estimated the next month's stock price. The system includes a user interface for displaying the stock price predictions for selected companies.

5 Conclusion

This paper introduces a new paradigm for predicting stock prices, leveraging recent advances in language models to analyze financial textual data. While the model shows some inaccuracy, not yet suitable for production, it paves the way for future NLP models to more accurately predict the stock market by incorporating the breadth of textual data.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*, 2020.

Appendix

A Demonstration

A.1 Tesla Example

Cohere Reranked Headlines

1. It seems the most impactful event occurred with a major Tesla peer. China's Nio reported its third-quarter results, revealing still robust (46% year-over-year) growth in EV sales. Net losses – both according to GAAP and non-GAAP (adjusted) standards – were deeper than they were in the year-ago period, but not worryingly so.
2. A futuristic new EV truck is not enough of an innovation or financial incentive to base an investment decision on TSLA stock.More From InvestorPlace ChatGPT IPO Could Shock the World, Make This Move Before the Announcement Musk's "Project Omega" May Be Set to Mint New Millionaires.
3. The long-awaited launch of the Cybertruck failed to boost Tesla stock after pricing for the electric vehicle came in much higher than initially promised.
4. This time has also come to mean something for Tesla investors: wondering about Elon Musk's potential sales of Tesla stock.
5. Tesla Cybertruck order holders disappointed by the high prices can look forward to a more affordable Tesla model coming soon.

Stock Price History and Prediction

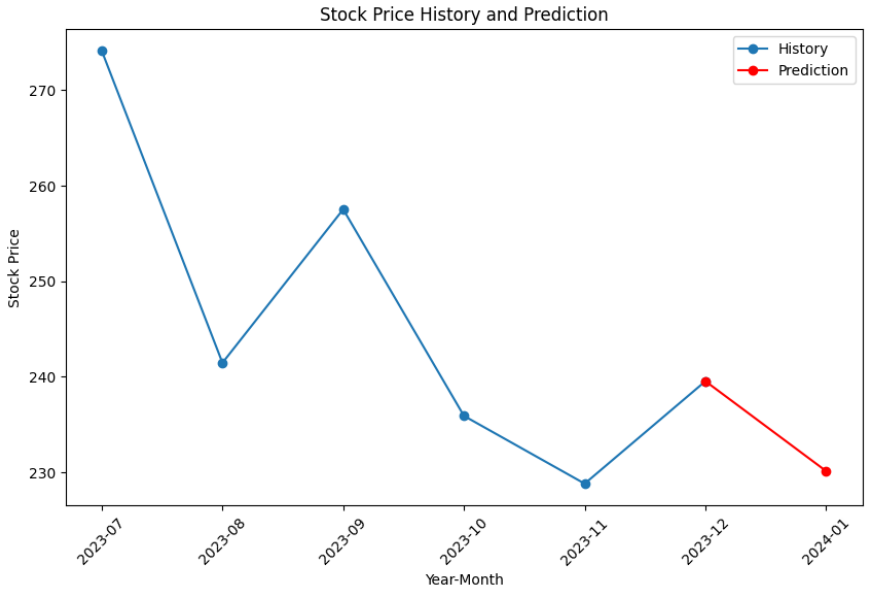

Figure 4: Tesla Stock Price History and Prediction for Jan 2024.

A.2 Apple Example

Cohere Reranked Headlines

1. New data from a major Apple supplier and Wall Street analysts pushed back against bearish calls around softening demand for the tech giant's devices and services. It's no wonder Apple (AAPL) shares have clawed their way back in recent weeks toward record highs.
2. In the world of finance, certain stocks can captivate investors and drive ... that have had a significant impact on market returns. These stocks include Apple Inc., Microsoft Corp., Amazon.com Inc., Nvidia Corp., Alphabet Inc., Tesla Inc., and Meta ...

3. No significant news for in the past two years. Key Stock Data P/E Ratio (TTM) The Price to Earnings (P/E) ratio, a key valuation measure, is calculated by dividing the stock most recent closing ...
4. Despite a turbulent market, the shares of Apple Inc ... Apple's stock has shown a commendable resilience, marking its second consecutive day of gains. Although it closed $6.99 below its 52-week high on July 19th, the company's stock price remains ...
5. The main news out on Apple in the early part of the month was its fourth-quarter earnings report. It showed solid numbers, topping estimates on the top and bottom lines, but the stock actually pulled back slightly on the news, falling 0.6% on Nov. 3 after two straight days of strong gains to open the month.

Stock Price History and Prediction

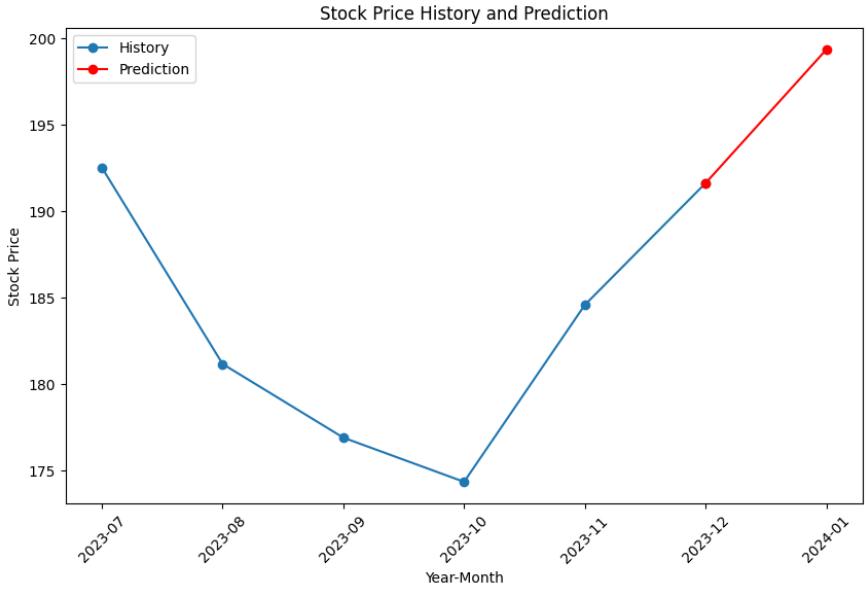

Figure 5: Apple Stock Price History and Prediction for Jan 2024.