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

Rithesh Bhat (https://orcid.org/0009-0007-0042-7839), Computer Science, University of Texas At Arlington, United States, rhb8955@mavs.uta.edu (mailto:rhb8955@mavs.uta.edu)
Bhanu Jain (https://orcid.org/0000-0001-8527-5908), Computer Science, University of Texas At Arlington, United States, bhanu.jain@uta.edu (mailto:bhanu.jain@uta.edu)

DOI: https://doi.org/10.1145/3652037.3652076 (https://doi.org/10.1145/3652037.3652076)
PETRA '24: The PErvasive Technologies Related to Assistive Environments (PETRA) conference (https://doi.org/10.1145/3652037), Crete, Greece, June 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 demonstrate 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.

CCS Concepts: • **Applied computing** → **Document analysis**;

Keywords: Artificial intelligence, neural networks, machine learning, trend prediction, logistic regression, Random Forest, Artifical Neural Network, stock price direction prediction, LLM, emotion analysis, sentiment analysis, Distilled LLM.

ACM Reference Format:
Rithesh Bhat and Bhanu Jain. 2024. Stock Price Trend Prediction using Emotion Analysis of Financial Headlines with Distilled LLM Model. In *The PErvasive Technologies Related to Assistive Environments (PETRA) conference (PETRA '24), June 26--28, 2024, Crete, Greece.* ACM, New York, NY, USA 7 Pages. https://doi.org/10.1145/3652037.3652076 (https://doi.org/10.1145/3652037.3652076)

1 INTRODUCTION AND MOTIVATION

In recent years, stock market analysis and prediction has become an exciting and a popular field. Given that the volatility and dynamics of stock market is not linear, predicting does not seem viable as explained in Efficient Market Hypothesis (EMH) and Random Walk Theory (RWT) [2] [8] [9]. The stock price prediction domain has undergone significant research and debate, with different perspectives emerging from [4] regarding the Efficient Market Hypothesis (EMH). EMH states that the price of a security reflects all the information available and everyone has access to the information making it impossible to predict using historical data alone; RWT, on the other hand, states that stock market prices are determined randomly, and hence prediction is infeasible. With the rise of social media and online platforms, news can reach millions of people instantly. In the world of finance any critical news can have significant impact on the stock price during the trading window or when the market opens for trading. Researchers have explored the relationship between investor sentiments expressed in online forums and stock price movements [12] Researchers have also used public sentiment in news outside of stock market field; During the pandemic period in 2020, which gave insights to make appropriate public health measures and responses to the outbreak by analyzing the news in social media; classify the sentiment (positive,negative,neutral) and then use TF-IDF to summarize the topics [14].

In this research work, we use ML classification models to predict the stock price direction. We use Logistic Regression, ANN, and Random forest algorithm. The results illustrate that emotion analysis-based attributes alone can give us as accurate prediction results as exclusively financial attributes based classification experiments can.

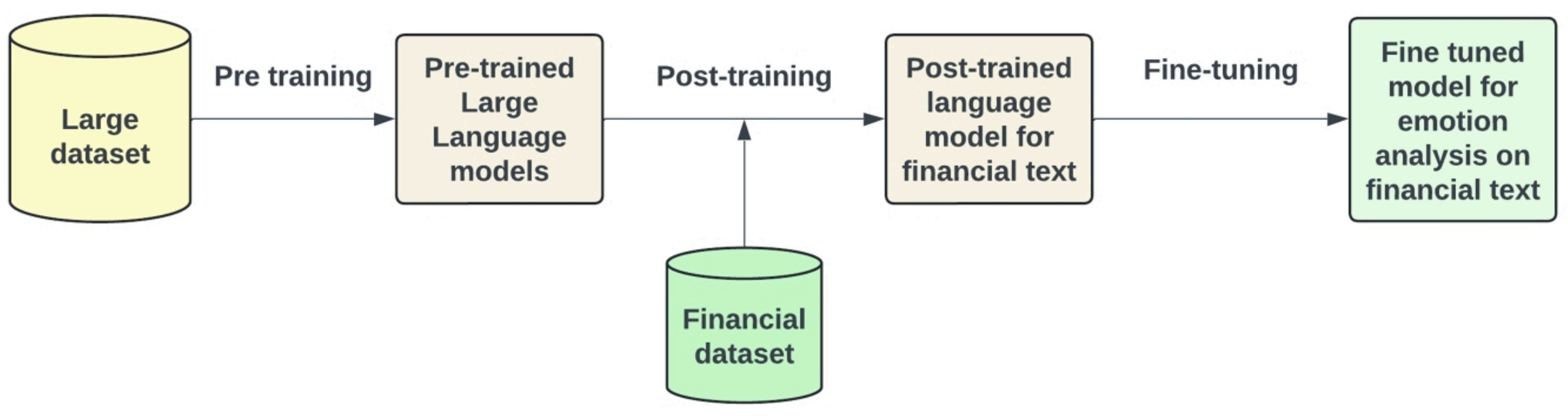

Figure 1: Fine tuning Distilled LLM Model to predict emotions embedded in the financial text

Our main contributions in this work are as follows:

- Used APIs from financial aggregators to create the required dataset to predict stock prices, thereby eliminated the need for web scraping to curate a financial dataset.
- Demonstrated and used the process of fine tuning a pretrained LLM model to efficiently predict the emotions for financial news headlines
- Leveraged Distilled LLM model to perform text-classification task instead of traditional NLP tasks for the same purpose
- Executed classification algorithms exclusively on emotion analysis attributes and then exclusively on financial attributes to predict the direction of stock price
- Analyzed and discussed the limitations and challenges faced in our approach.

In summary, the integration of Distilled LLM Model, emotion analysis of news headlines, and machine learning classification algorithms, present a promising approach in the field of stock price trend prediction. Our methodology paves an alternate path to predicting future stock prices using emotions embedded in financial news headlines as opposed to the traditional path of stock price prediction using financial data.

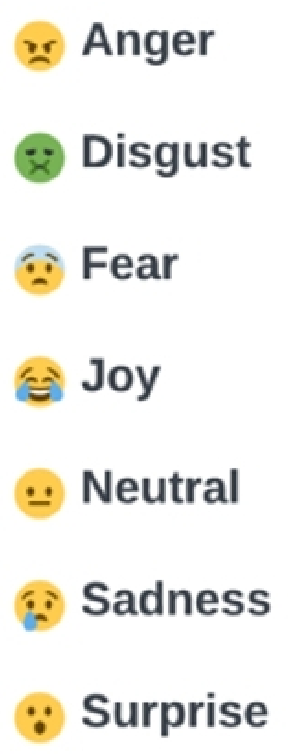

Figure 2: Emotions supported in Distilled text classification LLM Model

2 OVERVIEW OF DATASET AGGREGATION

2.1 Identification of Stocks

In this research, we selected a list of 32 mega-cap companies from the United States. Mega caps have a market capitalization of above 200 billion USD. We selected these companies because it is easy to find news articles about them and hence easy to study the information captured in the news headlines and scrutinize its correlation with the price trend of the company.

Subsequently, we gathered information pertaining to two areas related to the companies, namely financial news related to these companies and the financial attributes such as open price, close price, volume, day's high and low price of the stock. We stored these pieces of information in a database for downstream processing with the LLM model and other machine learning models. In the following subsections we elaborate upon the process used to extract these pieces of information from 2 different aggregators.

2.2 Techniques for financial news extraction

With the rise of GPT and 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 that results in blocking of the incoming ip addresses; and in some cases, the platforms have updated their policies on prohibiting the use of automated web scrapers since early 2023. To sidestep this problem as we collect news related to the shortlisted mega-caps, we retrieve news via API's from official news aggregators instead of scraping the news sites. There are many news aggregator platforms that offer news service in free as well as paid versions. In this paper, we have used newsapi.org as our news aggregator. They locate articles written worldwide and give headlines over API based mechanism. We created an account on their platform to get the APIKEY which needs to be passed as a header attribute when making the API call. The platform provides the user with 100 free API requests per day and gives access to worldwide news.

2.3 Techniques for fetching financial attributes related to the stocks

While collecting financial attributes of companies, we faced similar restricting web scraping challenge as we described earlier in Section 2.2. We overcame this challenge by using Alpha Vantage, a financial technology company, as the aggregator to fetch the stock price related information. AlphaVantage requires its users to create an account and then get their API_KEY. This API_KEY needs to be sent with an information request. AlphaVantage provides both real time and historical data upto 2 decades. For our use case we fetch the daily stock price and related attributes such as open price, close price, day's high and low, volume. Additionally, we also captured annual and quarterly earnings reports for each firm.

2.4 Libraries used during dataset collection

We have used 2 aggregators to fetch data. Each of these have their own Python package with ample utility functions to make the required API calls. We wrote a wrapper class that invokes the newsapi package internally. We invoke the get_top_headlines from the business category for the United States. We extract the following attributes from the API response: 'source' of news, the headlines of the article, url, if applicable - urlToImage, the date of article being published, description, and content of the article as provided in response. These pieces of information are then carefully processed and stored in the Postgres database. Alpha Vantage has a python package named alphavantage which gives us utility function to get the required data. The historical_data utility method gives us response of stock data up to 2 decades in a single API call which is a convenient and an efficient use of API calls. The package however did not give us any methods to fetch the earnings report for the firm, hence we had to switch to conventional way to fetch details from API by making explicit API call. For this research, we have started extracting News and stock price related information from August 2023.

2.5 Historical news dataset collection

We had limited news headlines extracted from news aggregators, since we were using free tier of newsapi, we could not extract data prior to one month of time. We were able to find the required historical data from Kaggle [1] for 6,000 stocks for the time span 2009 to 2020.

2.6 Challenges faced during dataset collection

Free tier of NewsAPI can fetch historical news for only up to one month and provide only 100 free API requests a day. Any historical news beyond one month time frame cannot be fetched using their API without subscribing to the paid plan. The responses from NewsAPI do not give the entirety of the text in the article for Description and Content section. Due to this limitation, we cannot perform emotion analysis on the article content as the full context is missing when the entire article text is not available. Analyzing these attributes can help us detect any false positives in the headlines versus the actual content of the article. The free tier of AlphaVantage gave us 25 API requests a day, which meant we could not monitor stock prices for more than 25 firms per day.

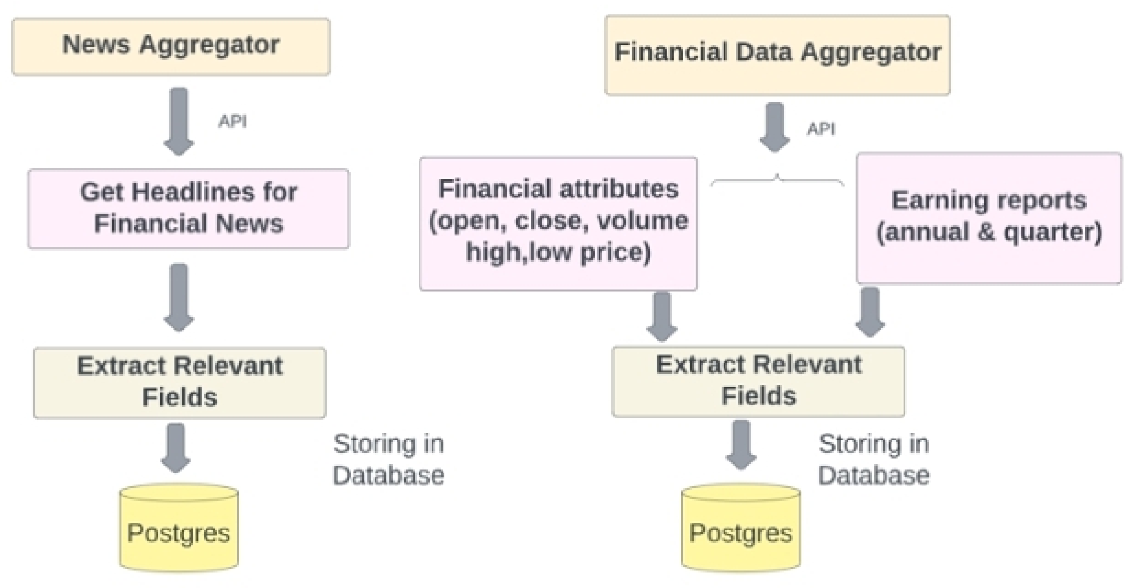

Figure 3: Emotions supported in Distilled text classification LLM Model

3 EMOTION ANALYSIS

Emotion Analysis refers to the task of predicting emotion from input text. In this work, we used financial news headlines to perform emotion analysis. We use emotion analysis as it has much more granularity and depth than Sentiment Analysis. We extract the financial news headlines and store them in our database. This is followed by accurately predicting the emotions encompassed in the headlines of the published articles. There are several ways to predict the emotions embedded in a given text. The traditional approach is comprised of performing NLP tasks such as tokenization, stemming, stop words removal, TF-IDF, and labelling the data for appropriate emotion, followed by using ML model such as Naïve Bayes, SVM etc. Training these model as per the traditional approach also involves performing hyper parameter tuning and post-processing. In contrast to the traditional approach, we use Distilled LLM model to do the text-classification instead of using NLP for the news headlines to get emotions and emotion strength.

3.1 Strategies to choose the appropriate LLM Model

There are many public platforms available that allow users to utilize the capabilities of LLM models (eg: OpenAI, claude) with some limitations mentioned below.

- Most of these offerings are paid and have limited number of requests you can do on the free tier.
- These requests are processed on cloud and the end users have limited control over how these requests are processed.
- We do not have control over the biases the model has and the ability to fine tune the model to adapt to a specific domain.
- General purpose LLM models are huge in size and can handle versatile tasks and thus can be slower compared to Distilled LLM's focused on specific task.

To overcome this, we downloaded the Distilled LLM model (light weight model) to our local system. We then trained and customized the base LLM model to domain specific text-classification

To find the appropriate LLM model, we chose hugging face platform. The platform hosts open source LLM models that users can fine tune and customize for their applications and can even share their own LLM models here. The significant gain achieved by using a pretrained LLM model for text-classification is that we do not have to reinvent the wheel. For example, in our case, we did not have to perform the NLP task. We used emotion-english-distilroberta-base [6], a fork from robertaBase [10] as our base LLM model. The main focus of this LLM model is to merely perform text-classification. Due to the LLM models distilled nature, it can only perform specific tasks, resulting in a lighter and computationally faster execution of those tasks.The model supports Ekman's six basic emotions (anger, disgust, fear, joy, neutral, sadness, surprise) and a neutral class as shown in Figure 3.

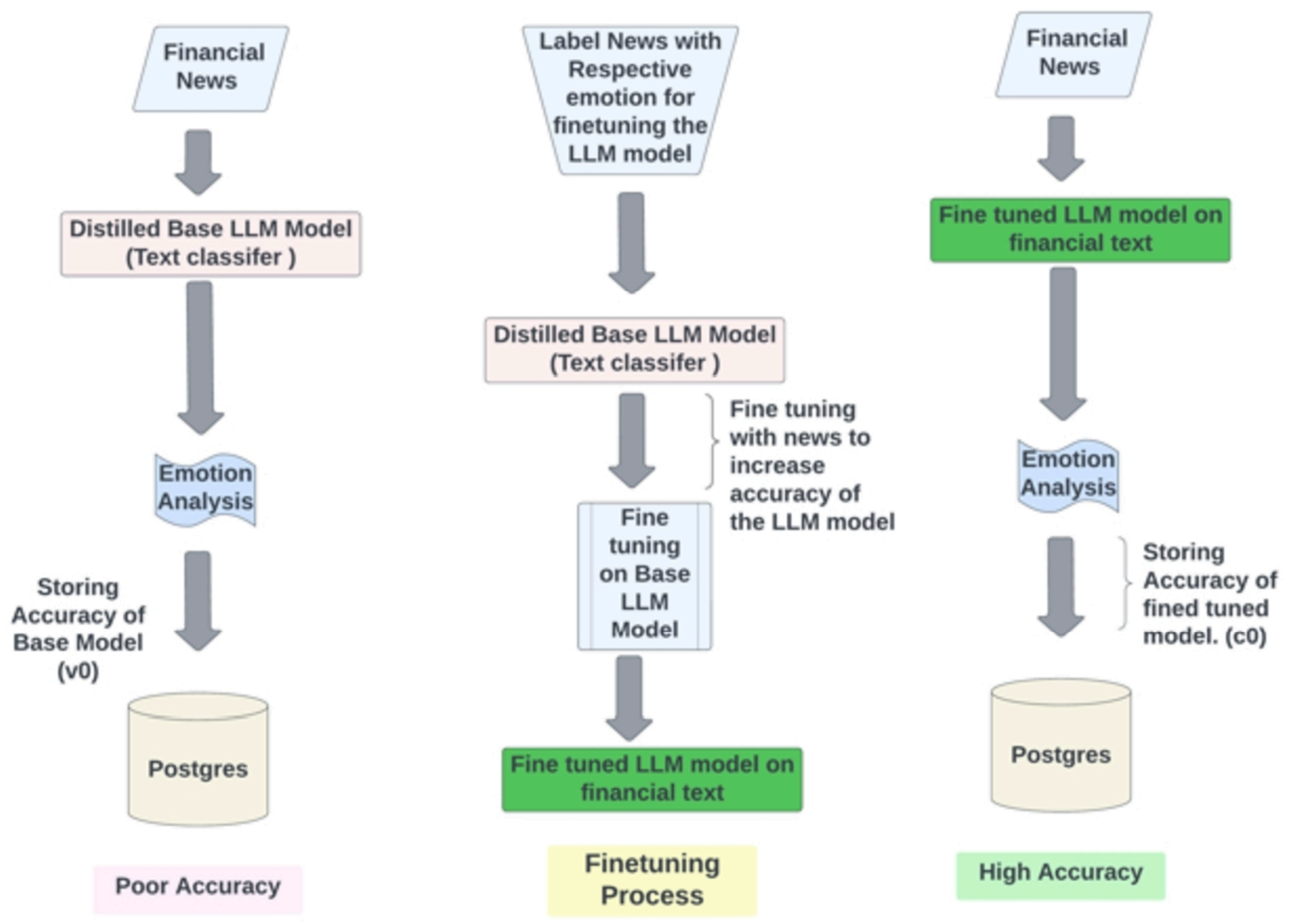

Figure 4: Fine tuning Distilled LLM Model to predict emotions encompassed in financial text

Table 1: News headline used for fine tuning Distilled LLM model

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

3.2 Training data used for Base LLM

The roberta-base model was pretrained on the reunion of five datasets, which in aggregate is close to 160 GB of text [7]. BookCorpus, OpenWebText (recreation of the dataset used to train GPT-2, English Wikipedia, CC-News (Dataset with 63 million English News Articles from September 2016 to Feb 2019) and Stories. The model was pretrained with the Masked language modeling (MLM) objective. Taking a sentence, the model randomly masks 15% of the words in the input and then runs the entire masked sentence through the model and has to predict the masked words [7]. The base model was then distilled into DistilRoBERTa-base by following the same procedure as performed in [11] which is to leverage knowledge distillation during the pre-training phase and shows that it is possible to reduce the size of a BERT model by 40% while retaining 97% of its language understanding capabilities and by being 60% faster.

Once we finalized on the fine-tuned checkpoint of DistilRoBERTa-base LLM Model [6] for our use case, the next task was to make it efficient in predicting the emotion of financial news. This was achieved by fine tuning a pre-trained model. This training consists of manually labeling the text with the appropriate labels and then training the base model with this so that the LLM model gets acquainted with the domain specific verbiage and can predict similar texts with much higher accuracy than before fine tuning the LLM model. See Figure 4 and table 2.

Table 2: Emotion Analysis Accuracy between Base LLM Model and Finetuned LLM Model for Financial News: Apple and Tesla

Financial News	**Published** Date	**Pre Training** Accuracy	**Post Training** Accuracy
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

3.3 Fine tuning the LLM model for financial Text

We take the LLM model [6], and then train it with the input features consisting of text and label with one of the seven emotions supported in the LLM model. This process further trains the pre-trained models on the domain specific dataset [15] For our experiment, we used just 76 news with varying emotions. Even with a small training dataset, the improvements in the result are significant. Table 2 showcases the efficacy of the fine tuned LLM model. We use this fine tuned LLM model to predict emotions of all the news captured in Section 2. With this Fine-tuned model, we can classify emotions in English text dataset. The model was trained on six diverse datasets and predicts Ekman's six basic emotions [6] and a neutral class as shown in Figure 3.

3.4 Limitations and bias of pretrained LLM model

The training data used for the base model was unfiltered raw text from the internet. The models were pretrained in a self-supervised manner, with no human labelling of text. Therefore, the model can have biased or inaccurate predictions.

4 USECASES

The use cases for utilizing emotion analysis are limitless [5]. This section provides an overview on how this dataset can be used for sentiment and emotion analysis, by furnishing several examples. The Distilled LLM model can be easily finetuned in the domain of choice. Fine tuning process is quick, and the size of distilled models are typically small. Emotion analysis can be used in analyzing customer reviews and feedback to understand the emotions towards products or services. They can also be used for monitoring social media platforms to gauge public emotions towards a brand, event, or even a topic. Analyzing public opinions and sentiment towards political figures, parties, or policies is another use case. Thematic analysis [3], analyzing patient reviews and feedback to understand their emotional experiences with healthcare providers are additional use cases. The dataset obtained in Section 2, can be used in many ways. In this paper, we offer another use case, where we use the dataset to predict the stock price direction using the ML classification algorithms. As mentioned in Section 3, after obtaining the data, the next step is to analyze the relationship between stock price fluctuation and emotions behind the headlines on a given day. This will be covered in Section 5.

5 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:

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

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.

Table 3: Next day closing price prediction based ononly financial attributes (F. Attributes) or only emotionanalysis based attributes (E. Attributes)

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

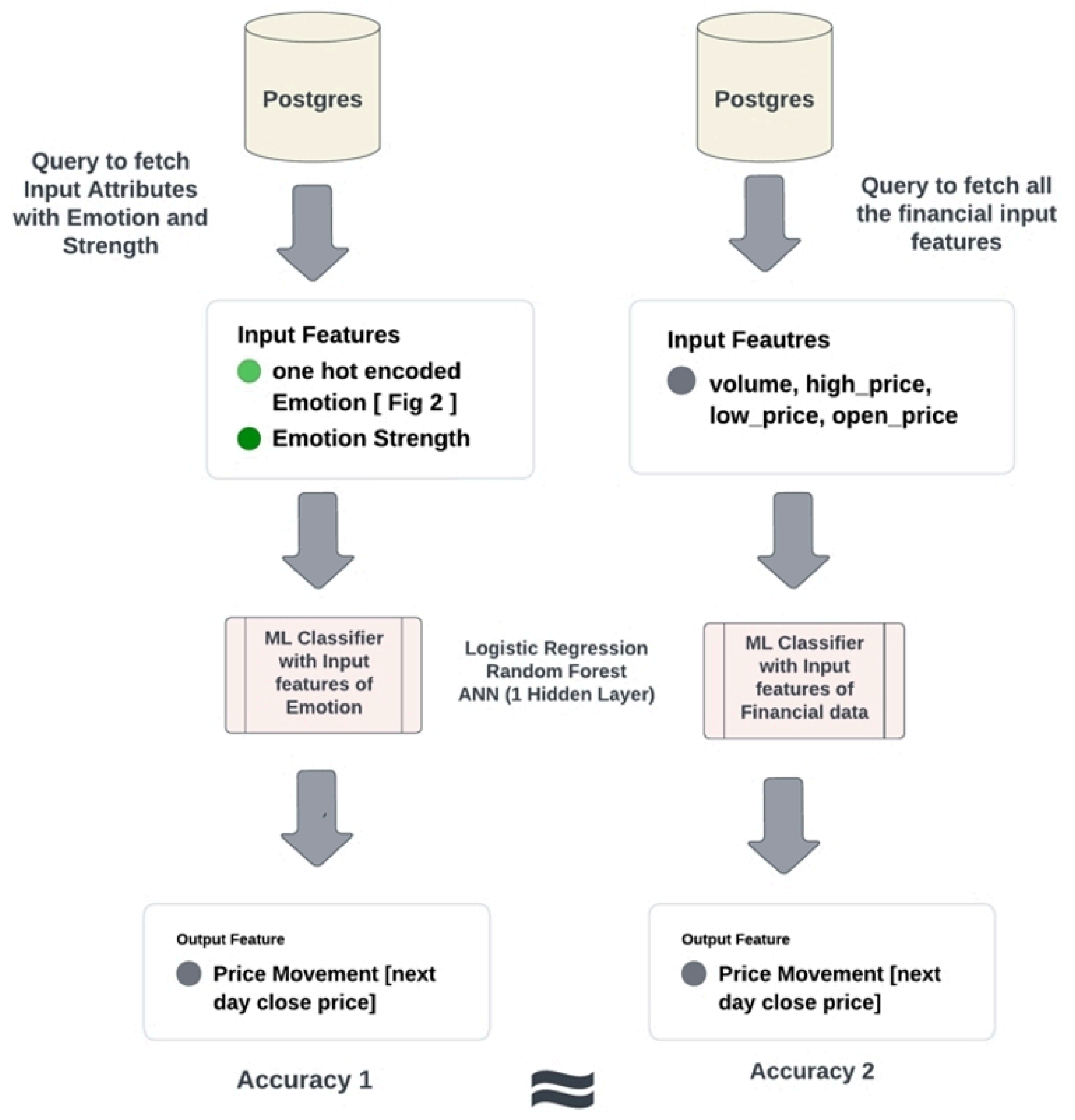

Figure 5: Overview of Separate ML Classifiers for predicting stock price direction

5.1 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.

6 CONCLUSION

In this paper we use a new approach for spiking the accuracy in predicting emotions for financial news using a Distilled LLM model and refining it by fine tuning the pre trained LLM model with manually labelled news headlines to further increase the accuracy of emotion analysis of news headlines. Distilled LLM models work exceptionally fast for the intended task and can be trained quickly. We added multiple attributes related to news headlines such as days from previous news, and corresponding attributes for the financial dataset too. We then use the emotion analysis-based attributes and financial attributes exclusively at a time to predict the next day closing price direction of the stocks in two parallel experiments. The dataset used for our experiments consists of four and a half months of news headlines that were collected ethically following the financial portals' rules. We believe that we can further improve the performance with a larger dataset. In the near future we would like to study the accuracy with a larger dataset with news collected from multiple sources and not just that from aggregators. In our next paper, we would like to extract news from popular social media platforms like X, formerly twitter, and reddit where investors pitch in their opinions and react to the news.

In this paper, we did not have stock news from all the days so the accuracy should improve further as more data becomes accessible. We could demonstrate that emotions captured from financial news headlines were sufficient in predicting the next day stock price prediction with almost as high of an accuracy as that from prediction using financial data alone for the same time frame in the two parallel experiments.

Finally, doing an emotion analysis on the content of the financial news articles beyond the analysis of headlines alone should be an interesting sequel to the current work and should further increase the accuracy of our current predictions. We would like to pursue that in our future work.

REFERENCES

[1] Miguel Aenlle. 2023. Daily Financial News for 6000+ Stocks. Available at **https://www.kaggle.com/datasets/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests (https://www.kaggle.com/datasets/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests)**. Navigate to

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

[3] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging large language model for thematic analysis. *arXiv preprint arXiv:**2310.15100 (https://arXiv.org/abs/2310.15100)*** (2023). Navigate to

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

[5] Yingqiang Ge, Wenyue Hua, Jianchao Ji, Juntao Tan, Shuyuan Xu, and Yongfeng Zhang. 2023. Openagi: When llm meets domain experts. *arXiv preprint arXiv:**2304.04370 (https://arXiv.org/abs/2304.04370)*** (2023). Navigate to

[6] Jochen Hartmann. 2022. Emotion English DistilRoBERTa-base. **https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/ (https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/)**. Navigate to

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arxiv:**1907.11692 (https://arXiv.org/abs/1907.11692) http://arxiv.org/abs/1907.11692 (http://arxiv.org/abs/1907.11692)** Navigate to

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

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

[10] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv* abs/1910.01108 (2019). Navigate to

[11] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR* abs/1910.01108 (2019). arXiv:**1910.01108 (https://arXiv.org/abs/1910.01108) http://arxiv.org/abs/1910.01108 (http://arxiv.org/abs/1910.01108)** Navigate to

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

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

[14] Tianyi Wang, Ke Lu, Kam Pui Chow, and Qing Zhu. 2020. COVID-19 sensing: negative sentiment analysis on social media in China via BERT model. *Ieee Access* 8 (2020), 138162–138169. Navigate to

[15] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, Sheng Li, Maosong Sun, Yang Liu, Hua Wu, Kang Liu, Wanxiang Che, Shizhu He, and Gaoqi Rao (Eds.). Chinese Information Processing Society of China, Huhhot, China, 1218–1227. **https://aclanthology.org/2021.ccl-1.108 (https://aclanthology.org/2021.ccl-1.108)** Navigate to

PETRA '24, June 26–28, 2024, Crete, Greece

ACM ISBN 979-8-4007-1760-4/24/06.
DOI: **https://doi.org/10.1145/3652037.3652076 (https://doi.org/10.1145/3652037.3652076)**