Computer Science
&
Data Science

Capstone Report - Fall 2024

Parameter-Efficient Fine Tuning of LLaMA for Sentiment Analysis for Enhanced Stock Price Prediction

Henry Zhang and Alec Sirkin

supervised by
Li Guo

Preface

Predicting stock prices has been a challenge for investors and analysts for as long as equities, derivatives, and other financial assets have been traded. With the recent uptick in research and development of large language models (LLMs), our aim was to see if we could utilize LLMs and machine learning to conduct sentiment analysis and gain additional insights into stock price movements.

As undergraduate CS majors with interests and work experience with GenAI and software engineering, as well as shared interests in finance, we aim to provide methods and in-depth analysis for fine tuning LLMs for domain specific sentiment analysis, and an implementation of said sentiment for stock price prediction.

This project is aimed at financial analysts, AI researchers, or anyone interested in the intersection between financial markets and AI. By fine-tuning a model and implementing an LSTM for predictions, we hope to give an indicator of performance with LLMs in sentiment analysis, as well as demonstrate the potential for AI to help with enhanced decision making with news sentiment and machine learning.

This project is not only a reflection of our academic journey but also a step toward bridging the gap between research and a concrete actionable product. We hope to provide insights, techniques, and findings for Parameter-Efficient Fine Tuning of LLaMA for Sentiment Analysis for Enhanced Stock Price Prediction in this report, and hope to expand on our findings in the future.

Acknowledgements

Special thanks to Li Guo, our faculty supervisor, for her close guidance and feedback throughout the entire semester. Additionally, the Machine Learning class taken in tandem with our capstone taught by her provided us with the foundational knowledge and practical skills to enhance our work. Her knowledge, effort, and care supervising our project helped shape the direction of our project and this would not have been possible without her.

Abstract

Predicting stock price movement using sentiment analysis is a challenge as it requires accurately understanding and interpreting sentiment of vast amounts of data with domain specific jargon, and being able to utilize said sentiment for accurate prediction. This is important because financial news often reflects events relevant to the market and investor behavior, which then adds another layer of information that can be used when predicting price movements. Recent research has highlighted the success of LLMs in applications of domain specific tasks, but fully fine-tuned LLMs for domain specific tasks are impractical for smaller projects due to computational load. Our work highlights the effectiveness of using QLoRA to fine-tune LLMs like LLaMA 3.1 for financial sentiment analysis, offering a more computationally efficient approach to domain-specific fine-tuning while still retaining performance improvements found with fine-tuning. Through QLoRA fine-tuning, we achieve a notably improved accuracy rate in classifying sentiment in financial news which is competitive against the performance of existing models fine-tuned models. Furthermore, we investigate the impact of incorporating sentiment data into stock price prediction. By adding sentiment as an additional input layer in an LSTM model, we observe improved accuracy in predicting stock price movements for certain tickers, demonstrating the value of sentiment analysis in enhancing stock price predictions.

Keywords

QLoRA; Sentiment Analysis; Fine-tune; PEFT; LLaMA; LSTM; Time series; Price prediction

Contents

1 Introduction

Context

In today's interconnected world, understanding the impact of public sentiment on financial markets has become increasingly critical. News articles and financial reports often contain valuable insights and trends that can influence stock prices. Sentiment analysis—a process of determining the emotional tone behind a series of words—can provide investors with a new dimension of decision making by analyzing how different types of text content affect market behavior. Traditional sentiment analysis methods have been effective to a degree, but advancements in large language models (LLMs) and fine-tuning techniques offer the potential to greatly enhance accuracy and depth.

Sentiment analysis can be particularly challenging due to the diversity in language used across different platforms. News articles often employ formal and structured language, while financial reports contain technical jargon and quantitative data. Moreover, the dynamic nature of market reactions means that sentiment about specific stocks must be interpreted in the context of it's company and sector. Domain specific understanding of text is necessary for accurate sentiment analysis related to stock prices.

Recent research has demonstrated that LLMs can capture nuanced meanings and emotional subtleties in text much more effectively than previous approaches. Additionally, deep learning techniques such as Long Short-Term Memory (LSTM) networks provide methods for modeling sequential data, such as stock price time series, with the ability to incorporate additional data such as financial news sentiment for improved price movement predictions. By fine-tuning LLaMA 3.1 for domain-specific sentiment analysis on financial texts, and integrating sentiment predictions into an LSTM framework to predict stock price movements, we aim to determine and demonstrate methods that can more accurately interpret and correlate sentiment with stock price movements.

Objectives

The primary objective of this capstone is to demonstrate the viability of improved financial forecasting by using LLMs for sentiment analysis and the sentiment found into our model for price prediction. To do this, our project is split into two objectives to achieve our goal.

The first objective is conduct and improve sentiment analysis on financial text and news article headlines. To do this, we aim to use LLaMA 3.1, the most recent open source model by Meta,

and prompt it to give a sentiment score based on relevant financial texts. We want to benchmark performance with the base model, then improve sentiment analysis through different methods, such as few shot learning and adjusting parameters in the model. We then will train the model and fine tune it using QLoRA on domain specific human labeled financial headlines, and evaluate the performance improvement in the models ability to accurately classify sentiment in a financial context.

After achieving our first objective of fine tuning our model for improved sentiment analysis, our second objective is to incorporate the evaluated relevant sentiment into a forcasting model. We will integrate the assigned aggregate sentiment by day as input features into an LSTM framework, which will model our stock price time-series data. The objective here is to enhance the base LSTM model's ability to predict stock price movements by providing it an additional layer of news sentiment insights, which can offer additional information beyond just numerical closing prices for stocks. By doing so, we can also determine if news sentiment is a leading indicator of stock price movement, and determine if there is correlation between market sentiment as expressed in financial news headlines and the price of traded assets.

Our objectives will follow these steps:

1. Data Collection: Retrieve data from relevant sources (Financial News articles, financial statements)
    a) Real Prices and corresponding news sources
        i. EODHD Financial Data API
        ii. Yahoo Finance web-scrapping Python package
    b) Labeled Datasets:
        i. Sentiment Analysis for Financial News - Kaggle
        ii. Aspect based Sentiment Analysis for Financial News - Kaggle
2. Perform sentiment analysis and predictions using base LLaMA 3.1 and calculate relevant evaluation metrics
3. Using Labeled Datasets, prepare data and fine tune using QLoRA on LLaMA 3.1. Compare the accuracy between the base LLaMA 3.1 model and the fine-tuned LLaMA 3.1 model specifically tuned on the financial data.

4. Implement LSTM on time-series data stock prices. Incorporate sentiment scores as an additional input feature and evaluate performance using metrics such as RSME.

Data collection, training, and the main application will all be accomplished using python (incl: pytorch, numpy, pandas), Meta LLaMA 3.1, NYU HPC cluster for training the fine-tuned model, and Yahoo finance for assessment and grading.

As part of the project, we will also identify the limitations of our approach and highlight potential areas for future research, such as investigating alternative models or architectures that could further improve prediction accuracy and robustness. Additionally, we will evaluate the scalability of our approach for real-time prediction systems in financial markets.

2 Related Work

Researchers in various disciplines have attempted to solve the question of predictability of the stock market by gauging public opinion since at least the mid-20^{th} century. Understanding public sentiment adds another dimension to data investors collect to gain a competitive edge and make more informed decisions when determining to buy, sell , and trade financial assets. There is a proven strong relationship between stock market values and public sentiment, so it is possible to suggest that public sentiment could directly influence stock performance[1]. A study by Padhanarth et al. corroborates this, arguing that investor confidence affects the demand of individuals when purchasing stocks during peak trading hours [2]. In our current digital age, opinions of investors, influential figures, and the general public are not only easier to access, but also carry a much more widespread influence, and therefore a more significant impact/weight on financial markets.

Understanding public opinion and investor confidence revolves around Natural Language Processing and specifically sentiment analysis. Sentiment analysis has evolved and improved leaps and bounds from rudimentary rule-based systems to complex machine learning and large language models. These large language models have revolutionized sentiment analysis due to their ability to understand context in natural language [3], as well their ability to be utilized in many domains due to its generalized knowledge. Due to these advances in LLMs, the potential to refine sentiment analysis in specifically the financial domain is significant, promising to enhance decision-making processes in the face of rapidly changing market conditions.

Current studies in the field of sentiment analysis on financial data utilize many different meth-

ods to perform their sentiment analysis, such as support vector machines, naive Bayes classifiers, deep learning, and pre-trained models. Our focus is on pre-trained models, specifically LLaMA 3.1, released in 2024. The objective is to conduct sentiment analysis on investor sentiment from financial news headlines and articles to predict market movements and make buy, hold, or sell decisions.

Our first aim is to provide insight into how using LLaMA 3.1 to perform sentiment analysis to predict stock price movement compares to existing methods mentioned above, as well as other general pre-trained LLMs, such as older versions of LLaMA, Bert, and GPT models. This will involve in-depth research into evaluation metrics used in other studies, for the best comparison. Additionally, we will evaluate the performance increase when a base pre-trained model is trained on domain-specific data using existing research on fine tuning methods. This will provide insight into how LLaMA 3.1 can be optimized for specific domain use cases and the performance improvements that can be gained from doing so.

2.1 Overview of Public sentiment in Finance

It was not until the recent rise in popularity of peer-to-peer social media platforms and advancements in the field of Natural Language Processing in the early 2000s that researchers had seriously considered using shared public sentiment as an important variable in the equation that could solve stock market pricing [4]. In the decade following the founding of Twitter, the popular peer-to-peer social media platform that allows users—including high profile individuals such as politicians and CEOs—to broadcast short messages to their followers, a number of studies [1][5] sought to explore the potential link between sentiment found in tweets regarding certain companies, and the movement of said companies' stock market prices.

Earlier studies, such as [5], found that for a few selected publicly listed companies included in the S&P 500 index, strong correlations were observed between price movement and the change in aggregate sentiment scraped from various Twitter accounts mentioning said companies. They then applied an adapted support vector machine to assign the tweets a binary classification of negative or positive. In the end, their findings were inconsistent and found extremely weak or nonexistent correlation coefficients for the majority of companies studied. However, [5] lacked more modern, accurate means of analyzing natural language sentiment, and employed a polarizing positive-negative classifier, a tactic that newer studies, such as [4], warn against.

In more recent studies, namely [4] and [6], researchers not only found consistently stronger

correlations between aggregate sentiment and prices, but also shifted the focus of their research to which forms of media correlate the strongest with price movements, and importantly, whether sentiment is the independent variable, and price is the dependent variable, or vice versa. In order to judge whether one variable was leading the other and whether any correlation existed at all between either Twitter or traditional news sentiment and stock market prices of various companies, [4] used the NLTK library's VADER sentiment analysis function and a time-lagged cross correlation in the range of +/- 5 days and collected the 11 total correlation values for each comparison. Their findings indicated that most of the studied companies had some significant correlation with either Twitter sentiment or financial news headline sentiment (usually not both), but that the directions of said correlations varied widely between the companies. One glaring weakness in [4] is the use of VADER, which, while trained on social media content, is not domain specific and is not trained on a financial corpus.

It should also be noted that both [4], [6] were published during the COVID-19 pandemic and both tested their hypotheses during a period of considerable economic turmoil.

What is missing from the existent body of research, is rigorous study of how well novel Natural Language Processing technologies can generate more locally accurate sentiment analysis to specific companies that could potentially establish stronger correlations between their sentiment scores and price movements. In the year 2024, the abilities of large language models to generate accurate and meaningful sentiment scores has become a prominent subject of research in stock market price movement prediction, and has been met with increasing interest from industry practitioners [7][8].

2.2 Methods to assign and improve sentiment

Traditionally, sentiment analysis has been conducted as seen in [4],[6] where support vector machines, other supervised learning approaches, or simple lexicon look-ups are used as a polarizing classifier of positive or negative sentiment. Additionally, these classification methods tend to make use of more generalized corpuses, rather than domain specific ones which can provide more accurate accurate, and less skewed results. [7] observed that the sentiment scores generated by OPT, BERT, FinBERT, LLaMA 3 and RoBERTa LLMs on their dataset formed a normal distribution around a median of 0.5 and similar standard deviations, whereas a simple non-machine learning sentiment score assignment following the Loughran-McDonald Dictionary, a prominently used lexicon for finance-specific terms and their sentiments, resulted in a median score of 0.68, noticeably skewered towards positive scoring compared to LLM evaluations, signifying that older

methods are not as accurate.

Several contemporary studies have highlighted the advantages that machine learning approaches, and particularly those which utilize NLP models, have over traditional LLM methods for assigning meaningful sentiment scores [7][8]. For example, [9] , a meta-study on investor sentiment scoring and aggregation methods, found that among studies which calculated investor sentiment scores using traditional methods, considerable error and poor performance could arise if an unsuitable algorithm was used. Financial texts, as with many other domain-specific texts, have jargon, implicit meanings, and occasional sentiment shifts. One of the greatest potential advantages that LLMs provide is their ability to adapt to changing sentiment and learn without needing an explicit, static lexicon for domain-specific buzz words, by applying fine tuning techniques using curated or available datasets.

Recent studies on improving performance of large language models (LLMs) have shown significant promise for enhancing their adaptability and applicability in domain specific applications such as finance. Techniques to optimize LLM performance include fine-tuning and in-context learning(ICL). Of these two, ICL incurs more computational and memory costs, because all examples in prompts must be processed every time a new prediction is made. Fine-tuning can not only reduces computational costs, but also has been found to significantly outperform few-shot in-context learning (ICL) in classification tasks with GPT-3 [10].

Traditionally, models are fine tuned fully. However with increasing model size, such as our LLaMA 3.1 model , fine tuning can be computationally challenging. Hence, parameter efficient methods like Low Rank Adaptation of LLM's(LoRA), where pretrained weights are fixed and additional new parameters are added, were devised and found to have no significant difference in actual performance [11]. There are many improvements and advancements in how to conduct LoRA paremeter efficient fine tuning(PEFT), such as GLoRa[12]. , QLoRA[13], and (IA)3, to name a few. For our use case, we will be implementing Quantized Low Rank Adaptiation(QLoRA) for fine tuning as outlined in [13] , as it is found that fine tuning on small but high quality datasets with QLoRA can replicate the effects of full finetuing, producing quality results while using smaller consumer level GPUs.

In the financial domain specifically, recent studies have focused on tailoring LLMs for sentiment analysis. For instance, enhancements made to the LLaMA 2 7b-hf model resulted in a remarkable increase in accuracy when grading financial sentiments, climbing from 37.3% to 89% after implementing PEFT using LoRa techniques and the Simple Fine-tuning Trainer (SFTTrainer)

from huggingface to the base model [8]. This transformation emphasizes the importance of model adaptations tailored to the unique lexicon and subtleties of financial texts. The integration of Parameter-Efficient Fine-Tuning (PEFT) methods further solidified the model's effectiveness, particularly in recognizing negative and positive sentiments.

Moreover, the ongoing exploration of incorporating real-time data feeds and multimodal data sources reflects a growing recognition of the need for dynamic and comprehensive sentiment analysis frameworks in finance. Future research directions include examining contextual and temporal factors that influence market sentiments, as well as broadening the model's applicability through cross-lingual adaptations. These developments not only highlight the transformative potential of fine-tuning LLMs for financial applications but also set the stage for proactive sentiment analysis tools that could redefine investment strategies.

Our review of existing literature has yielded a promising opportunity to engage in research regarding how new machine learning technologies can be used to fill in the gaps where less robust methods of sentiment scoring and aggregation have failed or not prevailed as well in prior research.

Existing research seems to highlight some correlations between investor sentiment and overall market movements, and even the movement of individual stock market prices. However, traditional approaches to classification appear to have limited the predictive capabilities compared to cutting-edge NLP technologies in 2024. LLaMA 2 models have already demonstrated their sentiment analysis capabilities in prior studies, which makes low rank adaptations on the recently released LLaMA 3.1 for finance domain-specific sentiment analysis as a parameter-efficient fine-tuning method a promising subject for further research. Despite few published works regarding fine-tuning LLaMA 3.1 for sentiment analysis tasks, to our knowledge, there is little or no research into parameter-efficient fine-tuning of LLaMA 3.1 specifically for sentiment analysis of specific financial news for a stock ticker for the prediction of absolute stock price movement.

3 Solution

In this section, our implementation and decision making for data preparation, improvement and evaluation of sentiment prediction with LLaMA, real-time data workflow and architecture, and model choice and considerations are explained here.

3.1 Sentiment analysis with Base model

We selected Meta's LLaMA 3.1 8b Instruct large language model, the most recent cutting edge model released in 2024, for its open source nature, which provides us with more flexibility in fine tuning. The 8B parameter model was chosen as it is suitable for efficient deployment on consumer sized GPUs, with the option to scale to larger LLaMA 70B or 405B models in future implementations. In order to test the base performance of sentiment assignment for financial news headlines, we utilized a publicly available labeled dataset on financial news categorized by sentiment. Financial-Phrasebank consists of 4840 rows of financial news sentences on different industries, company sizes, and news sources, annotated by 5-8 human annotators. Data is labeled into positive, negative, and neutral.

To test the baseline performance of the model, we used a random sample of 1000 of the dataset and generated predictions for sentiment with the prompt:

```
"Classify the sentiment of this headline as one of: positive, neutral, or negative.
Respond with only one word.
{ Headline }".
```

Listing 1: Prompt for classifying sentiment.

We then parsed the one word response from the LLM, and evaluated it against the labeled values to calculate our metrics. Some of the arguments when using the LLM such as temperature, tokens, were tweaked to produce the best results.

To extract the most performance out of the baseline model, we also experimented with in-context learning, testing the results from zero-shot prompting to few shot prompting. This was done by modifying the prompt to provide definitions for positive, neutral and negative labels, as well as adjusting the number of labeled examples given in the prompt, with HTML tags as a practice to make sections of the prompt clear to the LLM. We tested this modified prompt from one shot learning all the way to ten shot learning, meaning one to ten examples were provided in the prompt when evaluating performance

"Analyze the sentiment of the following financial news headline to determine its impact on the company's performance. You should respond with a classification of the sentiment as one word: **positive**, **neutral**, or **negative**.

Headline: {Headline}

Sentiment: [answer]

<Explanation> - **Positive** indicates favorable news that suggests growth, profit, or success.

- **Neutral** indicates news that doesn't have a clear positive or negative impact.

- **Negative** indicates unfavorable news that suggests losses, layoffs, or other adverse effects.

</Explanation>

<Examples>

1. Headline: 'According to Gran, the company has no plans to move all production to Russia, although that is where the company is growing.'

Sentiment: neutral

2. . . .

</Examples> "

Listing 2: Improved prompt with in context learning

3.2 Fine-tuning model

With a thorough baseline model testing above, we then focus on our objective of enhancing performance through fine-tuning using QLoRA. Implementation of this is split into two steps: Data preperation and applying QLoRA.

We use the labeled dataset Financial-Phrasebank, a CSV consisting of 2 columns, Financial news text and its corresponding sentiment label. Of the 4840 unique rows, we split the data into a training, evaluation, and testing set using an 80-10-10 split. Since we are using the LLaMA Instruct model, a chat model trained on a prompt question and answer format, we will need to convert the tabluar data into the same text format that LLaMA 3.1 Instruct was trained on. The format shown in listing 3 is based on the official Meta prompt template documentation for LLaMA 3.1, with row["headline] and row["sentiment] injected into the text from our CSV data.

"<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are a financial analyst assistant dealing with financial news <|eot_id|>

<|start_header_id|>user<|end_header_id|>

"Classify the sentiment of this headline as one word: positive, neutral, or negative.

Headline: **row["headline"]**

Sentiment:" <|eot_id|>

<|start_header_id|>assistant<|end_header_id|>

Sentiment: **row["sentiment"]**<|eot_id|>"

Listing 3: Format for training data

When conducting QLoRA, we used SFTTrainer from trl, a python package that is a wrapper around the transformers.Trainer class, and bitsandbytes, a wrapper for quantization. With this we can implement QLoRA, and easily configure hyper-parameters such as the learning rate, rank, number of epochs, etc. For our implementation, we used the recommended hyper-parameters from the QLoRA paper setup details.[13]. We did configure and test different epochs, and charted training and evaluation loss using Weights & Biases.

3.3 Incorporating real data

After fine-tuning and evaluating our LLM for sentiment analysis using the labeled dataset, the next step involves integrating the model's sentiment predictions with stock price data. To achieve this, we require historical financial price data for specific stock tickers, as well as corresponding financial news headlines for sentiment analysis.

We leverage **yfinance**, a web-scraper API capable of fetching comprehensive historical and current financial data for a given ticker in the form of a dataframe. Specifically, we extract the daily closing prices for the selected tickers. For financial news headlines, we use the **EODHD Financial News API**, which retrieves relevant news for specified tickers within a given time frame, returning the data in JSON format. We choose to fetch two years worth of data, from 2022-10-01 to 2024-10-01, which consists of over 7,000 financial news headlines and 500+ days of sequential closing prices for a given stock ticker.

To combine the sentiment of financial news articles by date, we assign sentiment scores—Positive, Negative, and Neutral—to each of the 7,000+ headlines in our dataset. To facilitate aggregation, these sentiments are mapped to numerical values: Positive as 1, Neutral as 0, and Negative as -1. This allows us to calculate the average sentiment score for each date by aggregating these numer-

ical values across all headlines for that day. The resulting aggregate sentiment score is then used as an input feature in the LSTM model discussed later. Next, we join the aggregate sentiment scores with the daily closing price data from **yfinance** using the date as the key. This creates a unified dataset to operate on with both sentiment and price information. Figure 1 illustrates the complete data pipeline.

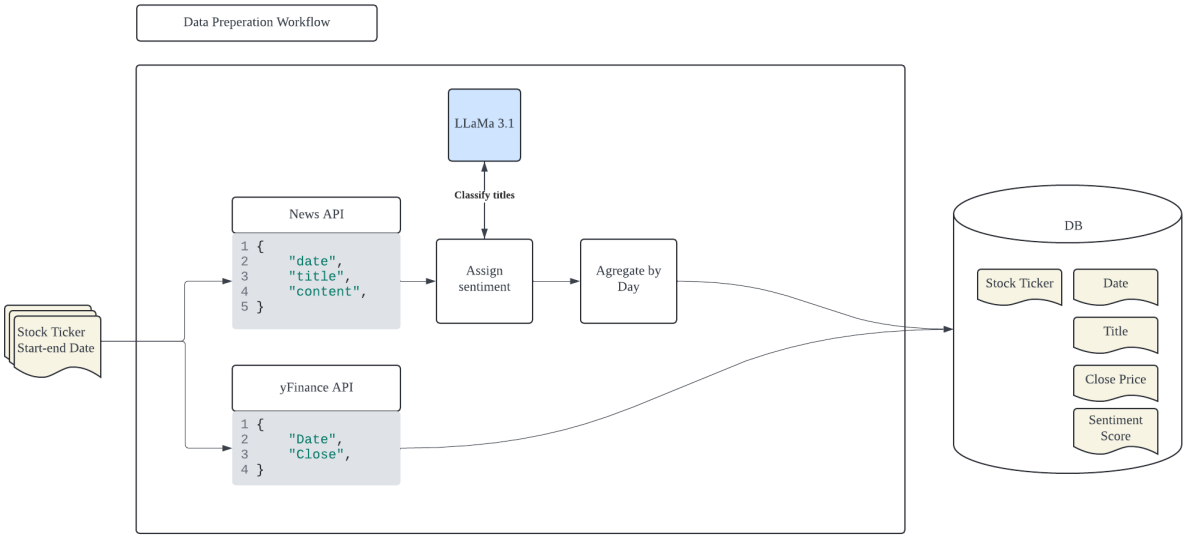


Figure 1: Architecture of our data preparation workflow

Before implementing our LSTM with our daily financial news sentiment and closing prices, we determine if the aggregated sentiment is correlated to the price movement at all. We compute the rolling average of the the sentiment score for the past 28 days to smooth short-term fluctuations and capture long-term trends, plot and compare that to the closing stock prices, and calculate the Pearson correlation coefficient to quantitatively confirm a correlation between sentiment and price.

3.4 Modeling our Data for Price Movement Prediction

Our approach was to use a simple Long-Short Term Memory model with a modest look back window of 60 time steps of trailing inputs to predict outputs at the next time step. We aggregated two years of price data from 10-01-2022 to 10-01-2024 for three publicly traded companies with sufficient price history: Microsoft (MSFT.US), Build Your Dreams (1211.HK), and J.P. Morgan Chase (JPM.US). The base model was comprised of a single input layer, two 32 dimensional hidden layers, a single fully-connected layer which mapped to a single dimensional output. The

price history of each company was cleaned and normalized so that all values were squeezed between [-1, 1]. The architecture is shown in Figure 2a. In the second LSTM, we added an additional input layer for aggregate news sentiment within that given time step, as taken from the mean sentiment assigned to all news articles by the QLoRA fine-tuned LLM with even weight considered for each article. The architecture of the enhanced LSTM model, shown in Figure 2b, differs only in the dimensionality of the input layer. We utilized mean squared error for the loss function and Adam as the optimizer for this regression task.

(a) Base Model

(b) Additional Input Layer

Figure 2: LSTM Model Architectures

4 Results

The base model, trained with few-shot learning, and two of the three fine-tuned adaptations (epochs=1,2) trained on our financial news headlines dataset yielded higher overall accuracy compared with the untrained base model. Moreover, the accuracy on classification of neutral and negative labeled data was markedly higher for the two aforementioned fine-tuned models than for the model trained with few-shot learning and the base model.

The LSTM models which we used as a measure of usefulness for improved sentiment extraction from finance focused news media also showed promising results. The enhanced model with the additional input layer consisting of each trading day's mean sentiment score outperformed the baseline model which relied on only price for predictions. Our research demonstrates a net positive effect of sentiment on accuracy in time-series based stock price prediction.

4.1 Baseline Performance of LLaMA 3.1 8B Model Without Fine-tuning

The base model yielded an overall accuracy rate between 72.1% and 73.4% when tested with varying temperatures (0.1 through 0.8), showing a relatively weak correlation between the temperature setting and the model's ability to provide accurate sentiment, once it had been prompted for a structured response. The exact accuracies across the split of test data are shown in Table 1,

with the best results coming from using a low temperature of 0.1, producing more predictable results.

Temperature	Accuracy
0.1	**0.734**
0.2	0.733
0.5	0.721
0.8	0.733

Table 1: Overall accuracy of base model by temperature setting.

4.2 Performance of Model Trained with Few-Shot Learning

The injection of a limited number of examples into a similarly formatted prompt resulted in noticeably higher classification accuracy for the base model. The accuracy scores of the base model's classification accuracy with few shot learning ranging from 0 to 10 shots is shown in Table 2, with the highest accuracy reported when two examples are provided, and accuracy dropping when too many examples were added. Quality of examples provided was not tested, and this could have an impact.

Examples in prompt	0	1	2	3	4	5	6	10
Accuracy	0.734	0.757	0.762	0.758	0.759	0.758	0.749	0.736

Table 2: Number of examples in prompt vs accuracy for base model

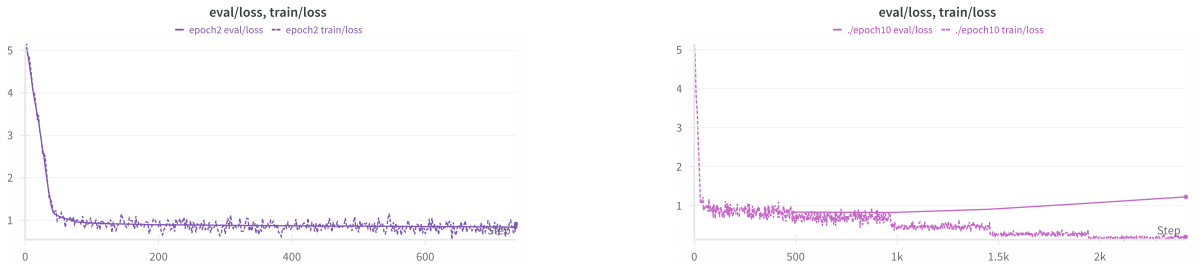
4.3 Performance of The QLoRA Fine-tuned Model Adaptation

The evaluation metrics for all of the fine-tuned adaptations which we trained with our proposed QLoRA method were competitive with, or exceeded existing public financial sentiment models such as BERT models on Hugging-face. While we used the hyper-parameters listed in [13], we tested out different epochs and the effect on model performance. The number of training epochs and according accuracy scores are outlined in Table 3. From an epoch of one to two, performance continued to increase as the training and evaluation loss both continue to decrease seen in 3a. However, with an epoch of 5, as seen in 3b, the model shows clear signs of overfitting, with the rising evaluation loss as the training loss decreases.

As such, the highest overall accuracy, neutral label accuracy, and negative label accuracy were all recorded when the model was trained for two epochs. All trials were done with the learning rate, LoRA rank, LoRA alpha all held constant at $2e^{-4}$, 64, and 16, respectively.

Number training epochs	1	2	5
Overall accuracy	0.849	**0.872**	0.740
Accuracy (Pos.)	0.646	0.701	**0.861**
Accuracy (Neu.)	0.965	**0.968**	0.667
Accuracy (Neg.)	0.786	**0.821**	0.804

Table 3: Number of training epochs vs accuracy scores for fine-tuned model



(a) Epoch of 2

(b) Epoch of 5

Figure 3: Training and Evaluation Loss

Fine-tuning shows an 11.9% increase in overall accuracy compared to few shot learning and presents higher precision and F1 scores overall and on the classification of each label, presented in Table 4 and Table 5. The confusion matrices for few-shot learning with the base model and for our QLoRA fine-tuned approach also indicate that the fine-tuned model misclassified significantly less neutral labels as positive than the best base model with few shot learning. The QLoRA fine-tuned model also misclassified marginally fewer negative labels as neutral and neutral labels as negative than the base model with few-shot learning. The fine-tuned model misclassified more positives as neutrals than did the base model with few-shot learning.

Figure 4.

$$\begin{bmatrix} 93 & 1 & 50 \\ 0 & 44 & 12 \\ 8 & 2 & 275 \end{bmatrix}$$

(a) QLoRA Fine-tuning

$$\begin{bmatrix} 110 & 0 & 34 \\ 0 & 42 & 14 \\ 68 & 4 & 213 \end{bmatrix}$$

(b) Few-shot Learning

Figure 4: Pos./Neg./Neu. Confusion Matrices

Holistically, the model adaptation based on our QLoRA fine-tuning method outperformed the base model and the base model when prompted with Few-Shot learning.

Table 4: Performance Metrics for Few-Shot Learning

	Precision	Recall	F1-Score	Support
Positive	0.62	0.76	0.68	144
Negative	0.91	0.75	0.82	56
Neutral	0.82	0.75	0.78	285
Accuracy	0.75 (485)			
Macro avg	0.78	0.75	0.76	485
Weighted avg	0.77	0.75	0.76	485

Table 5: Performance Metrics for QLoRA PEFT Method

	Precision	Recall	F1-Score	Support
Positive	0.93	0.70	0.80	144
Negative	0.98	0.82	0.89	56
Neutral	0.84	0.97	0.90	285
Accuracy	0.87 (485)			
Macro avg	0.91	0.83	0.86	485
Weighted avg	0.88	0.87	0.87	485

4.4 Base LSTM Model vs. Addition of Sentiment Input Layer

After computing the correlation between the 28-day rolling average of the sentiment and closing price, we find correlation for certain industries and companies. Figure 5 shows the visible correlation between Close Price marked in circles and Sentiment score marked with x. Additionally, we calculated the Pearson correlation coefficient between the two columns, getting a score of 0.742 for MSFT, demonstrating a moderately strong positive relationship between the two variables.

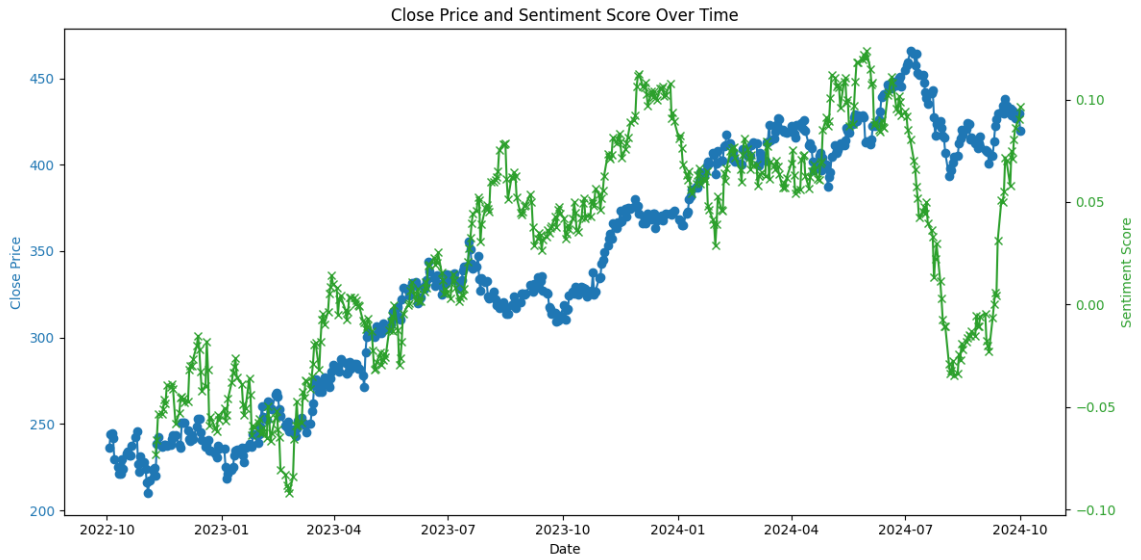
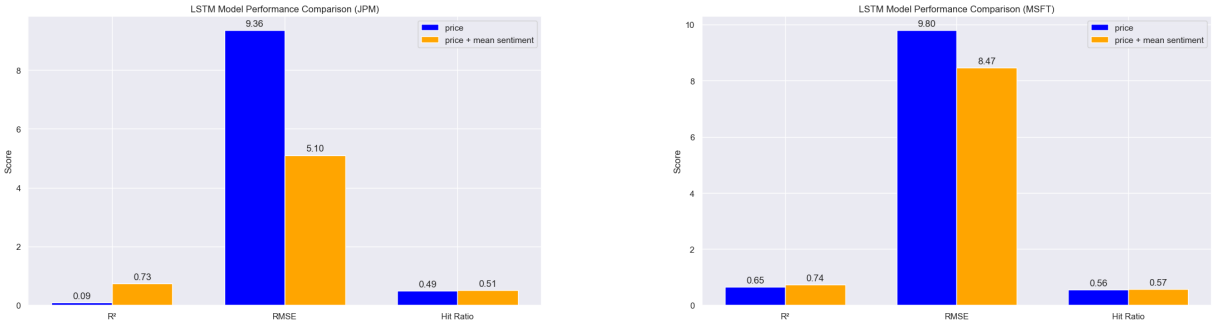


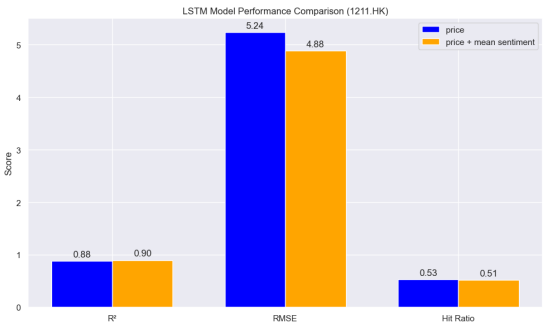
Figure 5: Comparison between 28-day rolling window sentiment average and Closing price for MSFT

The LSTM which had an additional input layer for the mean daily sentiment predicted by the QLoRA fine-tuned model performed better in terms of R^2 score, and root mean squared error for all three companies compared to the baseline model with only price as an input. The hit rate, which tracked the split of binary price movements (increase or decrease) that were accurately predicted, was similar between the models for all three companies. Hit rate showed a marginal improvement for two out of the three companies tested, but did not result in a performance increase in BYD, demonstrating that results are very dependent on company. All three metrics for all three companies are shown in Figure 6. The magnitude of the difference in R^2 scores between the baseline model and the enhanced model in the case of JPM indicates that at the very least, the enhanced model is a better fit for the testing data than the baseline model for specific companies. The R^2 marginally higher for the enhanced model for BYD and MSFT as well.



(a) JPM

(b) MSFT

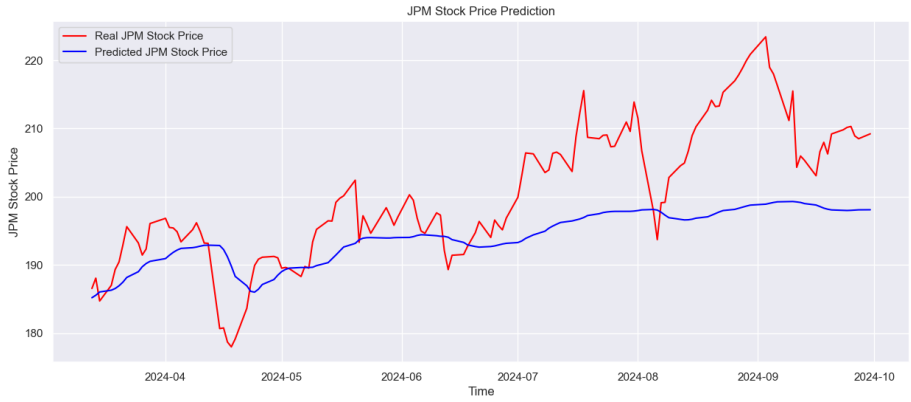


(c) BYD

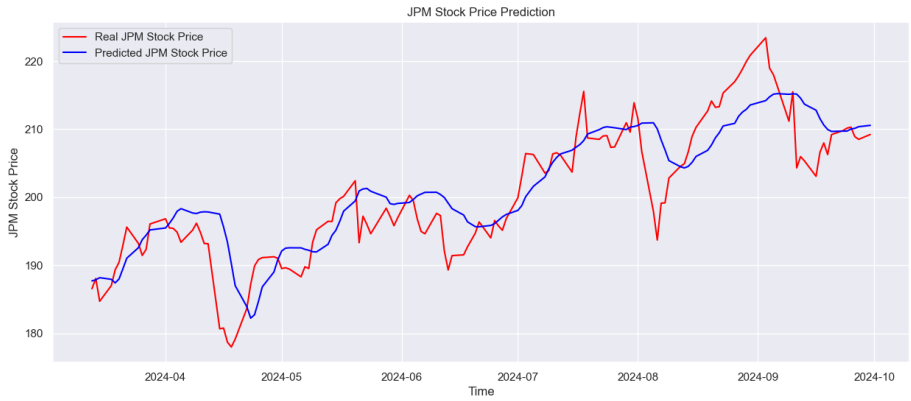
Figure 6: Performance Metrics for Baseline and Enhanced LSTM Models

Notably, for J.P. Morgan Chase, the enhanced model had roughly half the root mean squared error of the baseline model. The improved inference produced by the enhanced model on next day prices is illustrated in Figure 7. This figure makes it evident that there were factors or patterns not captured by the model during training. The near double RMSE and divergence towards the end of the testing period show that the enhanced model with sentiment as an input is significantly

more well adapted for certain companies and sectors than others.



(a) Baseline Model



(b) Enhanced Model

Figure 7: Next Day Price Predictions for Test Split of JPM Data

5 Discussion

Other studies and research focused solely on calculating sentiment or determining a correlation between sentiment and price, while our project attempts to use sentiment to derive actionable predictions by assuming correlation. While we have tested and found meaningful findings with regards to fine-tuning for enhanced sentiment analysis and incorporating sentiment with real stock price predictions, there are still key areas that need to be addressed in future research.

5.1 Imbalanced Dataset for Fine Tuning

The labeled dataset used to fine tune our model consist of 2308 neutral values, 1074 positive values, and 494 negative values. This dataset skewed towards neutral values can explain why model performance was much better at predicting neutral sentiment compared to positive and

negative news headlines. While our approach and model had performance that was comparable to existing models, we believe that there is room for improvement. We recommend taking actions to deal with the imbalanced dataset, such as modifying the loss function by adding weights proportional to the amount of positive, negative, and neutral data, or augmenting the data so that there are an equal amount of labeled data.

5.2 Limitations in Real Data sources

We did not dive deep into the correlation between sentiment and price movement, only checking to see if they were somewhat correlated. Further research into whether sentiment is a leading or lagging indicator of market movement and by what window of time steps will give a better understanding of the effectiveness of using sentiment for our predictions.

All findings related to real time data and news articles are dependent on the specific company, meaning some stock tickers could yield a stronger/more accurate result, while others would have a weaker result. Due to time constraints, we only tested a few companies, so more stock tickers, and aggregate results based on industry or company size could be worth exploring to get more generalizable results.

5.3 LSTM prediction model

Long short term memory recurrent neural networks are considered state of the art for time series prediction given their ability to remember specific information across time steps and forget selectively. While more robust than simple RNN architectures, LSTMs alone are not perfectly suited to predict sequence data when underlying patterns are highly correlated with other data which are not provided or available, or when the fluctuations in the sequence are near random. We believe our research would have benefited from a more robust or ensemble model which included additional architectures to extract more abstract representations in the data for inference. We propose the addition of a hierarchical LSTM ensemble with the addition of a 1 dimensional CNN model for feature extraction which together would be more capable of identifying patterns at both the most granular level, as well as across longer groups of time steps, improving overall inference.

6 Conclusion

Our work successfully achieved the original goal of improving sentiment analysis performance through fine-tuning a LLM and implementing a method to utilize real-time financial news headline sentiment in tandem with stock prices for prediction. Using APIs and data pipelines, we demonstrated the feasibility of a real-time stock price prediction system that integrates sentiment analysis for better performance. Key findings and objectives reached from our capstone:

1. Sentiment Analysis Performance: Fine-tuning with QLoRA improved accuracy from 76% to 87%, indicating enhanced understanding of financial contexts, while matching performance of existing sentiment classification models.

2. Correlation Between Sentiment and Stock Prices: Aggregated daily sentiment scores showed a positive correlation with stock price movements, suggesting that sentiment is correlated and can be used to make relevant decisions

3. LSTM Model for Stock Price Prediction: Our LSTM Model when provided an additional aggregated daily sentiment as an input had a lower RMSE, and a marginally higher improvement in performance when predicting price movement, demonstrating that sentiment can enhance prediction rates.

We demonstrate that fine-tuning large language models like LLaMA 3.1 with domain-specific data using QLoRA significantly enhances performance in financial sentiment analysis. Moreover, integrating sentiment data into stock price prediction models, such as a LSTM, can improve predictive accuracy.

Future work expanding on this project could be developing a fully operational real-time sentiment analysis and stock prediction application to provide immediate practical applications for traders and analysts. After demonstrating that daily sentiment contributes to better predictions, exploring more advanced models while incorporating other market indicators, such as trading volume, volatility indices, and macroeconomic factors, could generalize and improve predictive accuracy even further.

Future research can refine and extend our methodologies, contributing to more accurate, real-time, and actionable insights from sentiment analysis in financial forecasting.

References

[1] B. Hasselgren, C. Chrysoulas, N. Pitropakis, and W. J. Buchanan, "Using social media i& sentiment analysis to make investment decisions," *Future Internet*, vol. 15, no. 1, 2023. [Online]. Available: https://www.mdpi.com/1999-5903/15/1/5

[2] P. Padhanarath, Y. Aunhathaweesup, and S. Kiattisin, "Sentiment analysis and relationship between social media and stock market: pantip.com and set," *IOP Conference Series: Materials Science and Engineering*, vol. 620, no. 1, p. 012094, sep 2019. [Online]. Available: https://dx.doi.org/10.1088/1757-899X/620/1/012094

[3] S. Gupta, R. Ranjan, and S. N. Singh, "Comprehensive study on sentiment analysis: From rule-based to modern llm based system," 2024. [Online]. Available: https://arxiv.org/abs/2409.09989

[4] O. A. Smith S., "Comparing traditional news and social media with stock price movements; which comes first, the news or the price change?" *Big Data*, vol. 9, 2022.

[5] J. Smailović, M. Grčar, N. Lavrač, and M. Žnidaršič, "Predictive sentiment analysis of tweets: A stock market application," in *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, A. Holzinger and G. Pasi, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 77–88.

[6] D. Valle-Cruz, V. Fernandez-Cortez, A. López-Chau *et al.*, "Does twitter affect stock market decisions? financial sentiment analysis during pandemics: A comparative study of the h1n1 and the covid-19 periods," *Cognitive Computation*, vol. 14, pp. 372–387, 2022.

[7] K. Kirtac and G. Germano, "Enhanced financial sentiment analysis and trading strategy development using large language models," in *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, O. De Clercq, V. Barriere, J. Barnes, R. Klinger, J. Sedoc, and S. Tafreshi, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 1–10. [Online]. Available: https://aclanthology.org/2024.wassa-1.1

[8] P. Agarwal and A. Gupta, "Strategic business insights through enhanced financial sentiment analysis: A fine-tuned llama 2 approach," in *2024 International Conference on Inventive Computation Technologies (ICICT)*, 2024, pp. 1446–1453.

[9] Q. Liu and H. Son, "Methods for aggregating investor sentiment from social media," *Humanities and Social Sciences Communications*, vol. 11, p. 925, 2024.

[10] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. Raffel, "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning," in *Proceedings of the 36th Conference on Neural Information Processing Systems*, 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/0cde695b83bd186c1fd456302888454c-Paper-Conference.pdf

[11] C. Jeong, "Fine-tuning and utilization methods of domain-specific llms," *Journal of Intelligence and Information Systems*, vol. 30, no. 1, p. 93–120, Mar. 2024. [Online]. Available: http://dx.doi.org/10.13088/jiis.2024.30.1.093

[12] A. Chavan, Z. Liu, D. Gupta, E. Xing, and Z. Shen, "One-for-all: Generalized lora for parameter-efficient fine-tuning," 2023. [Online]. Available: https://arxiv.org/abs/2306.07967

[13] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," 2023. [Online]. Available: https://arxiv.org/abs/2305.14314

Appendix

1. Source code on Github: https://github.com/superkosat/Senior-Capstone
2. LLaMA 3.1 8B Finetuned model weights: https://huggingface.co/Nrezhang/FinancialLLama3.1