Department of Computer Science

香港城市大學
City University of Hong Kong

Financial KOLs' Opinion Mining based on Pre-trained Language Models

CS6534 Project Final Report

Student Name: WANG Yuchen

Student Name: 58183945

Supervisor: Prof. SONG Linqi

The 2nd Reader: Prof. Cheung Man Hon

Table of Contents

1. Introduction

1.1 Background

Individuals known as Key Opinion Leaders (KOLs) hold the power to shape the views and beliefs of their audiences by strategically sharing information, molding news stories, and offering their unique perspectives. As technology continues to evolve, the increasing number of social media users has created a fertile ground for KOLs to emerge and exert their influence on society [1]. In the realm of finance, social media influencers hold substantial sway in bringing together news, offering valuable perspectives, and forecasting market trends through their expressed sentiments. These opinions can rapidly spread and impact investors' approaches and expectations, as financial trading is largely based on anticipation [2]. This is particularly true for large-scale retail investors who usually have limited knowledge of analysis and rely heavily on expert advice.

It's important that relying solely on KOLs for investment advice may not always be the best course of action. While they may seem trustworthy, there's always the possibility that they could be misleading or even manipulative in their opinions. This can make it difficult for investors to discern the accuracy of their advice. In evaluating the predictive accuracy of authors on Seeking Alpha, who are considered KOLs, it has been established that at most 53% of stock performance predictions are accurate, which is only slightly better than random chance [3]. As a result, blindly following the advice of KOLs is considered risky as it can lead to unwelcome financial losses. In view of this, it is crucial to build a system using Natural Language Processing techniques to analyze the sentiment of such advice, assess its reliability and aggregate it to arrive at the combined advice of multiple highly credible financial KOLs. Such a system will effectively help retail investors filter out unreliable information and make informed investment decisions.

Sentiment analysis is a widely studied field that aims to determine the overall sentiment of social media posts. In the financial industry, sentiment analysis is commonly used to predict financial time series. However, the reliability of these predictions is often not justified. Most studies in this area focus on analyzing large, aggregated datasets from popular social media and finance-specific platforms like Twitter, Sina Weibo, Stock Twits, and Seeking Alpha [3]. While these platforms provide valuable insights, they do not consider the opinions of influential individuals.

1.2 Project Objectives & Scope

The primary objective of this project is to develop a sentiment analysis system for the

opinions of KOLs. This system focuses on a selected group of highly influential KOLs in the financial sector and analyses related social media tweets or financial articles. Through the sentiment analysis module, the system identifies the positive or negative emotions expressed by KOLs towards a particular stock, thereby identifying bullish or bearish expectations.

Based on the sentiment-derived expectations, an evaluation matrix is developed to measure the reliability of posts and KOLs over different time intervals and rank them accordingly. By aggregating the expectations of highly reliable KOLs, the system aims to predict the stock prices of certain stocks over a future period.

In order to guarantee the reliability of the sentiment analysis results, this paper will concentrate on examining the performance of Llama 2, a popular and sophisticated pre-trained large language model within the AI community, on the task of sentiment analysis of financial texts. Furthermore, another objective is to utilize the sentiment analysis results from this model to forecast the stock price of a specific stock over a defined timeframe and evaluate its predictive capabilities. Despite the system being designed for real-time application, the constraints of resources and time prevented the collection of sufficient data for a comprehensive evaluation. Consequently, our analysis focused on a representative sample of English-language financial influencers (KOLs), employing open-source historical web data to replicate the current scenario.

2. Related Work

This section aims to delve into the research findings relevant to this project. Initially, it will explore research that employs sentiment analysis techniques to extract valuable insights from financial social media content. Furthermore, it will introduce the currently popular pre-trained big language models and explore the feasibility of sentiment analysis for financial social media texts.

2.1 Sentiment Analysis on Financial Social Media

Prior research has leveraged various popular social media platforms to conduct sentiment analysis. Si et al. [4] developed a topic-oriented sentiment analysis framework specifically for tweets that include references to the Standard & Poor's 100 stocks (S&P100). This approach notably enhanced the accuracy of predicting movements in the S&P100 index prices. Similarly, Zhao et al. [5] demonstrated the effectiveness of using the general sentiment extracted from Weibo posts to improve the precision of daily price predictions for stocks in the new energy sector. These studies suggest that social media content holds significant predictive value for financial markets. However, these analyses predominantly utilize large, aggregated datasets primarily comprising contributions from retail investors, who typically possess limited financial expertise.

To extract meaningful insights and actionable investment advice from social media, it is beneficial to focus on prominent accounts known for their expertise. Research by Valle-Cruz et al. [6] pinpointed influential Twitter accounts such as The New York Times and Bloomberg. They discovered that during the H1N1 and Covid-19 pandemics, the sentiment expressed in tweets from these accounts closely mirrored stock market trends. However, their study primarily examined organizational accounts, neglecting the potential influence of prominent individual users. In a separate study, Wang et al. explored the relationship between the sentiment of online content and stock market performance, noting that posts by experts on Seeking Alpha generally performed better than those by regular users on StockTwits [3]. Despite this, the correlation between Seeking Alpha articles and stock performance remained modest until a select group of top authors was identified based on the volume of comments they attracted. The sentiment from these authors' posts showed a much stronger correlation with movements in stock prices. This approach could potentially be applied to other social media platforms with larger audiences, and additional criteria for ranking influential accounts could be considered.

2.2 Natural Language Processing and Pre-training Model

Natural Language Processing (NLP) is a technology that enables computers to process human language in the form of text or speech data. It involves understanding, interpreting, and manipulating human language, including the intentions and emotions of the speaker or writer [7]. Since the 1980s, NLP techniques have been extensively applied in the fields of financial market analysis and trading.

As research in the field of NLP deepens, pre-trained language models have opened vast opportunities for NLP [8]. The application of pre-training techniques in NLP allows well-trained language models to capture rich knowledge beneficial for downstream tasks, such as long-term dependencies and hierarchical relationships. These models are trained on any large corpus, meaning that there is virtually an unlimited amount of training data available for the pretraining process. Subsequently, the models are fine-tuned on a target dataset and task-specific layers are added.

A prime example of this is the Bidirectional Encoder Representations from Transformers (BERT) [9]. BERT is a powerful language model that has been pre-trained on masked token prediction and next sentence prediction. Unlike traditional unidirectional pre-training of language models, BERT trains deep bidirectional representations by jointly conditioning on both left and right context in all layers. Therefore, by adding an output layer, BERT can be successfully fine-tuned for a variety of downstream NLP tasks, including text sentiment classification. This approach marks a significant departure from earlier methods, leveraging the full context of input data for more nuanced and effective language understanding.

In addition, Liu et al. [10] found that BERT was significantly undertrained and introduced an improved pre-training method called RoBERTa. The adjustments made to the original BERT strategy involve the removal of the next sentence prediction objective. This change enhances RoBERTa's applicability to concise social media texts, such as tweets, which typically consist of a single sentence [11].

2.3 Large Language Models and Llama 2

Language models have been foundational to natural language processing tasks for a long time, particularly in predicting the next word in text sequences. Early models, such as statistical language models, employed probabilistic methods like n-grams to predict the next word based on the preceding n-1 words. This approach evolved into more sophisticated deep learning architectures, including Long Short-Term Memory (LSTM) networks.

In 2017, Vaswani et al. [13] introduced the Transformer architecture, catalyzing the development of large language models, such as OpenAI's ChatGPT and Meta's Llama 2. Compared to traditional transformer-based language models, models based on Generative Pretrained Transformers (GPT) possess key features—zero-shot and few-shot learning capabilities—that enable them to exhibit good performance even with no or minimal training data. Their advanced reasoning abilities allow them to generate new patterns and conclusions based on prompts and known facts, enabling the creation of analytical texts that contain unexpected yet valuable chains of thought.

This paper primarily investigates Llama 2, a collection of pre-trained and fine-tuned large language models developed by Meta Corporation, which was released to the public on July 18, 2023. These models vary in size from 7 billion to 70 billion parameters, including fine-tuned versions such as Llama-2-Chat, which have been optimized for conversational use cases [14]. Parameters represent weights obtained through training, subsequently used to predict subsequent tokens in a sequence. Llama 2's major objective is advanced natural language understanding [15], with applications spanning conversational AI, content generation, language translation, and information extraction. Llama 2 's comprehensive capabilities, encompassing contextual understanding, complex language processing, accuracy in sentiment prediction, and efficiency in handling large datasets, render it an indispensable tool in financial sentiment analysis. The model's effectiveness is further substantiated by various studies, including those by Pavlyshenko [16] and Breitung and Müller [17], reinforcing its status as a potent instrument for analyzing financial markets.

3. System Modeling and Structure

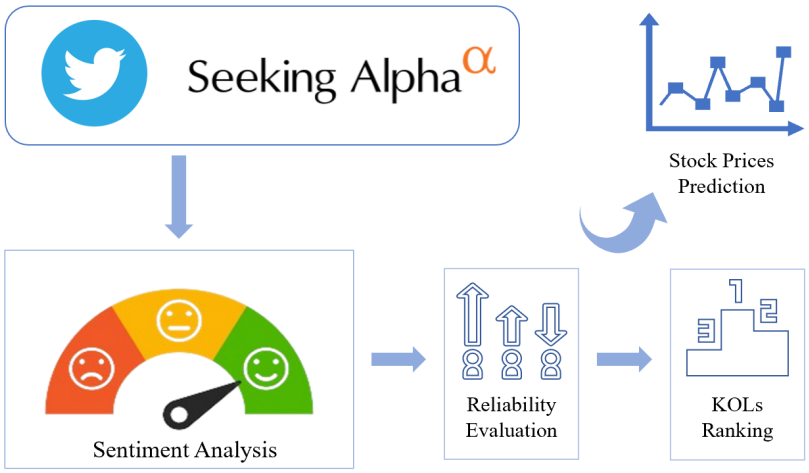

Figure 3.1 System design overview

Figure 3.1 provides an illustrative depiction of the architectural framework in this project. The system is divided into four modules: Data Collection, Sentiment Analysis, Reliability Assessment, and Stock Price Prediction. These modules form a complete logical loop, aiming to offer actionable insights derived from the emotional nuances present in social media and financial news disseminated by KOLs. This innovative methodology has the potential to transform the way investors interpret market sentiments and make informed decisions.

3.1 Data Description

To construct an accurate and comprehensive sentiment analysis module, we collected two categories of labeled financial text data. One category consisted of short textual content represented by tweets on Twitter, while the other comprises longer textual content primarily made up of financial news headlines and summaries. These datasets were employed for fine-tuning the model. In the evaluation and testing phase, a selection of influential financial experts on the Twitter and Seeking Alpha platforms was chosen for analysis of their tweets or news posts. This was done to assess and compare the model's performance after refinement. The detailed description of these datasets is below.

3.1.1 Twitter Financial News – For Fine-tuning

It is noteworthy that compared to formal written language, tweets exhibit a stark contrast in tone and style, often being more casual and straightforward. During the fine-tuning process, we selected the **Twitter Financial News** Dataset [18] from Hugging Face to fine-tune our pre-trained models. This English dataset contains labeled tweets

related to finance, with 11,932 tweets annotated with three labels: Positive, Neutral, and Negative. The distribution of each sentiment category in the dataset is shown in Figure 3.2.

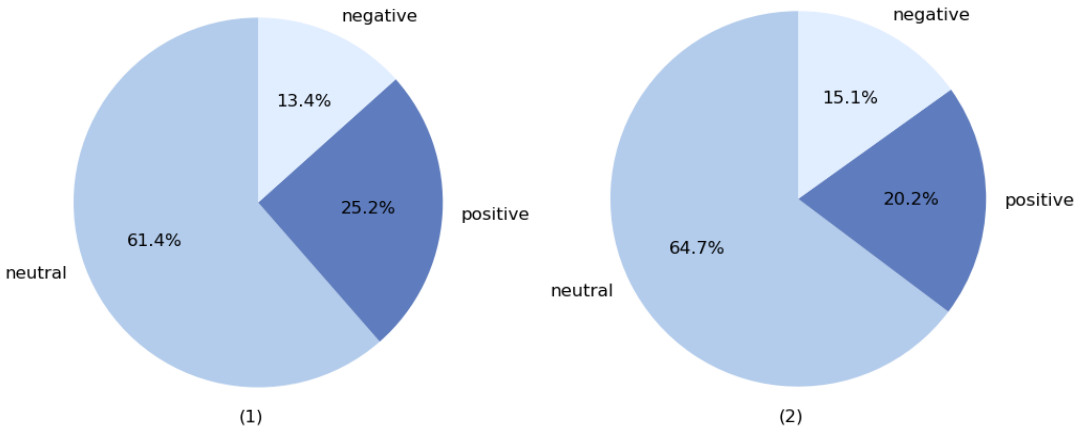

Figure 3.2 Distribution of different sentiment categories
in (1) Twitter Financial News and (2) Financial Phrase Bank

3.1.2 Financial Phrase Bank – For Fine-tuning

This dataset [19] is marked as serves as a valuable resource for fine-tuning. It boasts a collection of 4,840 English sentences, sourced randomly from financial news within the LexisNexis database. Each sentence is composed in formal written language and has been thoughtfully annotated by 16 experts in finance and business as either Positive, Neutral, or Negative. The dataset is divided into four subsets based on the level of agreement among annotators: **AllAgree**, **75Agree**, **66Agree**, and **50Agree**. For this project, we primarily utilized the **AllAgree** dataset for fine-tuning purposes. The distribution of each sentiment category in the dataset is shown in Figure 3.2.

3.1.3 Financial Tweets – for Evaluation

To effectively evaluate the fine-tuned model, we utilized StockerBot [20] to collect a tweet dataset that is independent of the fine-tuning data and related to real stocks from the period of February 26, 2023, to March 13, 2023. We compiled a list of influential Twitter finance accounts and a watchlist of stocks. StockerBot searched in real-time through the Twitter API for tweets posted by identified KOLs. If these tweets mentioned any stocks on the watchlist, StockerBot stored them in a Financial KOL Tweets' file.

The project selected 30 financial KOLs on Twitter, among which the notable ones are *MarketWatch*, *YahooFinance*, *TechCrunch*, and *TheEconomist*. The stock watchlist encompasses all companies listed on the New York Stock Exchange and NASDAQ, along with their stock codes. Table 3.1 presents a part of the stock symbols and stocks in the stock watch list.

Ticker	Name
FB	Facebook
AMZN	Amazon
TSLA	Tesla
NFLX	Netflix
NVDA	NVIDIA

Table 3.1 Some of the stock symbols and stocks in the stock watch list

3.1.4 Seeking Alpha Articles – for Evaluation

Seeking Alpha is a crowdsourced financial website that permits verified authors to publish articles concerning financial markets. Prior to data collection, we identified a list of Seeking Alpha authors based on the number of followers and their influence. As is shown in Table 3.2, seven popular authors who mainly wrote articles in stock analysis were selected in this project.

We used the Seeking Alpha RapidAPI [21] to gather financial articles from these selected authors between February 26, 2023, and March 13, 2023. For each article, its author, timestamp, article summary, and tagged stock symbols were stored in the dataset.

Author	Followers
Trapping Value	37.67K
Hoya Capital	32.67K
Taylor Dart	27.1K
Daniel Jones	26.3K
Stephen Simpson	18.81K
Nick Ackerman	11.69K
Gen Alpha	14.9K

Table 3.2 Seeking Alpha authors and their number of followers

```
stock = yf.Ticker("TSLA")
df_stock = stock.history(start="2024-01-01", end="2024-01-07")
df_stock
```

Executed at 2024.07.22 11:31:03 in 116ms

4 rows 4 rows × 7 columns

Date	Open	High	Low	Close	Volume
2024-01-02 00:00:00-05:00	250.080002	251.250000	244.410004	248.419998	104654200
2024-01-03 00:00:00-05:00	244.979996	245.679993	236.320007	238.449997	121082600
2024-01-04 00:00:00-05:00	239.250000	242.699997	237.729996	237.929993	102629300
2024-01-05 00:00:00-05:00	236.860001	240.119995	234.899994	237.490005	92379400

Figure 3.3 TSLA's stock prices obtained by Yahoo Finance API

3.1.5 Stock Price Data

During the evaluation process, we need to obtain the actual market price data of the specified stocks as the ground truth for comparison. The price data of these specified stocks and market indices were collected from the Yahoo Finance API [22]. As illustrated in Figure 3.4, for each trading day within the specified time period, seven values related to stock prices and trading volumes were retrieved. For a particular stock, the highest and lowest prices would be used to establish the fundamental facts of price changes in the KOL reliability assessment.

3.2 Sentiment Analysis

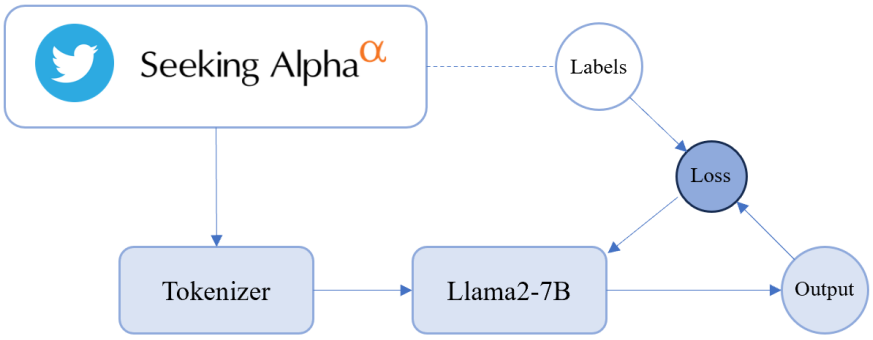

Figure 3.4 Sentiment Analysis Module

Sentiment	Label
Negative	0
Positive	1
Neutral	2

Table 3.3 Emotional Label Correspondence Table

Given the constraints of experimental hardware and the volume of training data available, we opted for the Llama 2 model with 7 billion parameters as our primary experimental subject. As is illustrated in Figure 3.4, to accommodate inputs with varying linguistic features, we utilized labeled **Twitter Financial News**, the **AllAgree** subset from **Financial Phrase Bank**, and a combination of both datasets to fine-tune the Llama-2-7B model. This fine-tuning process is aimed at constructing a sentiment classification model specifically tailored for financial text. The model processes encoded dataset texts as input and outputs three labels indicating sentiment—negative, positive, and neutral.

3.3 Reliability Evaluation

Upon receiving the sentiment analysis results from LLMs, a reliability assessment process can be undertaken. Based on the accuracy of sentiment analysis, we can compute reliability scores of every KOL to measure the prediction precision of the KOL's opinions. The calculation involves two sets of variables, S and P. S represents the sentiment indicator, with discrete values of 0, 1, and 2 (Negative, Positive and Neutral); whereas P denotes the price movement indicator for the discussed stock, capturing both the direction and magnitude of the price movement through discrete values of 0, 1, and 2 (Fall, Rise and Fluctuate). Additionally, different time frames are taken into account in determining the price movement. A comprehensive explanation will be provided in the methodology section.

3.4 KOLs Ranking and Stock Price Prediction

It has been demonstrated that leveraging sentiment information extracted from social media for trading decisions holds significant profit potential. Following the assessment of reliability among KOLs, we ranked their reliability based on the accuracy of historical stock trend predictions within varying time intervals. A higher reliability ranking indicates a closer alignment between the opinions expressed by these KOLs on social media regarding stock predictions and actual stock price movements during specific time windows.

Building upon this foundation, we can aggregate opinions from multiple influential KOLs regarding a particular stock to predict short-term price movements of selected stocks. The specific predictive models will be described in depth in the methodology section.

4 Methodology and Algorithms

4.1 Implementation Environment

Python is the go-to language for all stages of our implementations, covering data collection and pre-processing, model construction, KOL reliability evaluation, and building investment strategies. We fine-tuned our sentiment analysis model on the PyTorch framework.

For the task of training large language models, powerful GPUs and high-memory systems are indispensable. To accomplish this project, I initially utilized the high-throughput GPU cluster servers provided by the Computer Science Department at City University as my experimental environment. These servers were employed for data

processing and baseline model training.

The high-throughput GPU cluster (HTGC) is a specialized HTCondor cluster focused on GPU-related computational applications. The specific configurations of the HTGC are detailed in Table 4.1.

Configuration List	
GPU	7 x Nvidia V100
Maximum Memory Size Per Process	64 GB
OS	Ubuntu 20.04
GPU Memory	16 GB
CUDA Runtime Version	11.2
CUDA Driver Version	11.6
CUDA Capability	7.0

Table 4.1 HTGC1 Configuration List

Configuration List	
GPU	4 x Nvidia A40
Maximum Memory Size Per Process	64 GB
OS	Ubuntu 20.04
GPU Memory	45 GB
CUDA Runtime Version	12.2
CUDA Driver Version	12.1

Table 4.2 Configuration List of Server with A40

As the experiment progressed, I found it extremely challenging to continue training the llama 2 model in the above environment. Training the llama-2-7b model on a V100 graphics card with no AI acceleration capability took more than thirty hours for five epochs of training on 10,000 pieces of data, which made subsequent experimental progress very difficult. As a result, the subsequent experiments were conducted in a server environment equipped with four A40 graphics cards. The specific environment configuration is shown in Table 4.2.

4.2 RoBERTa-Large

Based on existing research, we selected the RoBERTa-Large pre-trained model, which has performed well in text classification and sentiment analysis tasks in past experiments, as the baseline model for this project. We need to provide a point of reference for the fine-tuning of the Llama-2-7B model, so that the experimental results can be quantified and comparable.

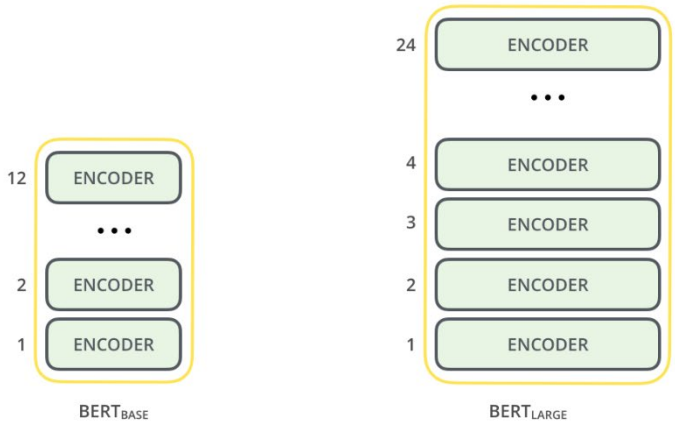

Figure 4.1 Comparison of BERT-Base and BERT-Large architecture

RoBERTa-Large is an advanced version of RoBERTa that uses a larger corpus and architecture, resulting in enhanced performance on various fine-tuning tasks. By replacing the top layer with a task-specific layer, it can be further trained for downstream tasks. The RoBERTa is trained it over the architecture of the BERT. Compared to BERT-Base used in RoBERTa, RoBERTa-Large with BERT-Large architecture has more encoder layers, which boosts the performance of the RoBERTa model on fine-tuning tasks.

4.3 Llama 2

4.3.1 Structure of Llama 2

The Llama 2 model is the focus of our research in this project. Llama 2 is built on an optimized auto-regressive transformer architecture, meaning it employs a neural network design of transformer tailored to generate outputs sequentially. This architecture has been optimized to enhance performance in language-related tasks. The Figure 4.2 shows the transformer structure for Llama 2 which contains the MLP layer, Norm layer and Multi-Head Attention layers.

The training of Llama 2 involved processing approximately two trillion tokens, sourced from a variety of public domains, which is roughly twice the amount handled by its predecessor, Llama 1. The focus of this extensive dataset was on data cleanliness and factual accuracy, which significantly enriches the model's knowledge base and minimizes errors. Additionally, Llama 2's capability to process up to 4096 tokens enables it to generate highly accurate and relevant responses to user queries. Comparative analysis presented in Table 4.2 delineates the distinctions between the attributes of the new Llama 2 models and their predecessors, the Llama 1 models.

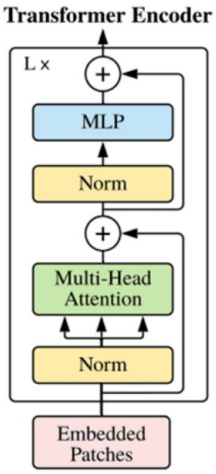

Figure 4.2 The transformer structure for Llama 2

	Training Data	Params	Context Length	GQA	Tokens	LR
LLAMA 1	*See Touvron et al. (2023)*	7B	2k	✗	1.0T	3.0×10^{-4}
		13B	2k	✗	1.0T	3.0×10^{-4}
		33B	2k	✗	1.4T	1.5×10^{-4}
		65B	2k	✗	1.4T	1.5×10^{-4}
LLAMA 2	*A new mix of publicly available online data*	7B	4k	✗	2.0T	3.0×10^{-4}
		13B	4k	✗	2.0T	3.0×10^{-4}
		34B	4k	✓	2.0T	1.5×10^{-4}
		70B	4k	✓	2.0T	1.5×10^{-4}

Table 4.2 Llama 2 family of models. Token counts refer to pretraining data only. All models are trained with a global batch size of 4M tokens. Bigger models like 34B and 70B use GQA for improved inference scalability. [14]

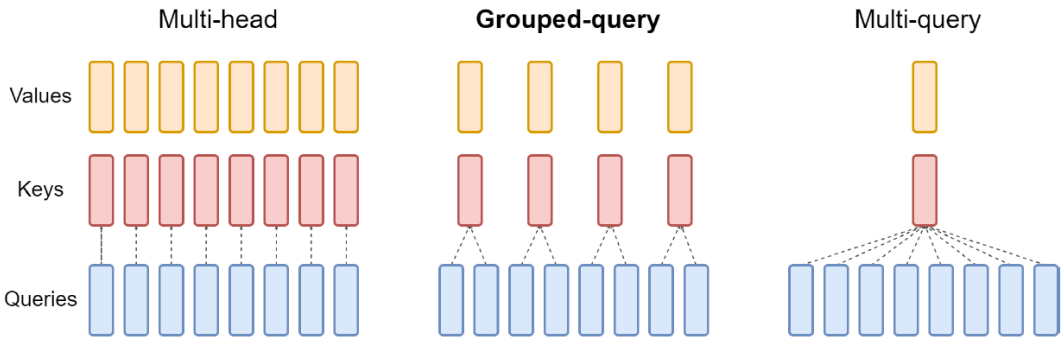

Figure 4.3 Overview of grouped-query method. Multi-head attention has H query, key, and value heads. Multi-query attention shares single key and value heads across all query heads. Grouped-query attention instead shares single key and value heads for each group of query heads, interpolating between multi-head and multi-query attention. [24]

To achieve its superior performance, Llama 2 incorporates several cutting-edge development paradigms, including rejection sampling, GQA, and GAtt. In detail, the enhancements encompass comprehensive data cleansing processes, refined data combinations, and an expansion of training data by 40%, alongside a doubling of the

context length. These modifications leverage grouped-query attention (GQA) to augment inference scalability within larger model frameworks. Figure 4.3 shows a comparison of grouped-query attention and multi-head/multi-query attention. These technological advancements significantly elevate the model's performance across diverse benchmarks.

To improve the reliability and safety of Llama 2 in real-world applications, a unique reinforcement learning human feedback (RLHF) model tailored for safety and practicality was adopted in the fine-tuning process of the model [14]. This dual-model strategy ensures that Llama 2 not only provides pertinent responses consistent with human preferences, but also maintains a high standard of safety by reducing the risk of generating harmful content. This makes Llama 2 a more reliable and safer choice for widespread applications.

4.3.2 Adaptations for Classification Task

During our experimental process, we observed that the Llama-2-7B model, even after fine-tuning, did not perform well on an independent test dataset. Consequently, to enhance the model's performance in this sentiment classification task, we implemented a small adjustment to the structure of Llama 2.

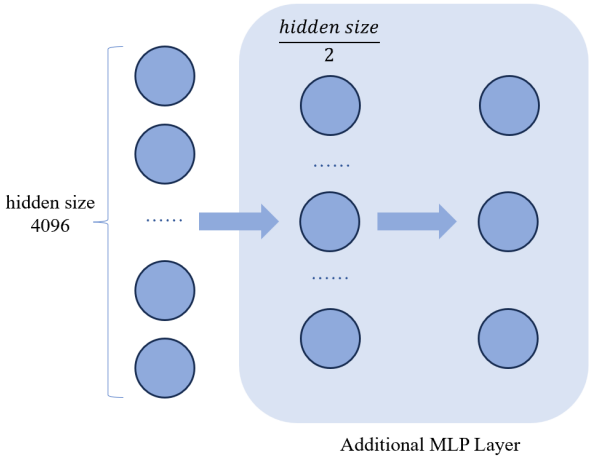

Figure 4.4 Structure of Additional MLP Layer

As depicted in Figure 4.4, we integrated an additional Multi-Layer Perceptron (MLP) layer into the output layer of the Llama 2 model. This adjustment was designed to bolster the model's capability for extracting features and to effectively reduce dimensionality, thereby facilitating more precise classification results. The MLP layer comprises a fully connected layer with a size equal to half of the hidden size, and it utilizes the ReLU activation function. This configuration not only amplifies the non-linear representation of features but also significantly enhances the model's ability to generalize across complex sentiment classification tasks.

4.4 LoRA

Many applications in natural language processing rely on adopting one large-scale, pre-trained language model that can be applied to multiple downstream applications. This adaptation is typically achieved through fine-tuning, which involves updating all the parameters of the pre-trained model. However, a major drawback of fine-tuning is that the new model contains just as many parameters as the original model. As larger models are trained every few months, this changes from a mere "inconvenience" for GPT-2 [25] or RoBERTa-Large [10] to a critical deployment challenge for GPT-3 [26] with 175 billion trainable parameters.

Hu, E. J. et al. [27] drew inspiration from the work of Li et al. [28] and Aghajanyan et al. [29] and found that the learned over-parametrized models in fact resided on a low intrinsic dimension. Hu, E. J. et al. hypothesize that the change in weights during model adaptation also has a low "intrinsic rank", leading to our proposed Low-Rank Adaptation (LoRA) approach. LoRA allows us to train some dense layers in a neural network indirectly by optimizing rank decomposition matrices of the dense layers' change during adaptation instead while keeping the pre-trained weights frozen. By freezing the core model, it facilitates rapid task-switching through the simple replacement of matrices A and B, as depicted in Figure 4.5. This approach substantially diminishes both the storage requirements and the overhead associated with switching tasks.

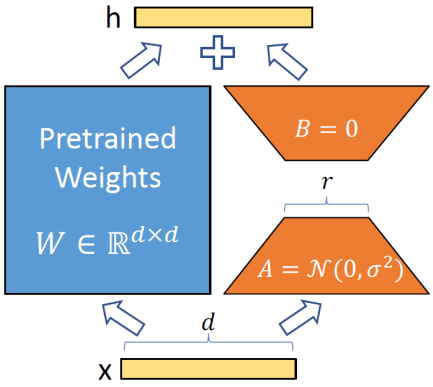

Figure 4.5 Reparameterization of LoRA, only training A and B [26]

Furthermore, LoRA enhances training efficiency and significantly reduces the computational demands typically required by hardware. This reduction is achieved by employing adaptive optimizers, which obviate the need for gradient calculations and maintenance of optimizer states for the majority of parameters. Instead, optimization is confined to the smaller, injected low-rank matrices, which are up to three times less demanding in terms of hardware resources. The design of LoRA is inherently linear, which allows for the seamless integration of trainable matrices with the pre-existing frozen weights during deployment. This integration ensures that there is no additional inference latency introduced when compared to a model that is fully fine-tuned. In the

forthcoming experiment, we will apply LoRA to fine-tune the Llama 2 model.

4.5 Reliability Evaluation

In order to quantify the prediction accuracy of the KOLs, an accuracy-based reliability score has been defined, which will be used to rank the reliability of different KOLs' social media opinions. In this algorithm, we need to calculate two sets of variables, S and P. S is the sentiment indicator, which takes discrete values of 0, 1 and 2, meaning negative, positive and neutral respectively. For each KOL, on each day, they post their opinions about several stocks via social media. At day i for a stock m, $\bar{s}_m^i$ indicates that day's rounded average sentiment about the stock. As the calculation of the average sentiment can result in floating numbers, the rounding procedure ensures that the average sentiment is a discrete integer.

In order to capture the trend movement of the price, a trend indicator Δ_m^i is defined.

$$\Delta_m^i = \max\_abs\left(\frac{c_m^{i+1}-o_m^i}{o_m^i}, \frac{c_m^{i+1}-o_m^i}{o_m^i}\right) \quad (1)$$

where c_m^i denotes the close price of stock m at day i, o_m^i denotes the open price of stock m at day i, t denotes the time window, *max_abs(a,b)* returns the number that has greater absolute value.

At day i for a stock m, $p_m^{i,t}$ denotes that day's price movement indicator of the stock.

$$p_m^{i,t} = \begin{cases} 0, & if\ \Delta_m^{i,t} < \alpha \\ 2, & if\ -\alpha \le \Delta_m^{i,t} \le \alpha \\ 1, & if\ \Delta_m^{i,t} > \alpha \end{cases} \quad (2)$$

where α is the significance level of the price trend and $p_m^{i,t}$ takes discrete value from $\{0,1,2\}$, where 0 means that the stock price has dropped significantly, 1 means that the stock price has risen significantly, and 2 means that the stock price fluctuates slightly. To ensure rationality, α is adjusted according to experiment.

For the time span of N days and defined time window t, we can get a series of average sentiment and a series of price movement indicator.

$$S = \left(\bar{s}_1^1, \dots, \bar{s}_j^1, \bar{s}_1^2, \dots, \bar{s}_n^2, \dots, \bar{s}_1^N, \dots, \bar{s}_u^N\right) \quad (3)$$

$$P^t = \left(p_1^{1,t}, \dots, p_j^{1,t}, p_1^{2,t}, \dots, p_n^{2,t}, \dots, p_1^{N,t}, \dots, p_u^{N,t}\right) \quad (4)$$

From these two series, we can obtain a confusion matrix.

Price	Sentiment		
	Negative 0	Positive 1	Neutral 2
Fall 0	T_0	T_{01}	T_{02}
Rise 1	T_{10}	T_1	T_{12}
Fluctuate 2	T_{20}	T_{21}	T_2

Table 4.3 Confusion matrix of two series

$$R^t = \frac{\sum_{i=0}^{2} T_i}{\sum_{i=0}^{2} T_i + \sum_{i=0}^{2} \sum_{j \in C, j \neq i} F_{ij}}, C = \{0,1,2\} \quad (5)$$

where R^t is the accuracy-based reliability score for time window t. The higher the value of R^t, the more accurate the KOL's prediction in that time window.

4.6 Stock Price Prediction Algorithm

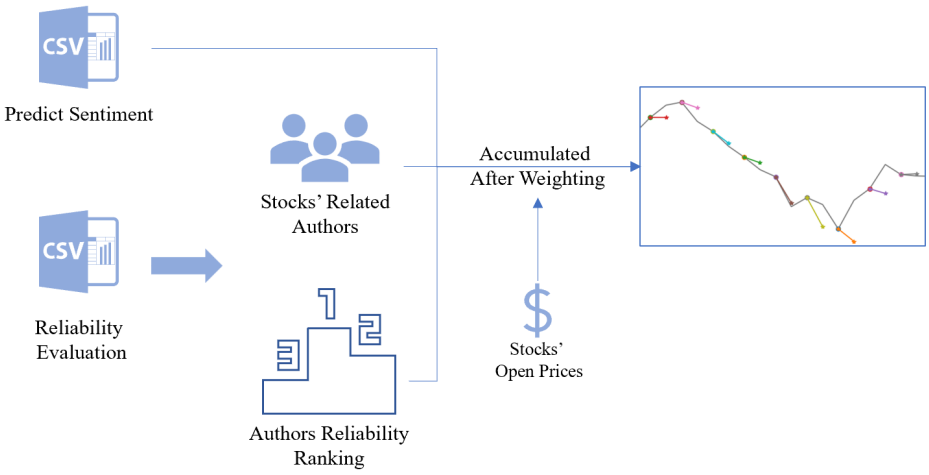

Figure 4.6 Data flow of stock price prediction

Taking into account the sentiment analysis results and the reliability ranking of KOLs, we are proposing an algorithm for predicting stock prices based on the sentiment expressed by KOLs. This algorithm aims to forecast short-term stock price movements, and the corresponding data flow is illustrated in Figure 4.6.

Suppose that for stock m, we obtain the comments of N financial KOLs on stock m over a period of time and judge their sentiment direction based on the sentiment analysis model to obtain the corresponding sentiment $s_m^i (i \in N)$. We assume that the impact of different sentiments on the stock price prediction is computed based on the daily opening price o_m^t. We then calculate the impact of different sentiments on the stock price prediction based on the daily opening price.

$$influence_m^t = \begin{cases} -\alpha \times o_m^t, & if\ s_m^i = 0 \\ \alpha \times o_m^t, & if\ s_m^i = 1 \\ random(-\beta \times o_m^t, \beta \times o_m^t), & if\ s_m^i = 2 \end{cases} \quad (6)$$

where α and β are the stock price influence factors, reflecting an expected value of a different sentiment on the stock price movement, in the experiment we set them to 0.03 and 0.01 respectively. s_i^m takes discrete value from {0,1,2}, where 0 denotes positive sentiment, 1 denotes negative sentiment and 2 denotes neutral sentiment.

Combining the results of the reliability analyses of the KOLs' respective *win* at different time windows, we can obtain stock price changes prediction of a certain KOL *i* for stock *m* on day *t*.

$$price_{m,i}^t = influence_m^t \times \frac{1}{1-e^{5-10\times reliable_i^{win}}} \quad (7)$$

where, $reliable_i^{win}$ is the reliability score of KOL *i* in the time window *win*. By adding up the predicted price increases or decreases of multiple KOLs for stock *m* on day *t*, we can get a prediction for the closing price of this stock on that day

$$close\_price_t = o_m^t + \sum price_{m,i}^t \quad (8)$$

5 Experiment and Evaluation

5.1 RoBERTa

5.1.1 SMART Framework for Fine-tuning

The SMART framework [12] is a powerful tool for fine-tuning pre-trained language models in a robust and efficient manner. By utilizing this framework, we have been able to achieve new state-of-the-art results on a variety of NLP tasks. To implement it in this project, we used the smart-pytorch Python library [30].

To ensure that the model complexity is effectively controlled during the fine-tuning process, we incorporated a smoothness-inducing adversarial regularization. This regularization ensures that even if a small perturbation is introduced to the input, the model output will not change significantly. As depicted in Figure 5.1, the decision boundary learned with this regularization (b) is smoother within the neighborhoods of training data points compared to the one learned without it (a). This characteristic is beneficial as it aids in mitigating overfitting and enhances the model's ability to generalize, particularly in resource-constrained domains like the financial sentiment

analysis task described in this study.

A set of Bregman proximal point optimization techniques has been suggested to avoid aggressive updating. These methods incorporate a trust-region-like regularization at each iteration to stabilize the fine-tuning process. The updates are then performed within a smaller range of the previous iteration to achieve better results.

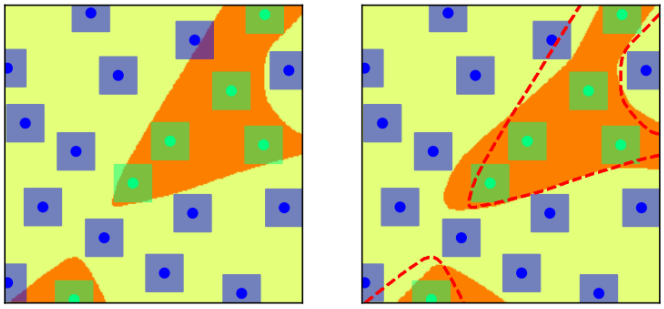

Figure 5.1 The Effect of Smoothening Adversarial Regularization on the Decision Boundary [12]

5.1.2 Experiment Result

Based on previous research, we need further pre-training techniques to enhance the performance of the RoBERTa-Large sentiment analysis model on financial tweet text. In the further pre-training process, the RoBERTa-Large model was trained with the Kaggle Financial Tweets [23] dataset. This dataset contains 28,275 unlabeled English financial tweets.

Following further pre-training, we used **Twitter Financial News** dataset and the **Financial Phrase Bank** dataset to fine-tune the pre-trained RoBERTa-Large model respectively. We divided the dataset into training and testing sets in an 8:2 ratio and conducted fine-tuning for 16 epochs. During each epoch, we assessed the model's performance and saved the model at the end of each epoch.

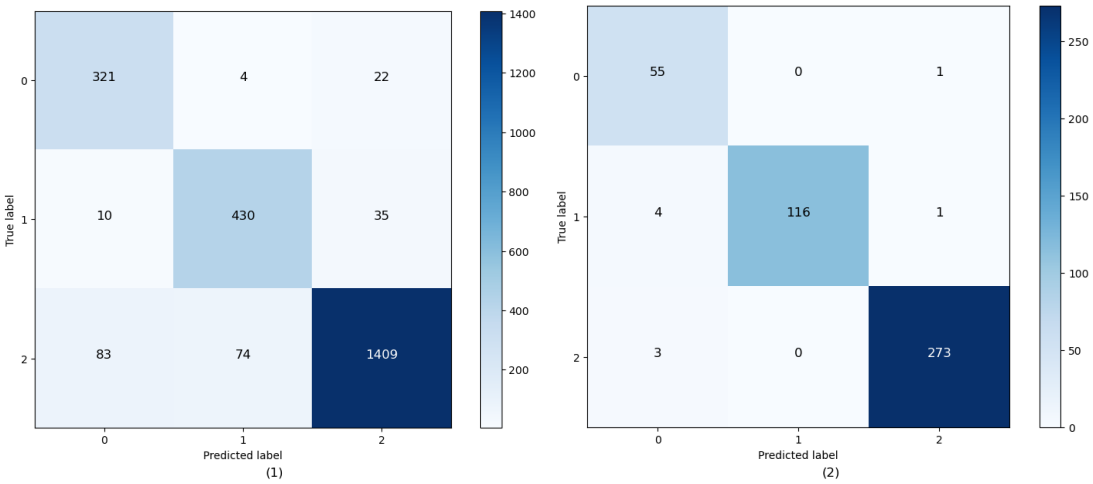

Figure 5.2 Confusion Matrix of RoBERTa-Large on test data
of (1) Twitter Financial News and (2) Financial Phrase Bank

After the fine-tuning process, we selected the best-performing model based on the classification accuracy on the test set. We further analyzed the performance of these models using confusion matrices, which provided detailed categorization of the model's predictions for different classes. Figure 5.2 displays the confusion matrix of the best model and Table 5.1 displays the performance of the evaluation metrics for the two models.

By examining the confusion matrix and evaluation metrics, it is clear that RoBERTa-Large exhibits exceptional performance in sentiment classification tasks on these two financial text datasets. We used these results as reference points for research on the Llama-2-7B model.

	Micro-F1	**Macro-F1**
RoBERTa-Large-Twitter	0.904	0.882
RoBERTa-Large-PhraseBank	0.980	0.967

Table 5.1 Evaluation Metrics of RoBERTa-Large on test data of Twitter Financial News and Financial Phrase Bank

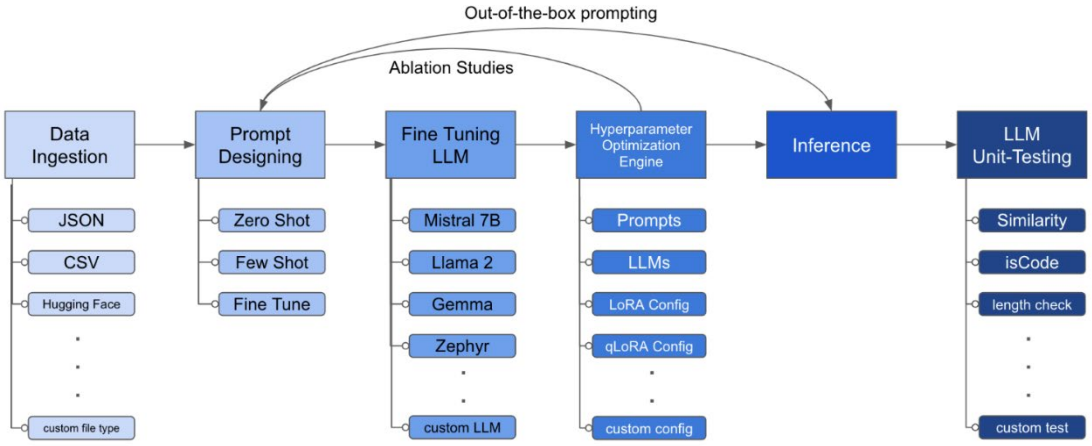

Figure 5.3 The structure of the LLM Fine-tuning Toolkit [31]

5.2 Llama-2-7B

5.2.1 LLM-Finetuning-Toolkit for Fine-tuning

During the fine-tuning process of Llama 2, we utilized the LLM Fine-tuning Toolkit [31], a versatile, open-source command-line interface tool designed for streamlined fine-tuning of language models, to fine-tune the Llama-2-7B model with our financial data. This toolkit supports a configuration-based approach, enabling researchers to specify experimental parameters through a single YAML configuration file. This file governs various aspects of the experimentation pipeline, including the selection of prompts, the choice of language models, the optimization strategies employed, and the

evaluation of model performance. The architecture and functionality of the LLM Fine-tuning Toolkit are depicted in Figure 5.3.

In this experiment, we primarily utilized the code from the LLM Fine-Tuning module, specifically the fine-tuning section for Llama 2. Modifications were made based on this foundation.

5.2.2 Experiment Result

During the fine-tuning, we ensured that training data settings were consistent with those used for fine-tuning the RoBERTa-Large model. This consistency guarantees comparability between the experimental results of the two models. We selected the best-performing models on each of the two datasets with different fine-tuning parameters for research.

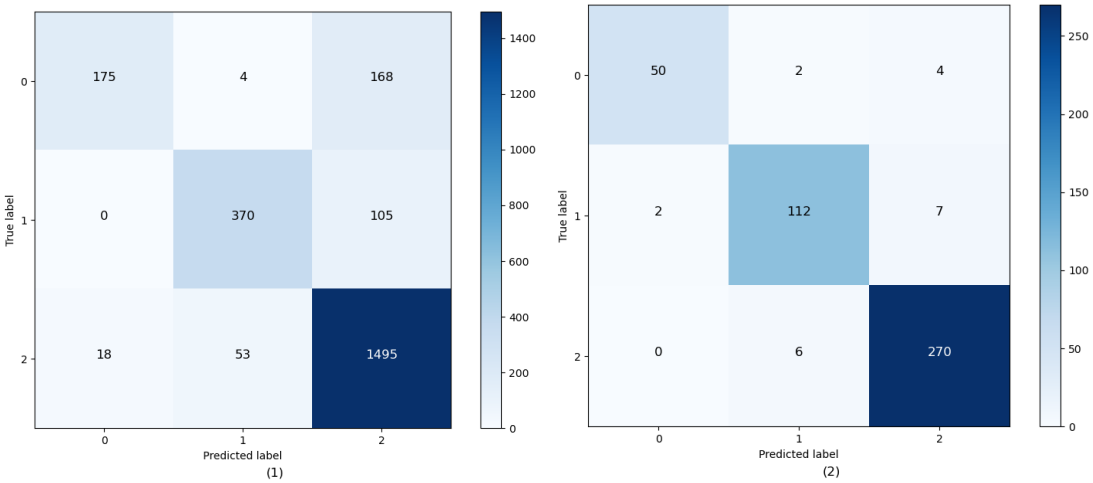

Figure 5.4 Confusion Matrix of Llama-2-7B on test data
of (1) Twitter Financial News and (2) Financial Phrase Bank

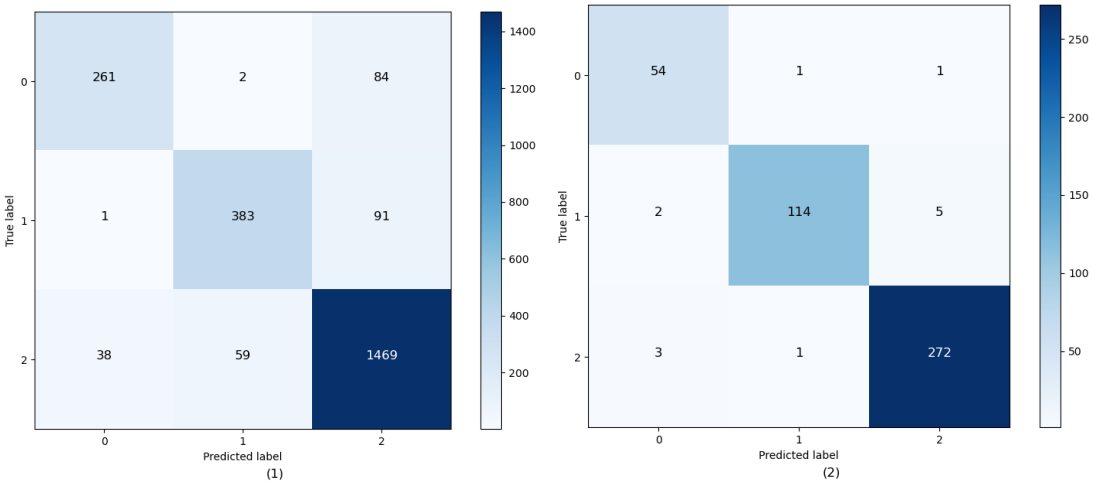

Figure 5.5 Confusion Matrix of Llama-2-7B-MLP on test data
of (1) Twitter Financial News and (2) Financial Phrase Bank

The confusion matrix for the Llama-2-7B model, classifying **Twitter Financial News** and data from the **Financial Phrase Bank**, is depicted in Figure 5.4. It is observable that an excessive number of texts with negative and positive sentiments were incorrectly classified as neutral, resulting in a classification performance that is significantly inferior to that of RoBERTa-Large.

It is evident from the matrices that the instances of positive and negative financial texts being misclassified as neutral have significantly decreased. This adjustment has led to a notable improvement in the overall classification accuracy compared to the performance prior to the adjustment.

Nevertheless, when evaluated in comparison with RoBERTa-Large, the adjusted Llama-2-7B model demonstrated micro and macro F1-scores that were found to be in close alignment with the performance of RoBERTa-Large. Nevertheless, it failed to exceed the recognition accuracy of RoBERTa-Large. The performance of the models on the two datasets is presented in Table 5.2.

	Micro-F1	**Macro-F1**
RoBERTa-Large-Twitter	0.904	0.882
Llama-2-7B-Twitter	0.854	0.788
Llama-2-7B-MLP- Twitter	0.885	0.852
RoBERTa-Large-Phrasebank	0.980	0.967
Llama-2-7B-Phrasebank	0.953	0.941
Llama-2-7B-MLP-Phrasebank	0.972	0.962

Table 5.2 Evaluation Metrics of RoBERTa-Large and Llama-2-7B on test data of Twitter Financial News and Financial Phrase Bank

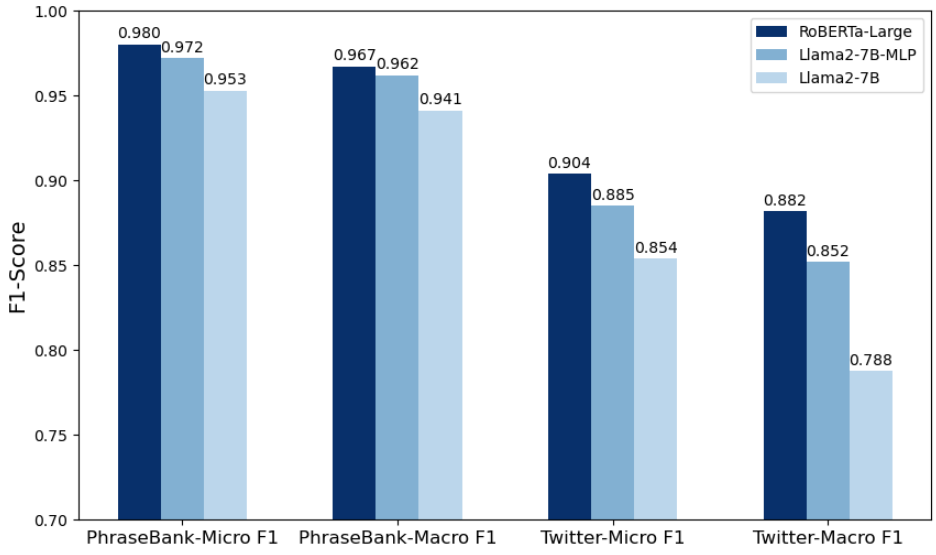

Figure 5.6 F1-scores of different models on the dataset of Twitter Financial News and Financial Phrase Bank

Figure 5.6 provides a more intuitive visualization of the comparative F1-score performance of the models when trained on different datasets. The accompanying confusion matrices reveal that both the RoBERTa-Large and Llama-2-7B-MLP models exhibit a significant number of misclassifications between neutral and polar sentiments on the **Twitter Financial News** dataset. In contrast, such misclassifications are markedly reduced on the **Financial Phrase Bank** dataset.

It is reasonable to conclude that the characteristics of tweets, such as their brevity, informal language, use of slang, emojis, and hashtags, play a significant role in the observed differences in classification accuracy. Unlike tweets, the texts in the **Financial Phrase Bank** are composed of two to three logically coherent sentences that clearly convey the intended meaning, therefore encapsulating more definitive sentiments. This clarity likely enhances the models' ability to detect sentiment, resulting in higher accuracy and better performance metrics for the model.

5.3 Mixing Training Data

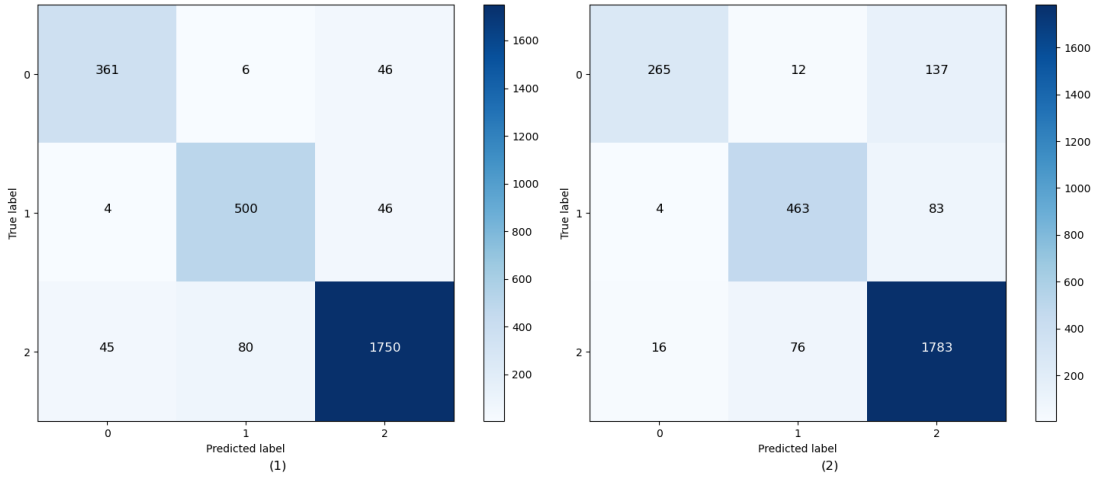

Figure 5.7 Confusion Matrix of (1) RoBERTa-Large and (2) Llama-2-7B-MLP

Building on the foundation of previous experiments, we further explored the performance of the RoBERTa-Large and Llama 2 when processing multilingual-style financial text inputs. To conduct this investigation, we combined and randomly shuffled data from **Twitter Financial News** and the **Financial Phrase Bank**. Subsequently, we redistributed this amalgamated dataset into a training set and a test set, maintaining an 8:2 ratio. Throughout multiple training iterations, we selected the model that exhibited the best classification performance. The confusion matrix representing its performance on the test set is depicted in Figure 5.7.

Table 5.3 presents the evaluation metrics for the RoBERTa-Large and Llama-2-7B-MLP models. After integrating mixed datasets, the performance of RoBERTa-Large exhibited a slight improvement when compared to the results observed during training solely on **Twitter Financial News**. However, this performance fell short of the

outcomes achieved with the **Financial Phrase Bank** dataset. In contrast, the performance of Llama-2-7B-MLP on the mixed dataset was not better than its performance on both two separate datasets. Figure 5.8 further presents these findings by providing a visual comparison of both models across different datasets.

	Micro-F1	Macro-F1
RoBERTa-Large-Twitter	0.904	0.882
Llama-2-7B-MLP- Twitter	0.885	0.852
RoBERTa-Large-PhraseBank	0.980	0.967
Llama-2-7B-MLP- PhraseBank	0.972	0.962
RoBERTa-Large-Mixed	0.920	0.899
Llama-2-MLP-Mixed	0.884	0.839

Table 5.3 Evaluation Metrics of RoBERTa-Large and Llama-2-7B-MLP on different datasets

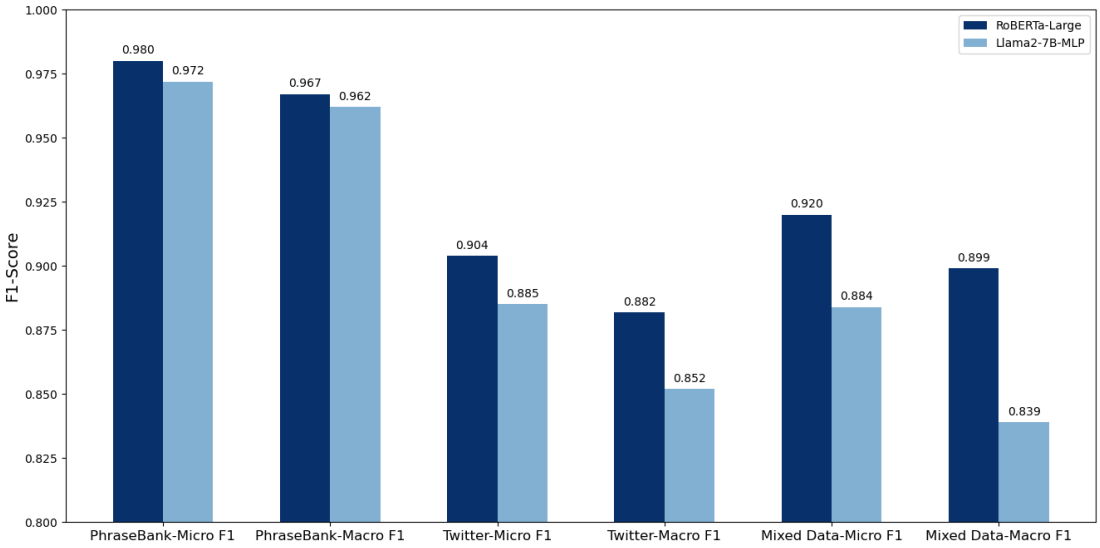

Figure 5.8 F1-scores of RoBERTa-Large and Llama-2-7B-MLP on different datasets

5.4 Reliability Evaluation and KOL Ranking

In our study, we selected the best-performing RoBERTa-Large and the Llama-2-7B models from multiple training iterations, for sentiment analysis on two distinct datasets: **Financial Tweets** and **Seeking Alpha Articles**. Utilizing the output from these analyses, we conducted a reliability evaluation of selected KOLs. Based on the reliability scores, we ranked these KOLs.

Figures 5.9 and Figure 5.10 illustrate the results derived from the sentiment analysis conducted using the RoBERTa-Large and Llama-2-7B-MLP models, respectively. The following figures display the TOP 3 KOLs in different time windows, ranked according to the reliability scores calculated from the sentiment analysis outputs.

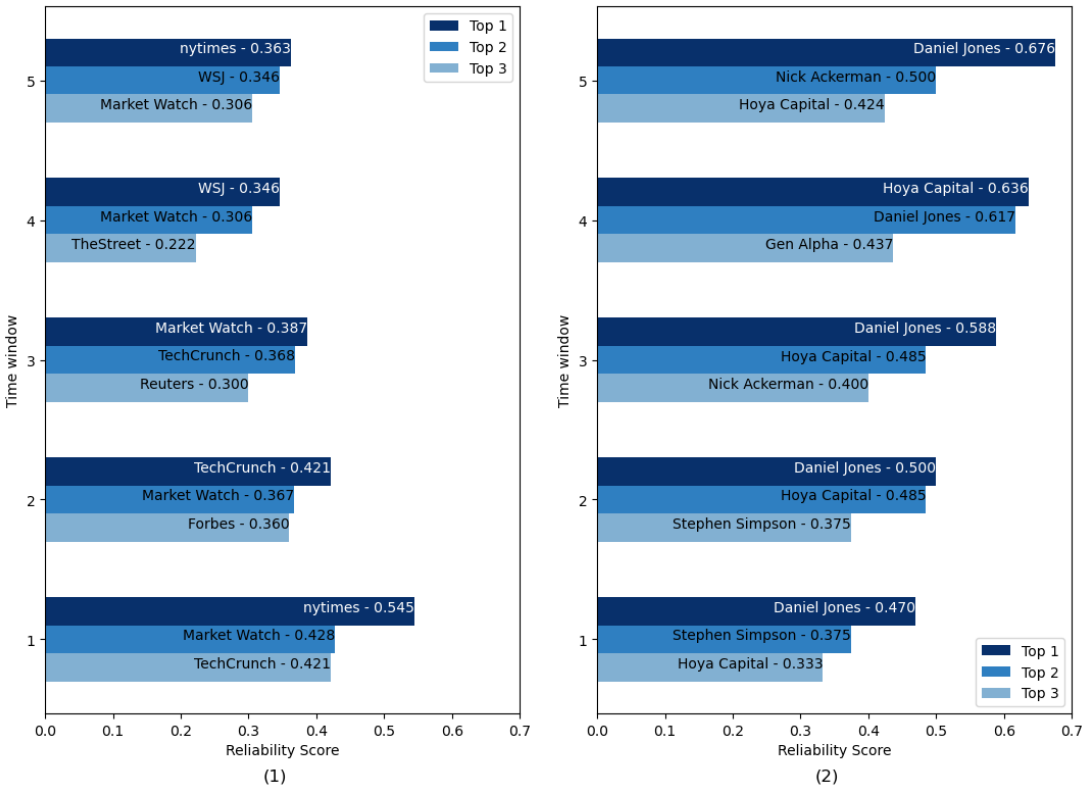

Figure 5.9 TOP 3 KOLs Rankings of (1) Twitter Financial News and (2) Financial Phrase Bank based on RoBERTa-Large

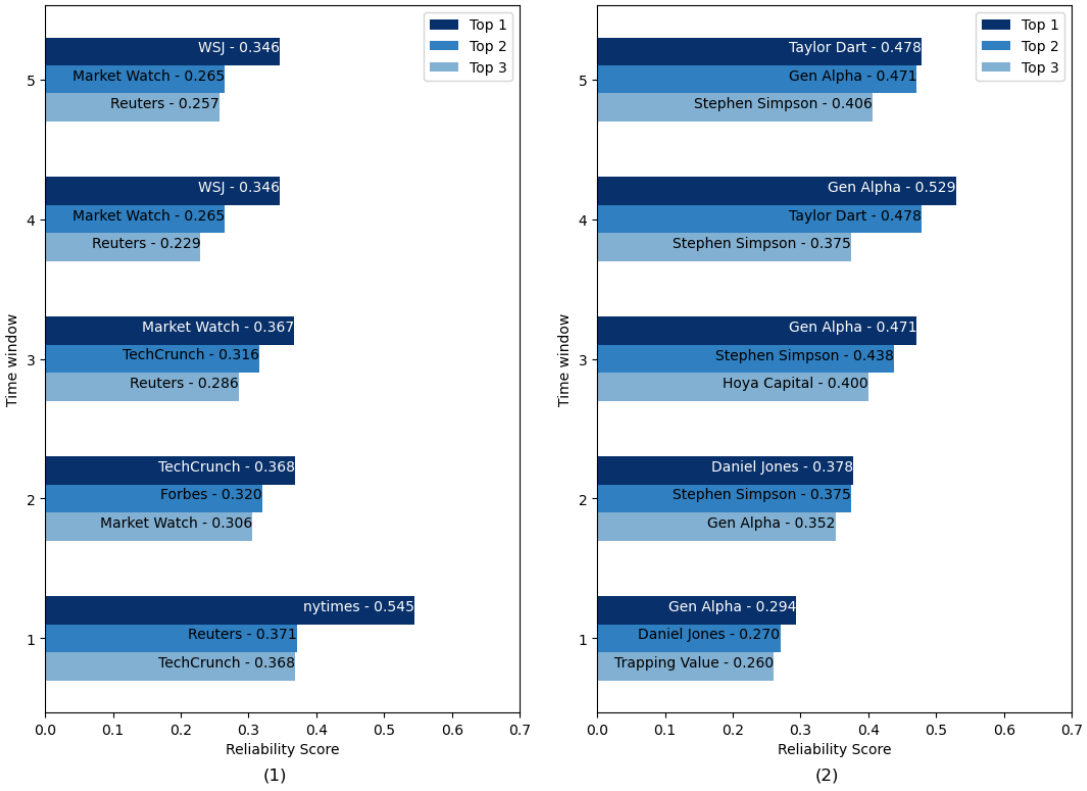

Figure 5.10 TOP 3 KOLs Rankings of (1) Twitter Financial News and (2) Financial Phrase Bank based on Llama-2-7B-MLP

From the ranking chart, it is observable that the overall credibility of authors in the **Seeking Alpha Articles** dataset surpasses that of authors in the **Financial Tweets** dataset. Furthermore, the credibility of these authors tends to increase over time with the expansion of the time window.

This trend suggests that the expectations of authors in **Seeking Alpha Articles** regarding long-term stock performance align closely with the actual stock price movements. In contrast, the sentiment expressed by authors in the **Financial Tweets** dataset correlates more closely with short-term stock price trends. This distinction is crucial for investors and analysts who rely on social media and financial journalism to make informed decisions about stock investments.

Upon comparing the results derived from RoBERTa-Large and Llama-2-7B-MLP, it is observed that although the lists of authors deemed highly reliability scores by both models are quite similar, the reliability scores of individual authors vary significantly across different time windows. Given the absence of a standardized benchmark for evaluating the quality of these reliability rankings, it is necessary to integrate these findings into practical stock price predictions to assess the credibility of the reliability rankings of KOLs.

5.5 Stock Prices Prediction and Analysis

Due to the scarcity of data in the **Seeking Alpha Articles** where the same author expresses opinions on the same stock at different time points, and the low overlap between the types of stocks covered in the **Seeking Alpha Articles** and those mentioned in **Financial Tweets**, our analysis primarily utilizes data from **Financial Tweets** for predicting and analyzing stock prices.

Additionally, there are a lot of stocks associated with KOLs opinions that have neutral sentiment, which hindered our ability to effectively predict stock price movements based on these tweets. As a result, our analysis concentrates on a select few stocks. These stocks are associated with multiple tweets from the same author, which exhibit distinctly positive or negative sentiments. This approach allows us to conduct a more targeted and meaningful analysis of the potential impact of sentiments on stock prices.

5.5.1 TSLA and CVX

In this section, we analyze the prediction of stock prices for TSLA and CVX.

Firstly, we employed a stock price prediction algorithm to forecast the prices of TSLA over the period from March 27, 2023, to March 17, 2023. The comparative results between the predicted prices from two different models and the actual stock prices are illustrated in Figure 5.11.

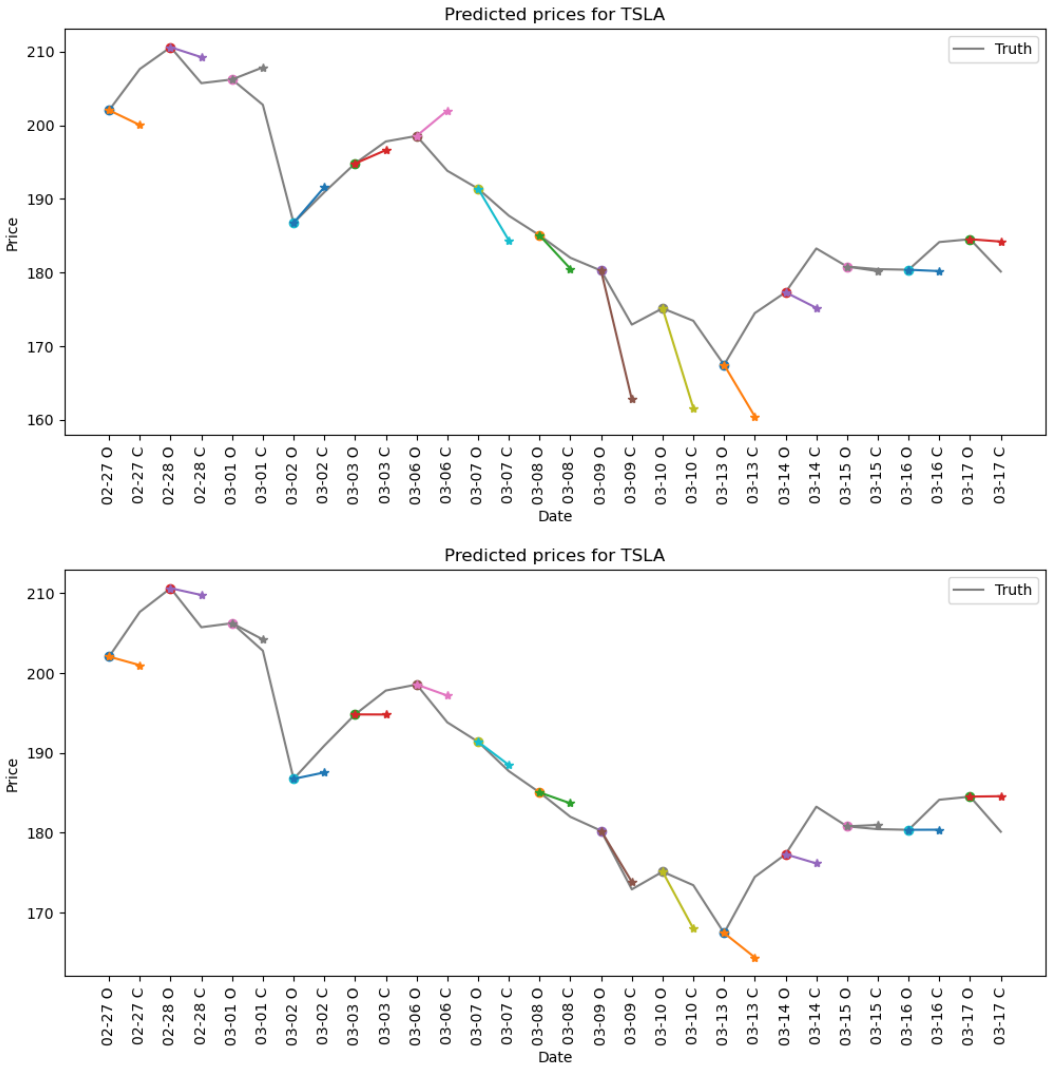

Figure 5.11 Analysis of TSLA stock price prediction. The top figure shows the analysis based on the output of the RoBERTa-Large model and the bottom figure shows the analysis based on the output of the Llama-2-7B-MLP model

It is observed that the outputs from the sentiment analysis of both models are remarkably consistent in this case study. According to the sentiment analysis, the stock price prediction algorithm was able to accurately forecast the trend of stock price changes for most of the period under review. However, there were variations in the specific amounts of price increases or decreases.

To compare the performance of two types of predictive results, we employed two statistical metrics: Mean Squared Error (MSE) and Mean Absolute Error (MAE). MSE measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.

$$MSE = \sum_{i=1}^{n}(y_i - y_i^p)^2 \quad (9)$$

MAE, on the other hand, measures the average magnitude of the errors in a set of predictions, without considering their direction. It provides a linear score that reflects the average size of the errors in a set of predictions, thus giving a straightforward

interpretation of the total error amount.

$$MAE = \sum_{i=1}^{n} |y_i - y_i^p| \quad (10)$$

The results of the evaluation, as shown in Table 5.4, indicate that the predictions derived from the sentiment analysis of the Llama-2-7B-MLP model are more aligned with the actual stock price trends. Compared to the predictions from the RoBERTa-Large model, Llama-2-7B-MLP demonstrates a smaller overall error.

	MSE	**MAE**
RoBERTa-Large	45.60	5.38
Llama-2-7B-MLP	21.59	3.90

Table 5.4 Comparison of Error Analysis on TSLA

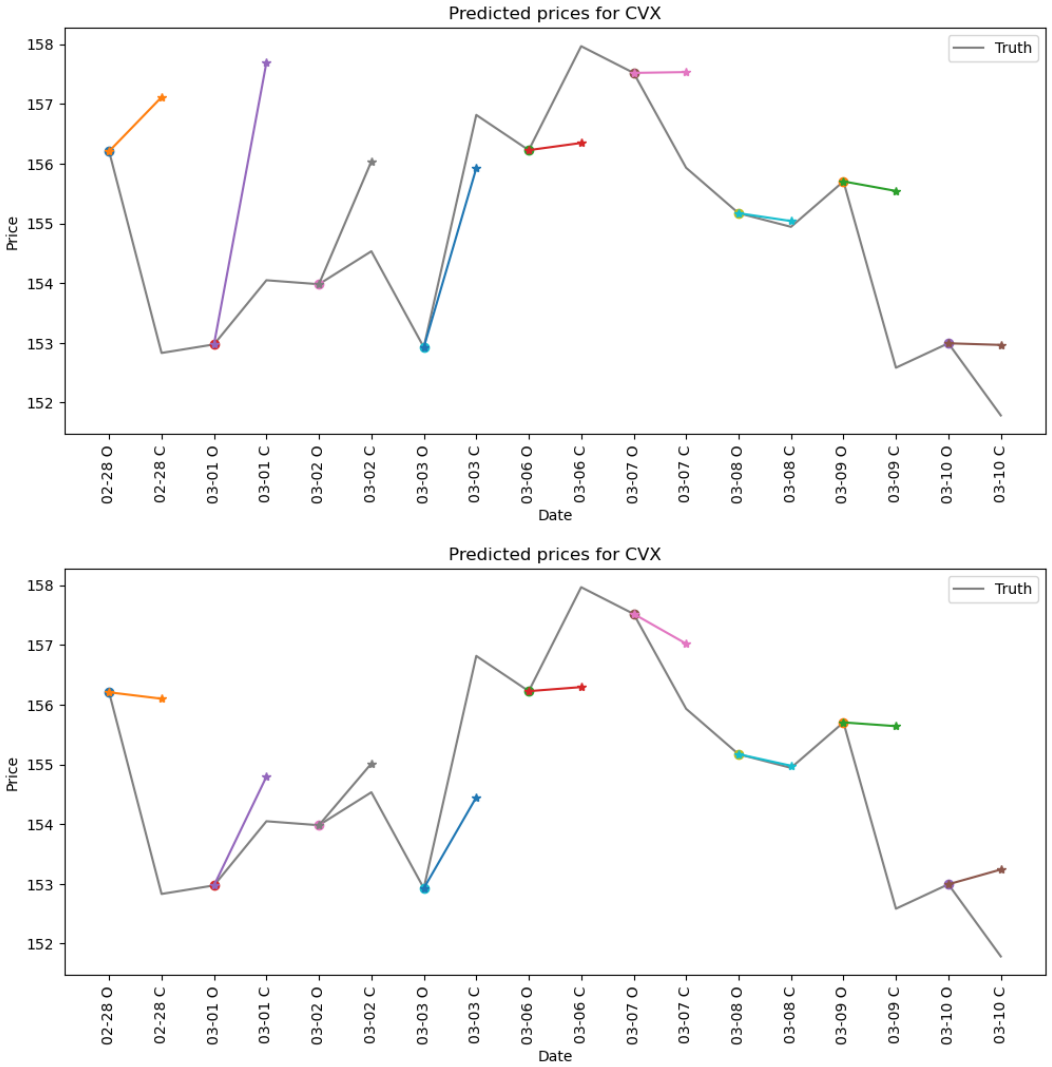

Figure 5.12 Analysis of CVX stock price prediction. The top figure shows the analysis based on the output of the RoBERTa-Large model and the bottom figure shows the analysis based on the output of the Llama-2-7B-MLP model

We also observed that multiple KOLs expressed opinions with either positive or negative sentiments regarding Chevron Corporation (CVX) over a specific period. Based on the timing of these expressed opinions, we conducted a predictive analysis of CVX's stock prices from February 28, 2023, to March 10, 2023. The outcomes of this analysis are illustrated in Figure 5.12.

Similar to the predictions for TSLA, both types of forecasts demonstrated a consistent judgment on the trend of stock price changes, accurately predicting the trajectory following the market opening to a certain extent. For stock price forecasting, the sentiment of KOLs and their reliability scores, as derived from the analysis using the RoBERTa-Large model, resulted in somewhat aggressive predictions. In contrast, predictions based on the Llama-2-7B-MLP model tended to be more conservative.

	MSE	**MAE**
RoBERTa-Large	5.93	2.07
Llama-2-7B-MLP	3.77	1.64

Table 5.5 Comparison of Error Analysis on CVX

Observations from the comparative stock price prediction charts suggest that the forecasts generated by the Llama-2-7B-MLP model align more closely with the actual stock price movements. The error analysis presented in Table 5.5 corroborates this assessment, indicating that the conservative approach of the Llama-2-7B-MLP model may provide a more accurate reflection of market dynamics.

5.5.2 NSC

In our analysis of the stock price prediction for NSC, we found a particularly interesting case. From February 26, 2023, to March 14, 2023, both the RoBERTa-Large and Llama-2-7B-MLP models predominantly classified the majority of social media comments made by several KOLs about this stock as neutral. However, a notable discrepancy was observed on March 9 and March 10, where the judgments of the two models diverged. This inconsistency evidently led to different predicted trends for NSC stock prices between March 9 and March 13.

Date	Text	RoBERTa	Llama 2
03-09	Norfolk Southern showed apparent lack of transparency in the days after a train carrying toxic chemicals?	neutral	negative
03-10	It was a disaster waiting to happen. Senators attacked Norfolk Southern response to last month toxic chemical.	neutral	negative

Table 5.6 Two specific tweets about NSC

To further investigate this phenomenon, we conducted a detailed analysis of the two specific tweets that contributed to the divergent model outputs. These two tweets are shown in Table 5.6, and we can see the difference between the two types of stock price forecasts affected by these two tweets in Figure 5.13.

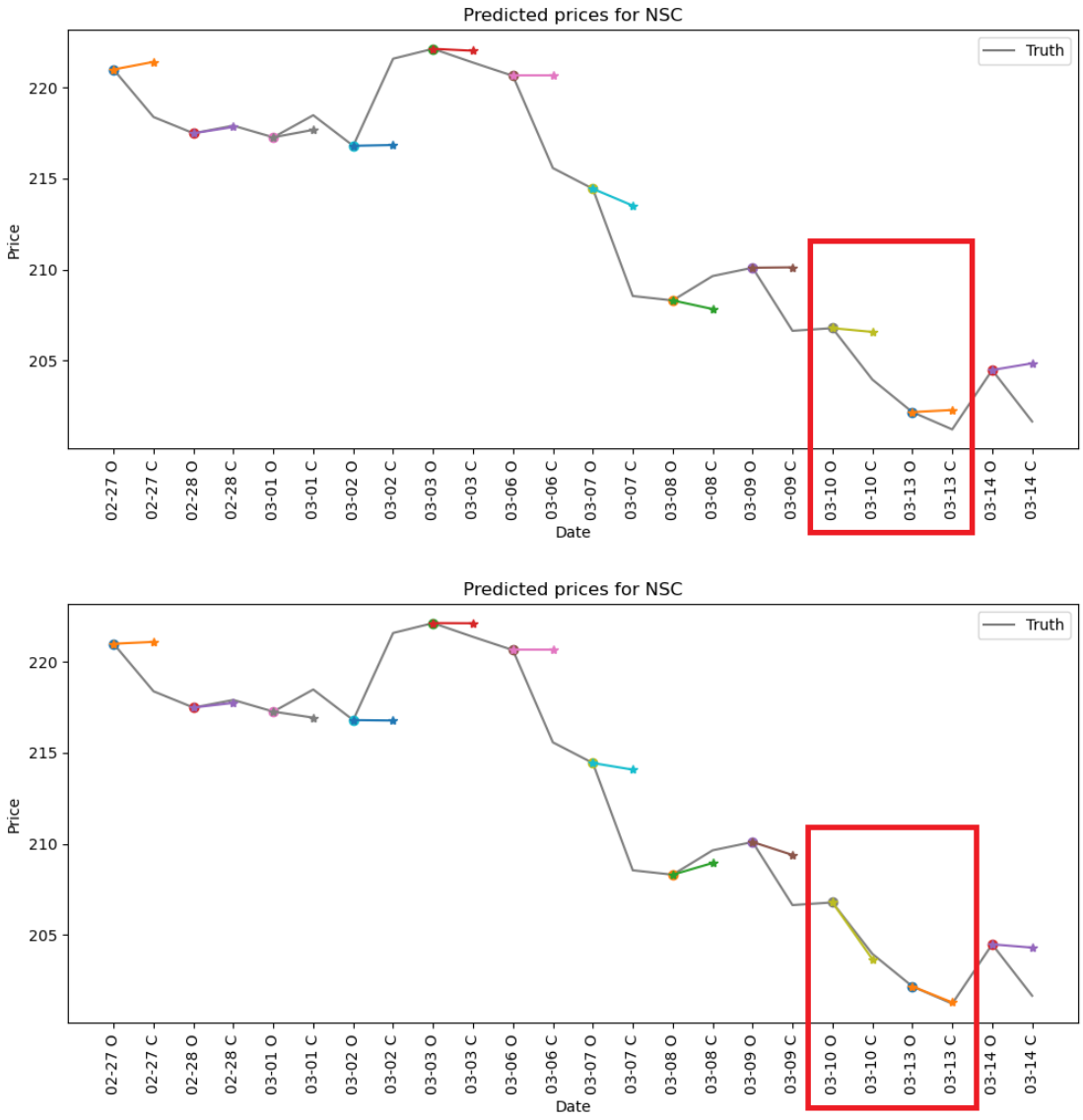

Figure 5.13 Analysis of NSC stock price prediction. The top figure shows the analysis based on the output of the RoBERTa-Large model and the bottom figure shows the analysis based on the output of the Llama-2-7B-MLP model

Through our analysis, it becomes evident that both tweets express criticism towards Norfolk Southern Corporation for its lack of transparency and inadequate response following a toxic chemical incident. Although the overall tone of the language used is straightforward and factual, in relation to the context of news events, the use of certain descriptive words clearly conveys a negative sentiment.

Combined with the analysis of stock price prediction, it's important to note that although the performance metrics suggest that Llama-2-7B is slightly less accurate than

RoBERTa-Large, this doesn't mean that Llama 2 is unreliable. The Llama-2-7B model not only analyzes the sentiment of text in tweets, but also considers contextual associations and relevant cultural and political factors that may impact the text. To a certain extent, Llama 2's judgement of the sentiment of the text is closer to the human way of thinking, as he makes associations and reasoning about humanistic and political kinds of factors. This is particularly crucial in financial analysis.

6 Conclusion

The primary focus of this study is to explore the use of pre-trained large language models for mining sentiment information contained in social media posts published by Key Opinion Leaders in the finance sector, and to perform quantitative analyses related to reliability ranking and stock price prediction. This paper makes two main contributions:

Firstly, we investigated the feasibility of applying Llama-2-7B model for sentiment analysis of social media texts written by financial KOLs. We used the RoBERTa-Large model, which currently excels in short text classification tasks, as a benchmark to evaluate the performance of the Llama-2-7B model on the same textual data. The results indicated that, after fine-tuning, the Llama-2-7B model, supplemented by an additional MLP layer for dimensionality reduction, achieved classification accuracy comparable to that of RoBERTa-Large. Furthermore, in subsequent evaluation analyses, we observed that Llama 2 was capable of deeper associations and reasoning with financial texts, considering underlying political and cultural factors to judge emotional tendencies, significantly enhancing the robustness of sentiment analysis.

Secondly, based on the sentiment classification results from a pre-trained large language model, combined with existing reliability ranking algorithms, we proposed a stock price prediction algorithm that leverages the sentiment of opinions made by KOLs. This algorithm aggregated the relevant statements published by multiple reliable KOLs and incorporated their reliability predictions across different time windows to forecast the daily closing prices of stocks. In the experimental analysis phase, this algorithm demonstrated commendable predictive performance on historical stock price data. This success to some extent validated the feasibility of using large language models to capture market sentiments for accurate stock price predictions, thereby substantiating the financial logic behind this approach.

7 Future Work

This study has come to an end for the time being, but I have the following two prospects for future research directions:

Firstly, in future studies, there is a critical need for more reliable sentiment-labeled financial social media texts to fine-tune large language models. Data remains the cornerstone of natural language processing, and high-quality data inputs are essential for training models that meet our expectations. The fine-tuning of large language models with this high-quality, sentiment-labeled data will enable more accurate interpretations and predictions of market movement.

Secondly, the stock price prediction algorithm presented in this paper is based on a preliminary concept using existing data. While this initial approach has provided valuable insights, there is substantial room for enhancement. For future research, it would be beneficial to integrate more specialized financial-economic considerations with KOLs' sentiments, such as incorporating macroeconomic indicators, and market sentiment indices. These indicators have a profound impact on market conditions and can significantly influence stock prices. By integrating these macroeconomic factors, the algorithm could better reflect the movement of the economic environment.

References

[1] Z. Wang, H. Liu, W. Liu, and S. Wang, “Understanding the power of opinion leaders’ influence on the diffusion process of popular mobile games: Travel Frog on Sina Weibo,” Computers in Human Behavior, vol. 109, p. 106354, Aug. 2020, doi: 10.1016/j.chb.2020.106354.

[2] R. J. Shiller, “Measuring Bubble Expectations and Investor Confidence.” National Bureau of Economic Research, Mar. 1999. doi: 10.3386/w7008.

[3] G. Wang et al., “Crowds on Wall Street: Extracting Value from Social Investing Platforms.” arXiv, Jun. 04, 2014. Accessed: Nov. 09, 2022. [Online]. Available: http://arxiv.org/abs/1406.1137.

[4] Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. 2013. Exploiting Topic based Twitter Sentiment for Stock Prediction. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 24–29, Sofia, Bulgaria. Association for Computational Linguistics.

[5] Zhao, Wanting, et al. "Sentiment analysis on weibo platform for stock prediction." Artificial Intelligence and Security: 6th International Conference, ICAIS 2020, Hohhot, China, July 17–20, 2020, Proceedings, Part I 6. Springer Singapore, 2020.

[6] Valle-Cruz, David, 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 14.1 (2022): 372-387.

[7] Darell, Johan, and Vivea Upadhyaya. "Leveraging LlaMA 2 for sentiment analysis." (2024).

[8] N. Kant, R. Puri, N. Yakovenko, and B. Catanzaro, “Practical Text Classification With Large Pre-Trained Language Models.” arXiv, Dec. 03, 2018. Accessed: Nov. 09, 2022. [Online]. Available: http://arxiv.org/abs/1812.01207

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv, May 24, 2019. Accessed: Nov. 09, 2022. [Online]. Available: http://arxiv.org/abs/1810.04805

[10] Y. Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach.” arXiv, Jul. 26, 2019. Accessed: Nov. 09, 2022. [Online]. Available: http://arxiv.org/abs/1907.11692

[11] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, “TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification.” arXiv, Oct. 26, 2020. Accessed: Nov. 09, 2022. [Online]. Available: http://arxiv.org/abs/

2010.12421

[12] H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and T. Zhao, “SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 2177–2190. doi: 10.18653/v1/2020.acl-main.197.

[13] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[14] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

[15] Insuasti, Jesus, Felipe Roa, and Carlos Mario Zapata-Jaramillo. "Computers' interpretations of knowledge representation using pre-conceptual schemas: An approach based on the bert and llama 2-chat models." Big Data and Cognitive Computing 7.4 (2023): 182.

[16] Pavlyshenko, Bohdan M. "Financial news analytics using fine-tuned Llama 2 GPT Model." arXiv preprint arXiv:2308.13032 (2023).

[17] Breitung, Christian, Garvin Kruthof, and Sebastian Müller. "Contextualized sentiment analysis using large language models." Available at SSRN (2023).

[18] N. A, “Zeroshot/twitter-financial-news-sentiment · datasets at hugging face,” zeroshot/twitter-financial-news-sentiment · Datasets at Hugging Face. [Online]. Available: https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment. [Accessed: 30-Mar-2023].

[19] Malo, Pekka, et al. "Good debt or bad debt: Detecting semantic orientations in economic texts." Journal of the Association for Information Science and Technology 65.4 (2014): 782-796.

[20] D. Wallach, “StockerBot.” Oct. 24, 2022. Accessed: Nov. 15, 2022. [Online]. Available: https://github.com/dwallach1/StockerBot

[21] “Seeking alpha API documentation free with API key & SDK: Rapidapi,” Rapid. [Online]. Available: https://rapidapi.com/apidojo/api/seeking-alpha. [Accessed: 30-Mar-2023].

[22] “Yfinance,” PyPI. [Online]. Available: https://pypi.org/project/yfinance/. [Accessed: 06-Feb-2023].

[23] “Financial Tweets | Kaggle.” https://www.kaggle.com/datasets/davidwallach/financial-tweets?select=stockerbot-export.csv (accessed Nov. 09, 2022).

[24] Ainslie, Joshua, et al. "Gqa: Training generalized multi-query transformer models from multi-head checkpoints." arXiv preprint arXiv:2305.13245 (2023).

[25] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

[26] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.

[27] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

[28] Li, Yuke, et al. "LoRa on the move: Performance evaluation of LoRa in V2X communications." 2018 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2018.

[29] Aghajanyan, Armen, Luke Zettlemoyer, and Sonal Gupta. "Intrinsic dimensionality explains the effectiveness of language model fine-tuning." arXiv preprint arXiv:2012.13255 (2020).

[30] Archinetai, "Archinetai/Smart-Pytorch: Pytorch – SMART: Robust and efficient fine-tuning for pre-trained natural language models.," GitHub. [Online]. Available: https://github.com/archinetai/smart-pytorch. [Accessed: 06-Feb-2023].

[31] "georgian-io/LLM-Finetuning-Toolkit" [Online]. Available: https://github.com/georgian-io/LLM-Finetuning-Toolkit