Spring 2024

Stock Price Trend Prediction using Emotion Analysis of Financial Headlines with Distilled LLM Model

Rithesh H. Bhat
University of Texas at Arlington

Stock Price Trend Prediction using Emotion Analysis of Financial Headlines with Distilled LLM Model

by

RITHESH HARISH BHAT

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

MASTERS THESIS

THE UNIVERSITY OF TEXAS AT ARLINGTON
May 2024

ABSTRACT

Stock Price Trend Prediction using Emotion Analysis of Financial Headlines with Distilled LLM Model

RITHESH HARISH BHAT

The University of Texas at Arlington, 2024

Capturing the volatility of stock prices helps individual traders, stock analysts, and institutions alike increase their returns in the stock market. Financial news headlines have been shown to have a significant effect on stock price mobility. Lately, many financial portals have restricted web scraping of stock prices and other related financial data of companies from their websites. In this study we demonstrate that emotion analysis of financial news headlines alone can be sufficient in predicting stock price movement, even in the absence of any financial data. We propose an approach that eliminates the need for web scraping of financial data. We use API based mechanism to retrieve financial news headlines. In this study we train and subsequently leverage light and computationally fast Distilled LLM Model to gather emotional tone and strength of financial news headlines for companies. We then use this information with several machine learning-based classification algorithms to predict the stock price direction based solely on the emotion analysis of news. We demon- strate that emotion analysis-based attributes of financial news headlines are as accurate in predicting the price direction as running the algorithms with the financial data alone.

To my mother Sharada Bhat and my father Harish Bhat

who gave me the wings to fly

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervising professor, Dr. Bhanu Jain, for her constant motivation, encouragement, and invaluable advice throughout my graduate studies.

I wish to thank my thesis academic advisor, Ginger Dickens, for guiding me on this thesis journey.

I extend my appreciation to Dr. Huber and Dr. Vassilis for their interest in my research and for taking the time to serve on my dissertation committee.

I would also like to extend my appreciation to Rely Health for providing invaluable experience in the LLM and AI domain during my CPT tenure. I wish to thank Prithvi Narasimhan and Soham Moore with Rely Health for their support and encouragement.

Sincere gratitude to all my managers and leads when I started my professional journey. I wish to thank Dennison Solomon, Neelakantha Subudhi, Ashish Kumar, Manikishore Sannareddy, Shrikant Jagtap, and Niteesh Gupta.

I am grateful to all the teachers who taught me during the years I spent in school, undergraduate studies in India, and Masters in the United States. I would like to thank Swathi Nayak for encouraging me to pursue graduate studies.

Finally, I express my deep gratitude to my brother for encouraging me during my undergraduate and graduate studies. I am also extremely grateful to my mother and father for their sacrifice, encouragement, and patience. I also thank several of my friends who have helped me throughout my career.

April 23, 2024

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

LIST OF ABBREVATIONS

1. NLP (Natural Language Processing)
2. ML (Machine Learning)
3. BERT (Bidirectional Encoder Representation from Transformers)
4. LLM (Large Language Model)

CHAPTER 1

INTRODUCTION

1.1 Introduction and Problem Statement

In recent years, the realm of stock market analysis and prediction has garnered considerable attention and interest. However, due to the non-linear nature of stock market volatility and dynamics, conventional prediction methods face challenges, as outlined in the Efficient Market Hypothesis (EMH) and Random Walk Theory (RWT) [2] [8] [9].

The discourse surrounding stock price prediction has been extensive and contentious, with contrasting viewpoints stemming from EMH [4]. According to EMH, stock prices incorporate all available information, rendering predictions based solely on historical data impractical. Conversely, RWT posits that stock market prices follow a random pattern, making prediction unfeasible.

With the advent of social media and online platforms, news dissemination has become rapid and widespread. In the financial domain, significant news events can swiftly impact stock prices, particularly during trading hours. Researchers have explored the correlation between investor sentiments expressed in online forums and stock price movements [12]. Moreover, sentiments expressed in news beyond the financial sector have also been leveraged for analysis. For instance, during the 2020 pandemic, researchers analyzed social media news to gain insights for appropriate public health responses by classifying sentiments (positive, negative, neutral) and summarizing topics using TF-IDF [14].

In our research, we employ machine learning classification models to forecast the direction of stock price trend. Specifically, we utilize Logistic Regression, Artificial Neural Networks, and the Random Forest algorithm. Our findings demonstrate that solely relying on emotion analysis-based attributes can yield prediction results comparable in accuracy to those obtained exclusively from financial attributes in classification experiments.

In our research, our primary contributions include the following:

- Utilizing APIs from financial aggregators to generate the necessary dataset for forecasting stock prices, thus eliminating the requirement for web scraping to compile a financial dataset.
- Demonstrating and applying the process of fine-tuning a pre-trained Large Language Model (LLM) to accurately predict emotions associated with financial news headlines.
- Employing a Distilled LLM model for text classification tasks rather than conventional Natural Language Processing (NLP) techniques, specifically for analyzing financial news.
- Employing classification algorithms separately on emotion analysis attributes and financial attributes to forecast stock price direction.
- Evaluating and addressing the limitations and challenges encountered in our approach.

  In essence, the combination of the Distilled LLM Model, emotion analysis of news headlines, and machine learning classification algorithms offers a promising strategy in predicting stock price trends. Our methodology offers an alternative approach to forecasting future stock prices by considering the emotional content embedded in financial news headlines, diverging from the traditional reliance solely on financial data for stock price prediction.

1.2 Structure of Thesis

This thesis is comprised of 7 chapters, each of which addresses a specific aspect of the research. Chapter 1 introduces the problem statement and provides an overview of the analysis while chapter 2 explores the existing body of work within the domain of stock price analysis. Chapter 3 thereafter details the methodologies employed to collect financial news and related attributes from aggregators like newsapi.org and AlphaVantage.

In Chapter 4, the focus shifts to Emotion Analysis, exploring the utilization of LLM models to extract emotions from financial news, along with discussions on fine-tuning these models. Chapter 5 is dedicated to presenting the experiments conducted and the results obtained from the classifier algorithms.

Chapter 6 explores potential use cases where LLM models can be applied for emotion analysis. Finally, Chapter 7 lays out the conclusions drawn from the research and outlines avenues for future work.

CHAPTER 2

RELATED WORK

Our research draws inspiration from [13] that investigates methodologies for enhancing stock market prediction through machine learning techniques. The referenced study explores a range of traditional and contemporary algorithms, including linear regression, Random Walk Theory (RWT), Moving Average Convergence / Divergence (MACD), as well as machine learning models such as Support Vector Machine (SVM), Random Forest (RF), and various neural network architectures.

In particular, the authors adopt a comprehensive approach by integrating both Artificial Neural Network (ANN) and Random Forest (RF) models to predict the closing prices of stocks. Leveraging historical data spanning a decade from prominent companies like Nike, Goldman Sachs, Johnson and Johnson, Pfizer, and JP Morgan Chase and Co., sourced from Yahoo Finance, they engineer a novel set of input features derived from Open, High, Low, and Close prices. These additional indicators aim to significantly enhance the accuracy of the predictive models.

The effectiveness of the proposed approach is evaluated using two key performance metrics: Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE).

Overall, this research paper serves as a foundational reference for exploring how machine learning techniques, particularly ANN and RF models, can be leveraged to improve stock market prediction accuracy.

We drew inspiration from this research paper and embarked upon a similar journey by employing ANN and random forest just like the experiments in it. We used the additional input features available in the dataset in our experiments and thereby made our models more robust and also achieved a lower root mean square error (RMSE) score than one quoted in the reference paper.

CHAPTER 3

OVERVIEW OF DATASET AGGREGATION

3.1 Identification of Stocks

In this research, we focused on a curated list of 32 Mega Cap companies based in the United States, characterized by a market capitalization exceeding 200 billion USD. We specifically chose these companies due to the abundance of news articles available about them, facilitating an in-depth analysis of the information conveyed in news headlines and its correlation with each company's price trends. Subsequently, we collected data in two main areas related to these companies: financial news coverage and key financial metrics such as opening price, closing price, trading volume, and daily high and low prices of the stocks. This data was systematically stored in a database for further analysis using the distilled LLM model and machine learning algorithms. The subsequent sections provide detailed insights into the methodology employed to extract this information from two different aggregators.

Table 3.1. Ticker and Company Name

Ticker	**Company Name**
AAPL	Apple
MSFT	Microsoft Corporation Common Stock
GOOGL	Alphabet Inc. Class A Common Stock
AMZN	Amazon.com, Inc. Common Stock
META	Meta Platforms, Inc. Class A Common Stock
TSLA	Tesla, Inc. Common Stock
UNH	UnitedHealth Group Incorporated Common Stock (DE)
V	Visa Inc.
WMT	Walmart Inc. Common Stock
JPM	JP Morgan Chase & Co. Common Stock
XOM	Exxon Mobil Corporation Common Stock
AVGO	Broadcom Inc. Common Stock
MA	Mastercard Incorporated Common Stock
JNJ	Johnson & Johnson Common Stock
ORCL	Oracle Corporation Common Stock
HD	Home Depot, Inc. (The) Common Stock
ADBE	Adobe Inc. Common Stock
CVX	Chevron Corporation Common Stock
MRK	Merck & Company, Inc. Common Stock (new)
COST	Costco Wholesale Corporation Common Stock
KO	Coca-Cola Company (The) Common Stock
PEP	PepsiCo, Inc. Common Stock
BAC	Bank of America Corporation Common Stock
CSCO	Cisco Systems, Inc. Common Stock (DE)
CRM	Salesforce, Inc. Common Stock

3.2 Techniques for Financial News Extraction

With the rise of GPT, similar large-scale LLM models, and the state-of-art technical capabilities to scrape the internet at a rapid pace, many news platforms have introduced rate limits on their website, resulting in the blocking of incoming IP addresses. In some cases, platforms have updated their policies

to prohibit the use of automated web scrapers since early 2023. To navigate this challenge while collecting news related to the shortlisted Mega caps, we adopted an alternative approach: retrieving news via APIs from official news aggregators instead of scraping the news sites. Various news aggregator platforms offer news services in both free and paid versions. In this paper, we opted for newsapi.org as our aggregator of choice. This platform aggregates articles from around the world and provides headlines via an API-based mechanism. To access this service, we created an account on their platform to obtain the API key, which needs to be passed as a header attribute when making API calls. The platform grants users 100 free API requests per day and provides access to worldwide news coverage.

Adapting to the changing landscape of online data accessibility, particularly in response to restrictions imposed by news platforms, has been pivotal in ensuring the continuity and reliability of our data collection process. By leveraging official news aggregators like newsapi.org, we not only circumvent potential IP blocking issues but also gain access to a broader range of news sources from across the globe. This strategic shift from web scraping to API-based retrieval not only aligns with evolving industry standards but also enhances the efficiency and scalability of our data gathering efforts. Furthermore, the availability of free API requests and the extensive coverage of worldwide news offered by the chosen aggregator empower us to maintain a comprehensive and up-to-date dataset, essential for conducting rigorous analysis in the realm of financial news emotion analysis.

Using newsapi.org as our main news aggregator emphasizes the importance of relying on established platforms that follow industry best practices. Partnering with a trusted service provider helps us avoid unreliable or unauthorized

```python
class NewsAPI:
    Rithesh
    def __init__(self, app: App):
        self.news_api_client = NewsApiClient(app.newsapi_key)
        self.app = app

    Rithesh
    def headlines(self):
        print(self.news_api_client.get_top_headlines(category='business', country='us', language='en'))
```

Figure 3.1. Code to extract news from newsapi.org public API.

sources, ensuring the credibility of the news articles we analyze. The API-based approach not only improves data retrieval reliability but also integrates smoothly with our existing data processing pipeline. This efficient workflow lets us collect, store, and analyze a vast amount of news data, providing valuable insights into the correlation between news sentiment and stock performance. Our adoption of API-based news retrieval is a strategic response to evolving data collection challenges, helping us extract maximum value from digital information while adhering to ethical and legal standards.

3.3 Techniques for fetching financial attributes related to the stocks

In gathering financial data, we faced challenges with web scraping similar to those described earlier. To overcome this, we used Alpha Vantage, a leading financial technology company, to retrieve stock price information. Alpha Vantage offers access to real-time and historical data spanning up to two decades, making it valuable for our research. Users need to create an account and get a unique API key, which must be included with each information request. This key ensures authorized access to the platform's data. We focused on retrieving daily stock price data and related attributes, including open and close prices,

daily highs and lows, and trading volume. We also collected annual and quarterly earnings reports for each firm.

Navigating data acquisition complexities in finance required a strategic approach prioritizing reliability, accuracy, and compliance. Alpha Vantage offers a comprehensive suite of stock market information tailored to our research needs. The platform's capabilities, including its extensive historical data archive and real-time data feeds, helped us construct a dataset covering key financial metrics and performance indicators for the selected Mega cap companies. Integrating Alpha Vantage's API into our data retrieval pipeline streamlined the data retrieval process, thus enabling efficient gathering and storage of financial data for analysis.

Incorporating earnings reports into our data collection framework added an additional layer of granularity to our analysis, allowing us to assess not only stock price movements but also fundamental factors driving corporate performance. By capturing both annual and quarterly earnings data for each firm, we aimed to gain deeper insights into their financial health and trajectory over time. This holistic approach to data collection underscores our commitment to conducting rigorous research grounded in comprehensive and reliable data sources. Moving forward, our collaboration with Alpha Vantage will continue to play a pivotal role in supporting our analysis of the relationship between financial attributes and news sentiment, ultimately enhancing our understanding of market dynamics and informing investment decision-making processes.

3.4 Libraries used during dataset collection

We employed two distinct aggregators, each with its own Python package equipped with utility functions tailored for making API calls. To streamline our data re-

trieval process, we devised a wrapper class that internally utilizes the respective Python packages. For news data, we utilized the "newsapi" package and invoked the "get_top_headlines" function within the business category specifically for the United States. From the API response, we meticulously extracted several key attributes, including the news source, headline, URL, URL to image (if available), publication date, description, and content of the article. These extracted pieces of information were meticulously processed and subsequently stored in a PostgreSQL database for further analysis.

In the case of stock price-related data, we leveraged the "alphavantage" Python package provided by Alpha Vantage. This package offers utility functions tailored for accessing the required financial data. Notably, the "historical_data" utility method proved to be particularly valuable, as it enabled us to retrieve historical stock data spanning up to two decades in a single API call. This efficient use of API calls streamlined our data collection process, allowing us to access a comprehensive dataset with minimal resource consumption. However, it's worth noting that while the package provided robust functionality for retrieving stock data, it lacked specific methods for fetching earnings reports for the firms under study. Consequently, we had to resort to the conventional method of making explicit API calls to retrieve this information.

For this research, our data collection efforts commenced from August 2023 onwards. By initiating data extraction from this point in time, we aimed to capture a relevant and contemporary dataset that aligns with the temporal scope of our study. This strategic decision ensured that our analysis encompasses recent market trends and developments, providing valuable insights into the dynamics of news sentiment and stock performance within the specified timeframe.

```
def store_compact_historical_data(self):
    symbols = self.app.get_symbols_from_stock_list(self.category)

    for i in symbols:
        ts = TimeSeries(key=self.app.vantage_key)
        d, meta_data = ts.get_daily(i, outputsize="full")
```

Figure 3.2. Leveraging Alphavantage libraries for financial attribute extraction.

3.5 Historical news dataset collection

The constraints imposed by the free tier of our chosen news aggregation platform prompted us to seek alternative solutions for accessing historical data extending beyond a one-month timeframe. To address this challenge, we identified a valuable resource on Kaggle— a renowned platform for sharing datasets—that offered a comprehensive dataset encompassing information on over 6,000 stocks [1]. This dataset spanned an extensive timeframe from 2009 to 2020, providing a wealth of historical data that complemented our news aggregation efforts. By leveraging this robust repository of information, we significantly enhanced the depth and breadth of our analysis, allowing us to gain insights into long-term trends and patterns in stock market dynamics. This strategic decision to augment our data collection efforts with supplementary historical data from Kaggle underscores our commitment to conducting thorough and comprehensive research, despite the limitations posed by the free tier of our news aggregation platform. Moving forward, the integration of this additional dataset will enable us to conduct a more nuanced and informed analysis of the interplay between news sentiment and stock performance, ultimately enriching the quality and relevance of our research findings.

3.6 Challenges faced during dataset collection

The free tier of NewsAPI offers the capability to fetch historical news for a limited duration of up to one month and imposes a restriction of only 100 free API requests per day. Beyond this one-month timeframe, accessing historical news data requires subscribing to the paid plan. One notable limitation of the NewsAPI responses is the inability to retrieve the entirety of the article text for the Description and Content sections. This constraint poses challenges for performing comprehensive sentiment analysis on the article content, as the full context is often missing when only partial information is available. Analyzing these attributes is crucial for detecting any discrepancies between the headlines and the actual content of the articles, which could potentially lead to false positives or misinterpretations.

Similarly, the free tier of AlphaVantage provides a limited allowance of 25 API requests per day. This restriction means that we could only monitor stock prices for up to 25 firms per day using the free tier. While AlphaVantage offers valuable financial data services, including historical stock price data spanning up to two decades, the limitations imposed by the free tier necessitated careful management of API requests to ensure optimal utilization of resources. Despite these constraints, leveraging the free tiers of both NewsAPI and AlphaVantage enabled us to access valuable data for our research, albeit with some limitations on data availability and API request quotas. Moving forward, consideration of these limitations will influence our data collection strategies and analysis methodologies, ensuring that we maximize the utility of the available resources while maintaining the integrity and reliability of our research findings.

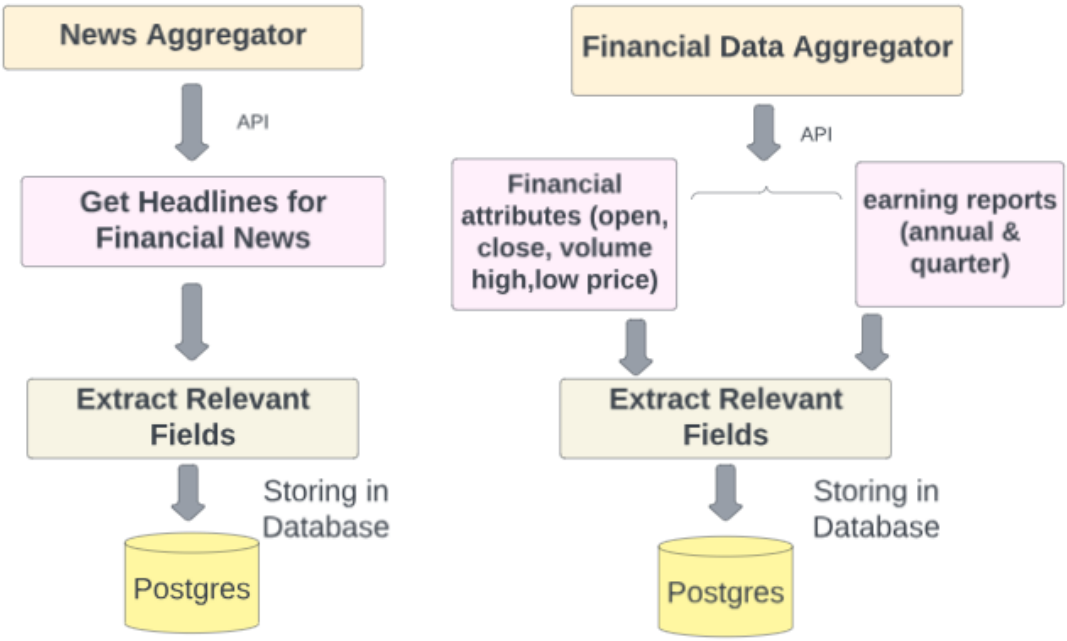

Figure 3.3. Data extraction overview with 2 aggregators.

CHAPTER 4

EMOTION ANALYSIS

Emotion Analysis, a process aimed at discerning emotions from textual input, serves as a powerful tool in understanding the nuanced sentiments embedded within financial news headlines. In our research endeavor, we opted for Emotion Analysis over Sentiment Analysis due to its ability to offer finer granularity and additional depth in capturing the spectrum of emotions expressed in the text. Our methodology involves extracting financial news headlines and meticulously storing them in our database for subsequent analysis. The essence of our approach lies in accurately predicting the diverse array of emotions embedded within these headlines, a task that traditionally involves a series of Natural Language Processing (NLP) techniques.

The conventional approach to emotion prediction typically entails a multi-step process, including tokenization, stemming, removal of stop words, and TF-IDF (Term Frequency-Inverse Document Frequency) analysis, followed by labeling the data with appropriate emotional categories. Subsequently, Machine Learning (ML) models such as Naïve Bayes or Support Vector Machines (SVM) are trained on this labeled dataset, with additional steps such as hyperparameter tuning and post-processing often necessary to optimize model performance.

However, in contrast to the traditional methodology, our approach diverges by leveraging the capabilities of Distilled Large Language Models (LLM) for text classification. Rather than relying on a suite of NLP techniques to extract emotions and their associated strengths from news headlines, we harness the

power of advanced LLM models. These models, pre-trained on vast amounts of textual data, exhibit a remarkable capacity to comprehend and contextualize natural language, thereby obviating the need for manual feature engineering or extensive preprocessing steps.

By employing Distilled LLM models for emotion analysis, we circumvent the complexities associated with traditional NLP-based approaches while achieving comparable, if not superior, levels of accuracy and efficiency. The inherent adaptability and scalability of LLM models enable us to effectively capture the nuanced nuances of emotion expressed within financial news headlines, providing valuable insights into market sentiment and investor behavior. This innovative methodology represents a paradigm shift in the realm of text classification, empowering researchers to unlock deeper layers of meaning within textual data with unprecedented precision and efficacy. As we continue to refine and optimize our approach, we anticipate further advancements in our ability to discern and interpret emotions in financial news with ever-increasing fidelity and insight.

4.1 Strategies to choose the appropriate LLM Model

Numerous public platforms offer access to Large Language Model (LLM) capabilities, such as OpenAI and Claude, but with certain limitations that warrant consideration. Primarily, these platforms typically operate on a paid subscription basis, with free tiers offering a limited number of requests. Moreover, as these requests are processed on cloud servers, end users have limited control over the processing environment and infrastructure. Additionally, users are not aware of the biases of the model and lack the ability to fine-tune it to adapt to specific domains or mitigate biases effectively. Furthermore, general-purpose LLM models, renowned for their versatility, tend to be larger in size and may

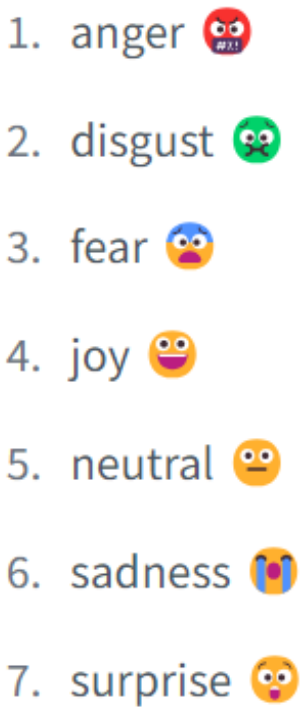

Figure 4.1. Emotions supported in Distilled text classification LLM Model.

exhibit slower performance compared to Distilled LLMs, which are specifically tailored for particular tasks.

To overcome these challenges, we opted to download a Distilled LLM model, often referred to as a lightweight model, to our local system. Subsequently, we trained and customized the base LLM model to suit the requirements of domain-specific text classification tasks. To identify the most suitable LLM model, we turned to the Hugging Face platform which hosts open-source LLM models that users can fine-tune and customize to meet their specific application needs. This platform also facilitates the sharing of custom LLM models among users, fostering collaboration and knowledge exchange within the community.

The adoption of a pre-trained LLM model for text classification yields significant advantages, the most important one among them being the elimination of the need to reinvent the wheel. By leveraging existing LLM models, such as "emotion-english-distilroberta-base," [6] a fork from RobertaBase [10], we overcome the need to perform intricate NLP tasks from scratch. This base LLM model, tailored specifically for text classification tasks, enables us to focus our efforts on domain-specific customization and fine-tuning, thereby expediting the model development process.

Notably, the distilled nature of LLM models ensures that they are optimized for specific tasks, resulting in lighter computational overhead and faster execution times compared to their general-purpose counterparts. In our case, the chosen model supports Ekman's six basic emotions 4.1 anger, disgust, fear, joy, sadness, surprise—alongside a neutral class, facilitating nuanced emotion analysis of financial news headlines. This tailored approach not only enhances the efficiency and accuracy of our emotion analysis but also underscores the versatility and adaptability of LLM models in addressing diverse text classification tasks with precision and efficacy.

4.2 Training Data Used for Base LLM

The RoBERTa-base model, renowned for its robust performance in natural language processing tasks, underwent extensive pretraining on a vast amalgamation of five datasets, collectively totaling close to 160 GB of text [7]. These datasets include BookCorpus, OpenWebText (a recreation of the dataset used to train GPT-2), English Wikipedia, CC-News (a dataset comprising 63 million English news articles from September 2016 to February 2019), and Stories. During pretraining, the model was subjected to the Masked Language Modeling

(MLM) objective, wherein a portion of the input sentence—approximately 15% is randomly masked, and the model is tasked with predicting the masked words within the context of the sentence [7].

Subsequently, the base RoBERTa model underwent a process of distillation to produce the DistilRoBERTa-base model, following a similar procedure outlined in [11]. This distillation process leverages knowledge distillation techniques during the pretraining phase, resulting in a reduction in model size by approximately 40%, while retaining 97% of its language understanding capabilities and achieving a 60% improvement in processing speed.

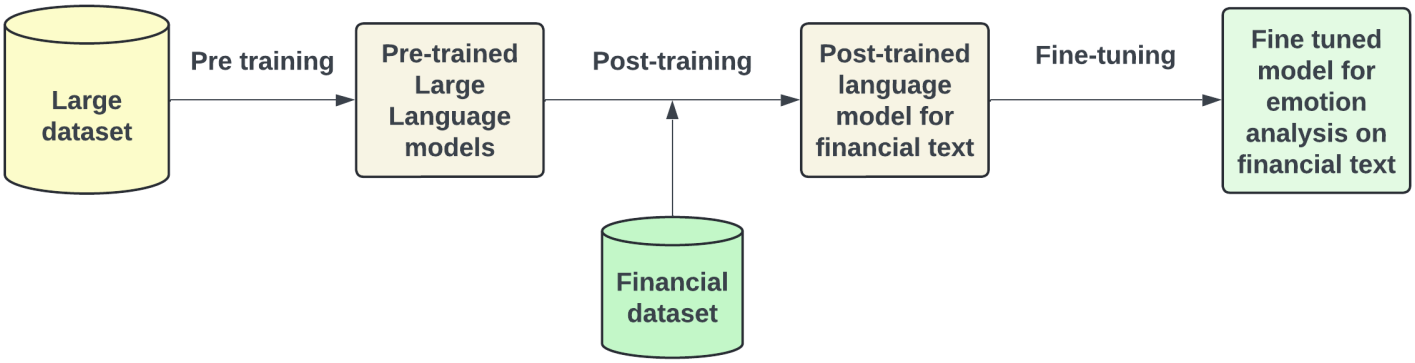

Figure 4.2. Fine tuning Distilled LLM Model to predict emotions behind financial text.

Having selected the fine-tuned checkpoint of the DistilRoBERTa-base LLM model [6] for our specific use case, our next objective was to optimize its performance in predicting the emotions expressed in financial news articles. This optimization process, known as fine tuning a pre trained model, involves manually labeling the text with appropriate emotion labels and subsequently training

the base model on this labeled dataset. By exposing the LLM model to domain-specific verbiage and emotion-laden textual data, fine tuning a pre trained model enables the model to better understand and predict similar texts with significantly enhanced accuracy. 4.2.

This iterative process of fine-tuning and optimization ensures that the DistilRoBERTa-base LLM model is finely attuned to the nuances of financial news language and emotions, thereby facilitating more accurate and insightful emotion analysis. The utilization of fine tuning a pre trained model techniques further enhances the model's predictive capabilities, enabling it to discern and interpret subtle emotional cues embedded within financial news headlines with precision and efficacy. This meticulous approach to model refinement underscores our commitment to harnessing cutting-edge technologies and methodologies to extract meaningful insights from textual data in the realm of finance.

4.3 Fine Tuning the LLM model for Financial Text

We embarked on a process of fine tuning a pre trained model , wherein we took the Large Language Model (LLM) [6] and trained it using input features comprised of text and corresponding labels representing all seven emotions supported by the LLM model [15]. Despite working with a relatively small training dataset, the improvements in the model's performance were striking. Specifically, we conducted training using a set of 76 news headlines and Even with a small training dataset, the improvements in the result are significant. Table 4.2 showcases the increase in accuracy of the fine tuned LLM model.

The process of fine tuning a pre trained model enabled us to refine the LLM model's ability to accurately predict emotions based on textual input. By exposing the model to labeled data encompassing diverse emotional states, we

Financial News	**Emotion**
Retail Revolution Unleashed: FedEx's Direct-To-Consumer Advantage	Joy
Apple Stock Dividend Analysis	Neutral
Nvidia's Smashing AI-Fueled Quarter Marks The Official Beginning Of A New Computing Era	Joy
Wall Street calls Apple's selloff on China concerns 'overblown'	Fear

Table 4.1. News headlines used for fine-tuning Distilled LLM model

facilitated its comprehension of the nuanced nuances inherent in human expression. Leveraging this fine-tuned LLM model, we proceeded to predict the emotions embedded within the news articles captured in section 2 of our study. With the aid of this fine-tuned model, we can effectively classify emotions within English text datasets, offering valuable insights into the emotions captured through textual content. The model's training involved exposure to six distinct datasets, each representing a diverse range of linguistic contexts and emotional expressions. As depicted in Figure 4.1, the model adeptly predicts Ekman's six basic emotions—anger, disgust, fear, joy, sadness, surprise—alongside a neutral class, thereby enabling a comprehensive analysis of emotional content within the news articles under scrutiny.

4.4 Limitations and bias of pretrained LLM model

The training data utilized for the base model was sourced directly from unfiltered raw text extracted from various sources on the internet. This approach enabled the model to learn from a diverse range of linguistic patterns and textual expressions encountered across online platforms. Importantly, during the pretraining phase, the models were trained in a self-supervised manner, mean-

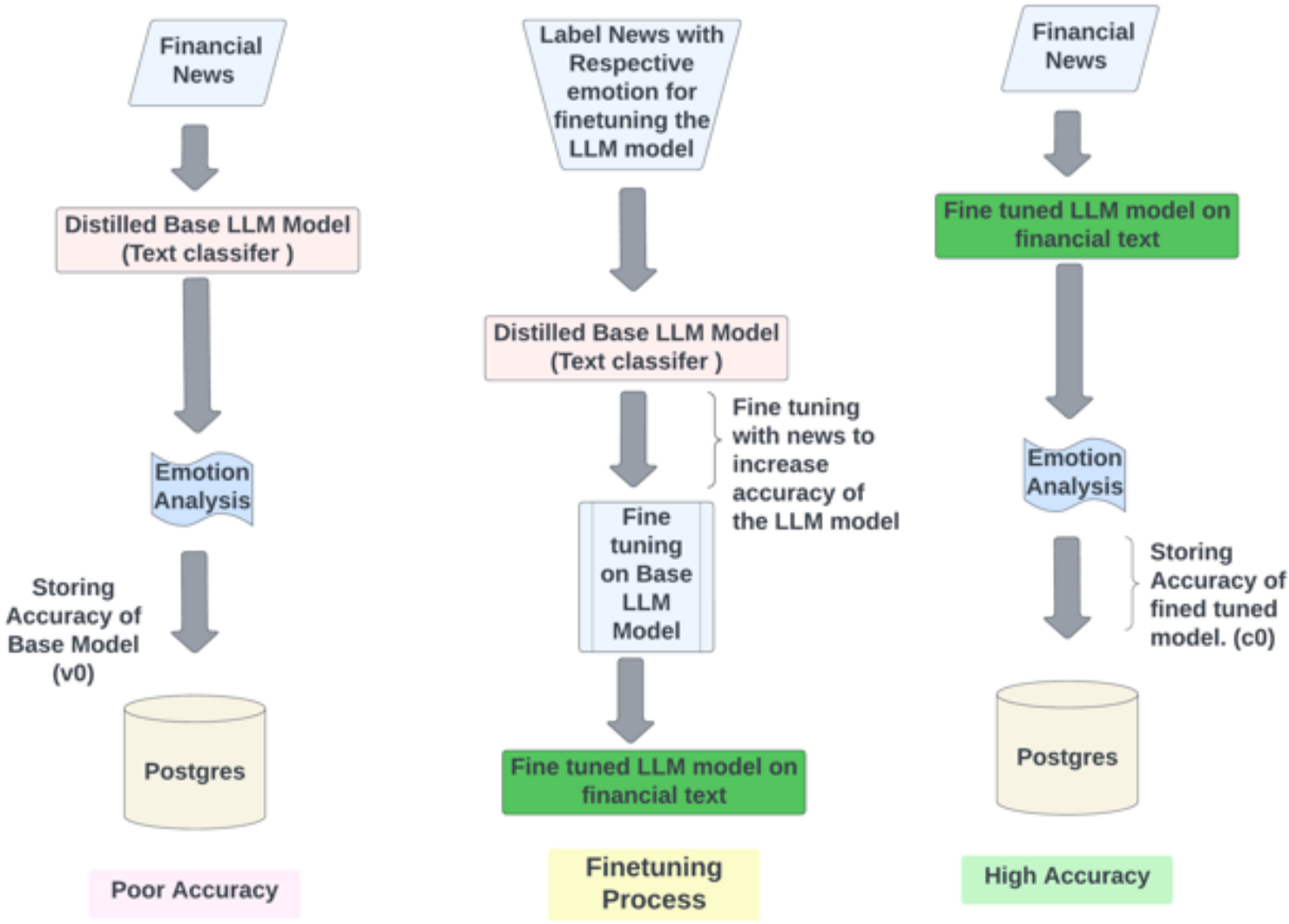

Figure 4.3. Fine tuning Distilled LLM Model to predict emotions behind financial text.

title character varying (512)	published_date timestamp without time zone	emotion character varying (64)	emotion_strength double precision	source character varying (512)	version charact
Is Apple Inc (AAPL) Fairly Valued? An In-depth Valuation Analysis	2023-08-24 16:15:37	neutral	0.8814097046852112	Yahoo Entertainment	c0
Deep Dive: Want to diversify your stock-market investments away from top-heavy indexes? Here's another easy ...	2023-08-29 13:49:00	joy	0.4959230124950409	MarketWatch	c0
What's Worth Streaming: What's worth streaming in September 2023? Here are your best bets amid slim pickings.	2023-08-31 11:12:00	joy	0.8622071146965027	MarketWatch	c0

Figure 4.4. emotion analysis results stored in database.

Financial News	**Date**	**Pre Training Emotion**	**Post Training Emotion**
Why Apple Should Snap Up ESPN: Top Analyst Breaks It Down	08/17/2023	Neutral 0.81	Joy 0.77
Apple in the Spotlight: From SSD Risks of Recycled Parts to UK Surveillance Law Controversies	08/21/2023	Fear 0.55	Fear 0.85
Oracle stock slumps, Apple shares steady and other stocks on the move	09/12/2023	Neutral 0.76	Joy 0.72
App Store anti-steering ban would be consumer-friendly, with little risk to Apple	12/14/2023	Neutral 0.87	Joy 0.68
After Over A 40% Rally In 2023, Will Antitrust And IPhone Issues Hurt Apple Stock?	01/11/2024	Neutral 0.31	Fear 0.50
Cathie Wood's Ark Invest Sells Coinbase and Robinhood Shares, Buys Tesla	01/13/2024	Neutral 0.49	Joy 0.92
Tesla Stock Makes Decisive Move From Key Level. Is This A Buy Signal?	01/11/2024	Fear 0.60	Joy 0.56

Table 4.2. Emotion Analysis prediction between Base LLM Model and Finetuned LLM Model for Financial News: Apple and Tesla

ing that there was no human intervention or labeling of the text. Instead, the model relied on self-supervision mechanisms, such as masked language modeling, to learn and infer patterns from the raw textual data.

While the self-supervised training methodology offers significant advantages in terms of scalability and efficiency, it also introduces inherent limitations and potential biases. Since the model learns solely from the raw textual data available on the internet, it may inadvertently capture and internalize biases, inaccuracies, and inconsistencies present within the training data. These biases can stem from various sources, including but not limited to cultural, societal, and linguistic biases inherent in online content.

As a result, the model's predictions may reflect these biases and inaccuracies, potentially leading to erroneous or skewed outcomes. For instance, if the training data disproportionately represents certain demographics, languages, or topics, the model may exhibit a propensity to prioritize or favor those characteristics in its predictions. Moreover, the absence of human supervision during training means that the model may lack nuanced understanding or context, leading to misinterpretations or misrepresentations of textual content.

It's important to acknowledge these limitations and exercise caution when interpreting the model's predictions, particularly in sensitive or high-stakes applications. Mitigating biases and ensuring the robustness and fairness of the model's predictions require ongoing efforts, including comprehensive data preprocessing, validation, and augmentation techniques. Additionally, supplementing the model's training with curated, labeled datasets and incorporating mechanisms for bias detection and mitigation can help enhance the model's reliability and trustworthiness in real-world scenarios.

CHAPTER 5

USECASES

The process of fine-tuning the Distilled LLM model for a specific domain is both swift and straightforward, owing to the model's compact size and adaptability. The potential applications of emotion analysis are vast and diverse [5]. This section presents several illustrative examples.

Emotion analysis holds significant promise in various domains, including:

Customer Sentiment Analysis: Emotion analysis can be invaluable for interpreting customer reviews and feedback, providing insights into the emotional responses towards products or services. By discerning the prevailing sentiments, businesses can tailor their offerings and customer experiences more effectively.

Social Media Monitoring: Monitoring social media platforms allows organizations to gauge public sentiments towards their brand, events, or topics of interest. Emotion analysis enables the extraction of valuable insights from the vast sea of social media data, facilitating informed decision-making and brand management strategies.

Political Sentiment Analysis: Understanding public opinions and sentiments towards political figures, parties, or policies is crucial for political analysts, policymakers, and strategists. Emotion analysis can help unravel the complex web of political discourse, providing nuanced insights into voter sentiments and preferences.

Usage in Healthcare domain: Analyzing patient reviews and feedback offers healthcare providers a deeper understanding of patients' emotional experiences

and perceptions of their services. Emotion analysis aids in identifying themes and trends in patient feedback, enabling healthcare providers to enhance patient satisfaction and quality of care.

Emotion analysis can be used to perform Thematic analysis [3] for analyzing patient experience with healthcare providers.

Additionally, the dataset obtained in chapter 3 presents numerous opportunities for exploration and analysis. In this paper, we propose another compelling use case: predicting stock price direction using machine learning classification algorithms.

CHAPTER 6

EXPERIMENT

We divide our approach into 2 steps: Data preprocessing and Machine Learning Algorithm execution. The details of the two steps in our approach are as follows:

6.1 Data Preprocessing

Step 1 is **Data Preprocessing**. This is comprised of selecting Input Features (emotions and emotion strength) for the first experiment and selecting input features(open_price, close_price, high,low, volume) for the second experiment. Both experiments have price_direction for the next day as Output Variable. Step 1 for both the experiment includes the following:

- For experiment 1: Create appropriate SQL query to fetch emotions and emotion_strength for a given time frame and include closing_price in the dataset.
- Since there were 7 emotions in string format, we performed one hot encoding to these emotions to convert them into boolean values. There were 7 input features (anger, disgust, fear, joy, neutral, sadness, surprise).
- For experiment 2: Create appropriate SQL query to fetch the stock price related information for a given time frame.
- For both experiments, in order to avoid over fitting we remove attributes such as company name, date etc.
- The input features were open price, volume, high, low price, rolling averages of close_price.

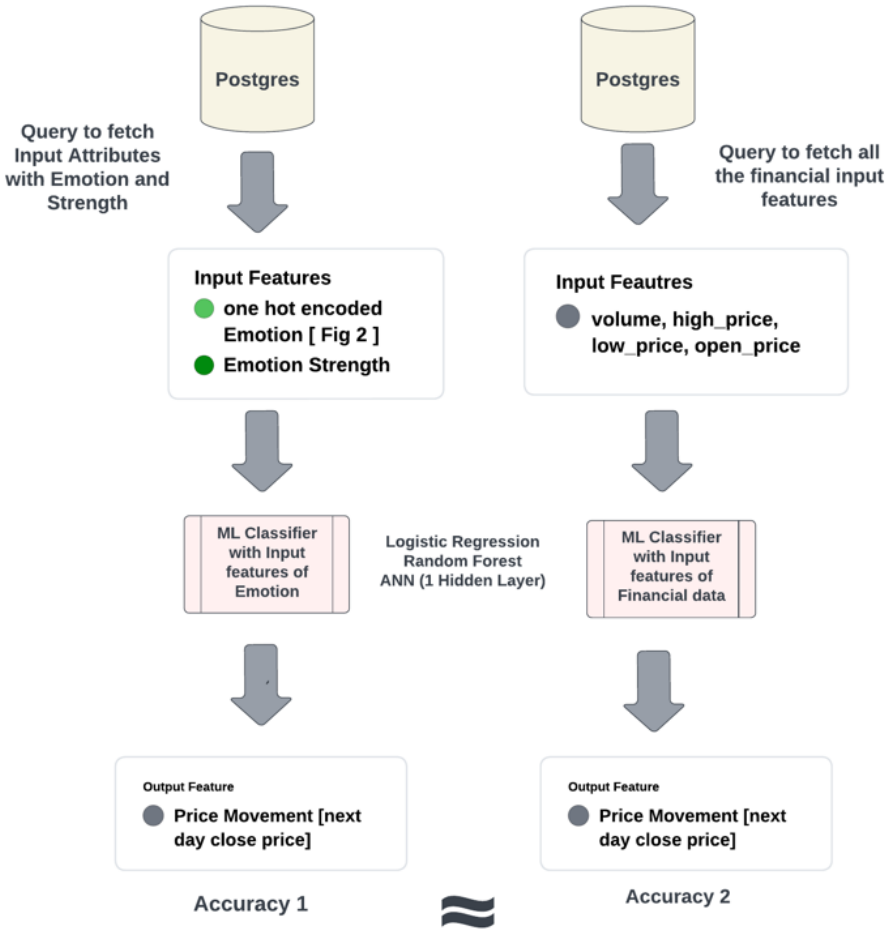

Figure 6.1. Overview of Separate ML Classifiers for predicting stock price direction.

– The dataset utilized for training exhibited skewness owing to the inherent volatility of stock price fluctuations. Nearly 5000 records indicated a next_day_close_price of 0, and close to 900 records showed a positive value (1). We employed the value 0 to signify a scenario where the closing price of the subsequent day is lower than that of the current day, and assigned

Classifier Algorithms	F Attributes	E. Attributes
Logistic Regression	0.87	0.87
Random Forest	.83	0.79
ANN(1 hidden layer)	.84	.67

Table 6.1. Next day closing price direction prediction based on only financial attributes (F. Attributes) or only emotion analysis based attributes (E. Attributes)

a value of 1 if the closing price of the following day exceeded that of the current day.

6.2 Algorithm execution

Step 2 is **Algorithm execution**. This is comprised of splitting the data into training and test sets. Step 2 includes the following parts:

- Split each dataset into training (80%) and test (20%) datasets.
- Use random_state as 42 in the classifier algorithms to get the same splits for every iteration, This helps in reproducing the accuracy for the same dataset.
- Run the three classification algorithms (Logistic Regression, Random Forest, Artificial Neural Network) for both experiments with the appropriate datasets.

6.3 Future enhancements for experiments

- In the current work we combined all the data pertaining to the twenty five stocks and used it for the training dataset. Since companies have unique patterns of price fluctuations, in future, we will segment the data to be company specific.

- A financial dataset downloaded from Kaggle [1] had over a million records. However, it contained global stocks not related to the United States. Even though we had access to the financial attributes (open_price, close_price, high_price, volume) of our twenty-five stocks since the early 2000s, on average, the earliest we could fetch the news information was from the year 2018 and later. Much of the requisite information needed to experiment accurately with the help of emotion analysis with the current dataset is missing. In the future, we will try to find more news articles from the past to augment the dataset.
- The advanced input features of the dataset and their explanation include:
  * Daily_Return: The daily percentage change in the closing price, indicating the price movement relative to the previous day's closing price.
  * SMA_50: The 50-day Simple Moving Average (SMA) of the closing price, providing a smoothed average price over the specified period.
  * Standard_Deviation: The standard deviation of the closing price over a 50-day period, measuring the dispersion of prices from the SMA_50.
  * Upper_Band: The upper band of the Bollinger Bands, calculated as 2 standard deviations above the SMA_50, indicating potential overbought conditions.
  * Lower_Band: The lower band of the Bollinger Bands, calculated as 2 standard deviations below the SMA_50, indicating potential oversold conditions.
  * EMA_12: The 12-day Exponential Moving Average (EMA) of the closing price, giving more weight to recent prices and providing insight into short-term trends.

* EMA_26: The 26-day Exponential Moving Average (EMA) of the closing price, offering insight into intermediate-term trends
* MACD_Signal: The MACD (Moving Average Convergence Divergence) signal line, calculated as the difference between EMA_12 and EMA_26, aiding in identifying trend reversals or momentum shifts.
* VWAP (Volume Weighted Average Price): The volume-weighted average price, providing insight into the average price weighted by trading volume over a specified period.
* ROC (Rate of Change): The rate of change of the closing price over a 14-day period, indicating the momentum of price movements relative to historical data.
* H-L (High-Low): The price range between the highest and lowest recorded prices over a specified time period.
* O-C (Open-Close): The difference between the opening and closing prices of a financial instrument within a specific time frame.
* 7 Days Moving Average (7 Days MA): The average price of a financial instrument over the past 7 days.
* 14 Days Moving Average (14 Days MA): The average price of a financial instrument over the past 14 days.
* 21 Days Moving Average (21 Days MA): The average price of a financial instrument over the past 21 days.
* 7 Days Standard Deviation (7 Days STD DEV): The measure of the dispersion of prices from their average over the past 7 days.

With the inclusion of these input features, we collected the dataset for Nike stock starting from 04/05/2009 to mirror the dataset used in [13]. We used Random Forest to compute the RMSE (Root Mean Squared Error)

as a metric to gauge the discrepancy between model-predicted values and the observed values. We achieved RMSE value of 0.5129202619530167. Notably, this was an improvement over similar experiments conducted in [13], for the stock Nike where only a subset of financial attributes, including H-L, O-C, 7-day moving average, 14-day moving average, 21-day moving average, and 7-day standard deviation, were included as input features and resulted in RMSE score of 1.10.

We couldn't incorporate emotion and emotional_strength into our Kaggle dataset based experiment because we lacked the necessary daily news data dating back to 2009. Once we acquire a suitable dataset containing news information, we intend to conduct more comprehensive experiments.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Conclusion

In this paper, we introduce a fresh methodology to beef up predicting emotions in financial news articles. Our approach taps into the capabilities of a Distilled Large Language Model (LLM), known for its swift and efficient processing of textual data. Additionally, we fine-tune the model's classification prediction by fine tuning a pretrained LLM model with domain specific input, a process where news headlines are manually labelled to refine emotion analysis prediction.

The adoption of Distilled LLM models comes with significant perks, especially in terms of rapid processing and training. This swiftness is crucial for handling large volumes of financial news data effectively. To enrich our analysis, we enhance the dataset with various attributes related to news headlines, including temporal factors like the duration since the last news article. Corresponding attributes are also added for the financial dataset, ensuring a comprehensive coverage of factors influencing stock price movements.

Our experimental setup comprises two parallel experiments, each focusing on different attribute sets: one exclusively based on emotion analysis attributes, and the other integrating financial attributes with emotion analysis. By dissecting and scrutinizing each attribute set independently, we aim to unravel their respective impacts on predicting the next day's stock closing price direction.

Throughout our experiments, we employ a dataset spanning four and a half months, providing ample time for rigorous analysis and evaluation. By explor-

ing the correlation between emotional sentiment in financial news and subsequent stock price movements, we glean insights that can guide decision-making processes in financial markets.

7.2 Future Work

In future work, there exists a vast scope for exploring company-specific patterns by leveraging emerging datasets available on platforms like Kaggle. Conducting in-depth company-wise pattern analysis can offer insights into the unique emotional impact of various events on individual stock performance. This exploration can discover relationships between specific company sentiment and corresponding stock price movements, providing a granular understanding of market dynamics at the micro-level. Additionally, integrating social media data from platforms like X (formerly Twitter) and Reddit presents an opportunity to extract investor opinions and reactions to news in real-time. By incorporating such data streams, predictive models can be further enriched, offering a more comprehensive understanding of market sentiment and enhancing stock price trend detection.

Furthermore, future research could go deeper into the content of financial news articles beyond headline analysis. We were not able to achieve this due to the limitations in newsapi.org API response. Emotion analysis applied to the entire article content could yield richer insights into the underlying sentiments shaping market movements. Additionally, the exploration of multimodal data integration, including textual articles, images, and videos, holds promise for capturing a more holistic view of market sentiment. By combining multiple data sources, researchers can enhance the sophistication of predictive models, enabling more accurate and robust stock price trend detection systems.

REFERENCES

[1] Miguel Aenlle. Daily financial news for 6000+ stocks. Available at https://www.kaggle.com/datasets/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests, 2023.

[2] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

[3] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. Llm-in-the-loop: Leveraging large language model for thematic analysis. *arXiv preprint arXiv:2310.15100*, 2023.

[4] Thomas Delcey. Samuelson vs fama on the efficient market hypothesis: The point of view of expertise. *Œconomia. History, Methodology, Philosophy*, (9-1):37–58, 2019.

[5] Yingqiang Ge, Wenyue Hua, Jianchao Ji, Juntao Tan, Shuyuan Xu, and Yongfeng Zhang. Openagi: When llm meets domain experts. *arXiv preprint arXiv:2304.04370*, 2023.

[6] Jochen Hartmann. Emotion english distilroberta-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/, 2022.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[8] Yu Ma, Rui Mao, Qika Lin, Peng Wu, and Erik Cambria. Multi-source aggregated classification for stock price movement prediction. *Information Fusion*, 91:515–528, 2023.

[9] Jigar Patel, Sahil Shah, Priyank Thakkar, and Ketan Kotecha. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert systems with applications*, 42(1):259–268, 2015.

[10] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

[11] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019.

[12] Robert P Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2):1–19, 2009.

[13] Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, and Arun Kumar. Stock closing price prediction using machine learning techniques. *Procedia computer science*, 167:599–606, 2020.

[14] Tianyi Wang, Ke Lu, Kam Pui Chow, and Qing Zhu. Covid-19 sensing: negative sentiment analysis on social media in china via bert model. *Ieee Access*, 8:138162–138169, 2020.

[15] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized BERT pre-training approach with post-training. In Sheng Li, Maosong Sun, Yang Liu, Hua Wu, Kang Liu, Wanxiang Che, Shizhu He, and Gaoqi Rao, editors, *Proceedings of the 20th Chinese National Conference on Com-*

putational Linguistics, pages 1218–1227, Huhhot, China, August 2021. Chinese Information Processing Society of China.