Stock Market Forecasting with Pretrained Deep Learning Models

Son Tung Do
Department of Computer and Information Sciences
Fordham University
Bronx, NY, USA
sdo3@fordham.edu

Anthony Chu
Department of Computer Science
Stanford University
Stanford, CA, USA
achu2025@stanford.edu

Yijun Zhao
Department of Computer and Information Sciences
Fordham University
Bronx, NY, USA
yzhao11@fordham.edu

Yanjun Li
Department of Computer and Information Sciences
Fordham University
Bronx, NY, USA
yli@fordham.edu

***Abstract*—To perform stock market-related forecasting, we developed a synchronized framework that employs pretrained deep learning models to automatically analyze real-time public information from social media and forecast market trends. In the proposed framework, market related tweets are aligned with stock market index data, textual entailment method for zero-shot classification is adopted to perform sentiment analysis of tweets, and market forecast is generated by applying pretrained time series foundation model. The models were trained and tested with approximately 188,000 tweets and three major indices from the New York Stock Exchange for the year 2021. The experiments demonstrated consistently strong performance. This framework can be extended to analyze text documents from various sources for forecasting.**

***Index Terms*—Sentimental Analysis, Social Media, Stock Market Change, Twitter, Large Language Models, Zero-shot Classification, Textual Entailment, Random Forest Classifier, Time-series Forecasting.**

I. Introduction

In the research area of stock market forecasting, the Efficient Market Hypothesis (EMH) [4] states that stock market prices are largely driven by new information and follow a random walk pattern. Since nowadays everybody shares their ideas, thoughts, and emotions with other users on social media such as X (formerly Twitter), it becomes one of main sources for people seeking new information. Adapting to this trend, financial market analysts need to constantly monitor and evaluate related information on social media to support stock buying and selling decisions.

In this procedure, the following three challenges should be addressed:

1) Accurately capturing the stock market related sentiment from social media: Sentiment Analysis is a text classification task that extracts the sentiment—positive, negative, or neutral—that a writer expresses toward an object/subject. Over the years, numerous sentiment analysis methods have been developed and studied [5], [8], [10], [11], [14], [23], [24], including lexicon-based approaches, machine learning classifiers, and more recently, deep learning models. However, conventional supervised classifiers have often struggled to deliver consistently convincing results due to the limited availability of training datasets, and the complexity and nuance of human languages. Zero-shot text classification method (0Shot-TC) [1] has been studied to tackle this problem. 0Shot-TC aims to associate an appropriate label with a piece of text, irrespective of the text domain and the aspect (e.g., topic, emotion, event, etc.) described by the label, with limited or no training datasets. Furthermore, a textual entailment paradigm is proposed to solve the 0Shot-TC problem with text documents [18]. In this model, human behavior is simulated as constructing a hypothesis for each sentiment class label, and pretrained large language models are utilized to generate the probabilities for entailment and contradiction, which are then converted to label probabilities [6]. One more advantage of the entailment method over the traditional classification method is that the meaning of class labels are encoded and processed with the dataset by the pretrained large language model as well. In this project, we aim to provide a deeper understanding of the sentiment trends in social media data by applying the textual entailment method.

2) Associating the sentiment of social media with stock market changes: Time-series data refers to a sequence of observations recorded at regular intervals, capturing trends and temporal dependencies in dynamic systems. Both tweets sentiment over time and real-time stock market changes can be modeled as time-series datasets. Existing approaches to aligning tweets sentiment with stock prices rely on the following two primary methods:

1) **Next Period Prediction:** This method aggregates tweets collected over a specific period and analyzes their sentiment to predict stock market movements for the following interval. It assumes that aggregated sentiment provides sufficient context for forecasting. Typical periods range from hours to several days [21], [23], [25].
2) **Same Period Prediction:** In this approach, the time interval of tweets is directly matched with stock market data. For example, tweets gathered during a 30-minute or 1-hour window are used to predict price changes for the same interval [8], [24].

In this research, we proposed a synchronized framework combines the strengths of both methods. Instead of using fixed daily or intra-period windows, we employ a dynamic time-

2690-828X/25/$31.00 ©2025 IEEE
DOI 10.1109/BigDataService65758.2025.00026

window strategy. Specifically, we investigated how to determine optimal time windows for relating sentiment analysis results to actionable market forecasting, considering the lag of the stock market's response to sentiment. Our goal is to identify patterns between changes in stock market prices and the sentiment conveyed through tweets within defined time windows. By doing so, we aim to make our findings practical and useful for financial market analysts. This hybrid approach enhances prediction accuracy by dynamically adapting to the temporal and contextual nature of both tweets activity and stock market movements.

3) Forecasting the stock market changes: Time-series forecasting involves predicting future values based on historical data. Traditional classification models have shown effectiveness in short-term forecasting [8], but they often struggle with capturing complex long-term dependencies. Recently, various deep learning models, including Recurrent Neural Networks (RNNs), Multi-Layer Perceptrons (MLPs), Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Generative Adversarial Networks (GANs) [19]–[25], have been explored for time-series forecasting. These models frequently outperform classical statistical approaches such as ARIMA and GARCH [3].

Motivated by recent advancements in large language models for Natural Language Processing (NLP), generative AI forecasting models have gained attention [26]. Large pretrained time-series foundation models offer several advantages over traditional approaches. First, their self-attention mechanisms, inspired by transformer architectures, improve long-range dependency modeling, addressing the limitations of classical models and recurrent networks. Second, they leverage vast amounts of historical and synthetic time-series data, allowing them to learn generalizable temporal patterns that can be transferred across domains. This pretraining enables robust zero-shot and few-shot forecasting, reducing the need for extensive labeled training data. Third, these models are highly flexible, capable of handling varying history lengths, prediction length, and temporal granularities within a unified framework. In this project, we evaluated both traditional classification models and a pretrained time-series foundation model to assess their forecasting performance of stock market.

II. Contribution/Methods

The design and implementation of this project involves four key components: A) Collecting high-quality training datasets, including tweets and stock market data; B) Building the synchronized framework to align tweets with corresponding stock market changes; C) Conducting sentiment analysis on the tweets; and D) Comparing the forecasting performance of traditional classification models and pretrained time-series foundation models.

A. Collecting Datasets

1) Tweets Dataset: Our first task was to collect high-quality tweets for training and prediction. It's impractical to analyze all tweets posted daily, especially since not all of them influence the stock market. To address this challenge, we proposed collecting only tweets containing financial keywords. The following two datasets were examined: the Financial PhraseBank, as detailed in [12], which includes 4,845 English sentences randomly selected from financial news on the LexisNexis database, and a collection of approximately 5,000 news article headlines with one or two sentences related to U.S. economic performance [13]. Both datasets were annotated by human experts to indicate their potential impact on stock market fluctuations. The most frequent words were extracted from the above mentioned two datasets, after removing English stop words, across three sentiment classes (Positive, Negative, and Neutral). The following unigrams and bigrams were identified as financial keywords: **stock market, economy, bond, inflation, economic**. Consequently, only tweets containing at least one of these financial keywords were collected for training and prediction purposes.

2) Stock Mareket Dataset: The stock market comprises a vast number of companies, making its measurement a significant challenge. Our methodology for assessing changes in the United States stock market relies on market index data rather than the performance of individual company stocks. To provide a comprehensive analysis, we collected historical data from three major indices: the S&P 500 (S&P), the Dow Jones Industrial Average (Dow Jones), and the NASDAQ-100 (NASDAQ). This approach allows us to capture a broad spectrum of market activity and trends, ensuring a more robust and representative assessment of the overall market dynamics.

B. Building a Synchronized Framework

Recognizing the stock market price adjustments will take time after the information from social media platforms is digested, we proposed a synchronized framework aligning tweets with stock performance as shown in Fig. 1. This framework is designed with three parameters: the duration of the tweets observation window – **F** minutes, the lag of the stock market's response to sentiment (Offset) – **O** minutes, and an additional period beyond the tweets window – **E** minutes. The idea is that tweets posted within **F** minutes are anticipated to start influencing stock market performance **O** minutes later, with this effect continuing for an additional duration of **E** minutes after the tweets window ends. Therefore, the size of the stock window, or the cumulative impact duration on stock market behavior, spans **F+E-O** minutes, which starts at **F-O** minutes after the tweets window starts.

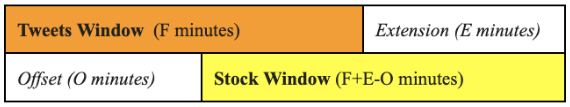

Fig. 1. The Synchronized Framework for Alignment

C. Sentiment Analysis of Tweets

To predict the stock market changes, we believe the following information is valuable: 1) How confident the collected tweets are about the stock market changes. 2) The overall sentiment towards the stock market. 3) The sentiment towards the stock market's past, present, and future, respectively. 4) Whether the tweets contain toxic information. Since the entailment method can work with any number of class labels, and more than one candidate label can be correct, we specified the following 11 non-exclusive labels: *stock market, toxic, positive, negative, neutral, positive about past, positive about present, positive about future, negative about past, negative about present*, and *negative about future*. The **BART** model [9] pre-trained on the Multi-Genre Natural Language Inference (MultiNLI) corpus [7] was adopted to conduct the zero-shot classification with the above mentioned 11 class labels. In this project, **ChatGPT** [15] of OpenAI was adopted for comparison. It classifies the sentiment of collected tweets as *positive*, *negative* and *neutral*.

D. Model Training and Forecasting

TimesFM, a time-series foundation model introduced in [2], is a decoder-style attention model pre-trained on a large corpus of time-series data, encompassing both real-world and synthetic datasets. Experimental results demonstrate that it can generate accurate zero-shot forecasts across various domains, forecasting horizons, and temporal granularities. In this study, we adopted **TimesFM** to perform zero-shot forecasting on the dataset generated through the alignment outlined in the framework of Fig. 1. To benchmark its performance, we tested Random Forest Classifier, Histogram-based Gradient Boosting Classification Tree as traditional classification models for comparison.

III. Experiment

This experiment is designed to address three key questions essential for enhancing the accuracy and reliability of stock market forecasting. First, we investigate whether the textual entailment method, utilizing a pretrained large language model, can effectively analyze tweets sentiment, ultimately contributing to accurate predictions at the end of the pipeline. Second, we compare the prediction performance of traditional classification models with that of a pre-trained time-series foundation model. Third, we studied the configuration of the proposed synchronized framework. To optimize this framework, we systematically fine-tune three critical parameters: the duration of the tweets observation window (**F**), the lag time between sentiment expression and market response (**O**), and the extended period beyond the tweets window (**E**). Our goal is to achieve high prediction accuracy while ensuring the model's robustness and practicality for real-world applications.

A. Collecting Datasets

1) Tweets Datasets: A dataset of 188,020 tweets originating from the United States in the year 2021 was collected. These tweets contain at least one of the aforementioned financial keywords (see Section II-A).

2) Stock Index Datasets: We collected historical data for the year 2021 from three major indices: S&P, Dow Jones, and NASDAQ. Each index dataset includes 1-minute interval data recorded daily (business days) from 9:30 AM to 4:00 PM US Eastern Time, resulting in approximately 99,000 entries per index.

B. Creating Training Datasets

To implement the alignment of tweets windows and stock windows, the parameter **F** is chosen from the set {10, 15, 20, 25, 30, 35, 40, 45, 50, 55}, the parameter **O** from {10, 15, 20, 25}, and the parameter **E** from {5, 10, 15, 20, 25, 30}, respectively. Since the size of the stock window, **F+E-O**, can not be negative or zero, we tested combinations that resulted in a positive stock window size ranging from 10 to 60 minutes.

As described in Section II-C, two sets of tweets sentiment information were derived. The first set, **ChatGPT**, was obtained using OpenAI's **gpt-3.5-turbo** model and includes the proportions of *positive, negative*, and *neutral* tweets within each tweets window, along with the total tweets count. The second set, **BART**, was generated using **bart-large-mnli** [7]. To refine the dataset, tweets with the *stock market* class label probability of less than 0.5 were removed. For the remaining tweets, the probabilities of the 10 class labels within each tweets window were averaged across all tweets in that window.

Considering market dynamics, which account for both opening and closing prices, stock market performance is represented by the average difference between the opening and closing prices within a given stock window, denoted as **Market_Change**.

Table I presents two types of datasets generated. For each target index — S&P, Dow Jones, and NASDAQ — data points from a 12-month period of the year of 2021 were merged for both dataset types. On average, each dataset contained 72,000 data points per stock window size, ranging from 10 to 60 minutes.

TABLE I
Features of Training Datasets

Type	# of Features	Features	Target
ChatGPT	4	positive, negative, neutral, Number of Tweets	Market_Change
BART	10	toxic, positive, negative, neutral positive about past, positive about present positive about future, negative about past negative about present,negative about future	Market_Change

C. Experiment Results

1) Performance of Traditional Classification Models: To evaluate performance of traditional classification models, we implemented a 10-fold cross-validation approach, using the accuracy as the primary evaluation metric. The average accuracy scores of each run are recorded. Two models, Histogram-based Gradient Boosting and Random Forest Classifier, each with **BART** and **ChatGPT** sentiment analysis methods, were tested. Fig. 2 shows the performance of forecasting index S&P

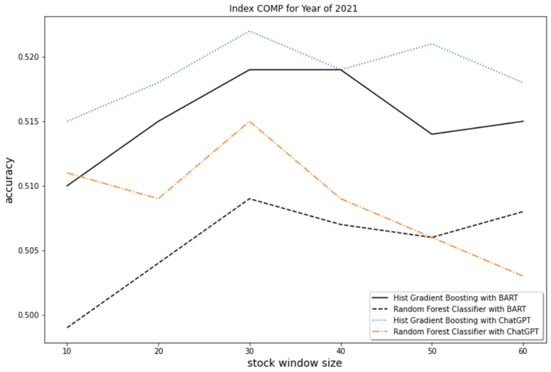

Fig. 2. Performance of Traditional Classification Models on S&P

across varying stock window sizes in the year 2021. Overall, both models do not perform well - the accuracy values are below 52%, while the selection of the optimal stock window size appears crucial, with 30 to 40 minutes yielding the highest accuracy for most models. The results suggest that HistGB outperforms RFC for stock market prediction across all stock window sizes and that **ChatGPT**-based sentiment analysis provides an advantage over **BART**-based analysis.

2) Performance of Time Series Model - TimesFM: We utilized the **timesfm-2.0-500m** checkpoint of **TimesFM** [2], a model designed for time series forecasting with support for context lengths up to 2048 timepoints and flexible horizon lengths, optionally incorporating a frequency indicator. In our experiment, we performed the high-frequency forecasting, the batch size was set to 128, with the context length, representing the number of past time steps used as input, ranging from 8 to 128, and the horizon length, indicating the number of future time steps being predicted, ranging from 12 to 1. Datasets of two types of tweets sentiment information were fed into the pre-trained model to predict the **Market_Change** of three indexes.

To fine-tune **TimesFM** model, different combinations of context length and horizon length were tested. Fig. 3 presents the performance of **TimesFM** with **BART** sentiment information in forecasting NASDAQ market changes for the stock window size of 20 minutes (F=20 min., O=10 min., and E=10 min.). The trend shows that as the horizon length increases, accuracy generally declines across all context lengths. However, longer context lengths (e.g., 128) tend to maintain higher accuracy compared to shorter ones, suggesting that incorporating a larger historical window improves the model's forecasting ability. Since the model takes longer to process a longer context length, we set the context length to 16 and the horizon length to 2 in the following experiments.

Datasets with **ChatGPT** and **BART** sentiment information were fed into the pre-trained model to forecast the **Market_Change** of three indexes. The best performance for each stock window size and each index is recorded. Fig. 4 presents the performance for forecasting changes of Dow Jones. The performance of both models indicates that larger stock window sizes improve forecasting performance, with the most significant increase observed between 10 and 20 minutes. At larger stock window sizes (50 min. – 60 min.), **TimesFM** with **BART** achieves an accuracy close to 100%, while **TimesFM** with **ChatGPT** slightly lags but remains highly accurate.

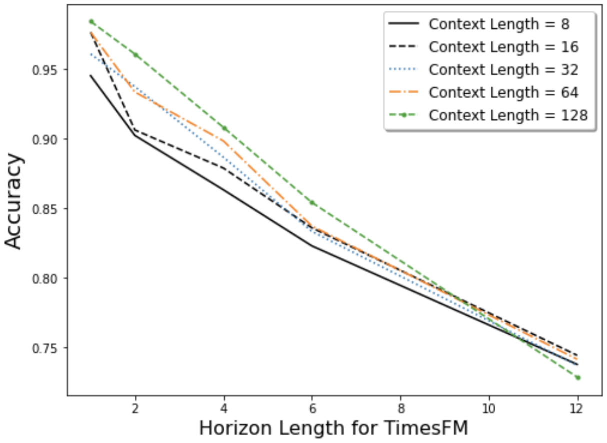

Fig. 3. Performance of TimesFM Forecasting NASDAQ

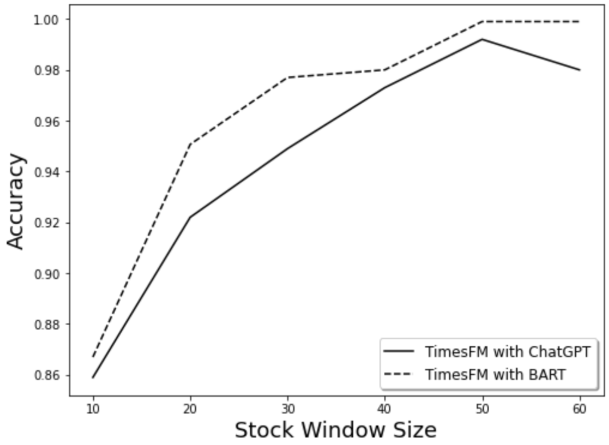

Fig. 4. Performance of TimesFM on Down Jones

The performance for forecasting NASDAQ and S&P changes follows similar trends. Overall, **TimesFM** with **BART** outperforms **TimesFM** with **ChatGPT** in forecasting stock market changes.

3) Study of Different Configurations of the Synchronized Framework: Table II presents the performance of the **TimesFM** model in forecasting S&P market changes across different stock window sizes (20 minutes and 30 minutes). Overall the model consistently achieves high accuracy across various parameter configurations, with values mostly above 0.961 and peaking at 0.999. The results suggest that increasing the stock window size improves predictive accuracy, particularly in cases with higher frame sizes and well-optimized offset-extension values.

TABLE II
PERFORMANCE OF TIMESFM ON FORECASTING S&P

Frame	Stock Window Size = 20 min.			Stock Window Size = 30 min.		
	Offset	Extension	Accuracy	Offset	Extension	Accuracy
15	10	15	0.961	10	25	0.969
20	10	10	0.961	10	20	0.969
	15	15	0.961	15	25	0.969
25	10	5	0.961	10	15	0.969
	15	10	0.961	15	20	0.969
	20	15	0.953	20	25	0.961
30				10	10	0.969
	15	5	0.961	15	15	0.969
	20	10	0.953	20	20	0.961
	25	15	0.945	25	25	0.969
35				10	5	0.969
				15	10	0.999
	20	5	0.977	20	15	0.984
	25	10	0.977	25	20	0.977

It is observed that accuracy tends to be higher when the tweets frame size is larger than 30 minutes in most cases. For frame sizes below 30 minutes, the forecasting performance remains stable regardless of further reductions in frame size. The results reinforce the effectiveness of **TimesFM** in leveraging time-series sentiment data for financial forecasting. Future experiments could focus on optimizing offset and extension values for lower frame sizes to enhance model robustness.

IV. CONCLUSION AND FUTURE WORKS

The findings of this study provide a foundation for advancing financial prediction methodologies by incorporating sentiment-driven time-series modeling. By aligning tweet sentiment with market fluctuations and applying both traditional classification models and pre-trained time-series foundation models, we systematically evaluated their predictive capabilities. Our findings indicate that the **TimesFM** model achieves high forecasting accuracy by capturing complex temporal dependencies, particularly when utilizing larger tweets frame window and well-optimized framework configurations. Additionally, we observe that sentiment analysis with pre-trained large language models, such as **BART** and **ChatGPT**, achieve high forecasting performance.

Future research could focus on combining information collected from more diverse social media platforms, integrating multimodal financial data to improve robustness, and exploring real-time applications of sentiment-aware market forecasting models. While the framework achieves strong predictive results, it currently lacks interpretability, which is often desired in financial applications where users need to understand the rationale behind forecasts. Future work will incorporate tools like SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), or attention heatmaps to identify influential inputs and make the model's predictions more transparent and practically useful.

REFERENCES

[1] M. Chang, L. Ratinov, D. Roth, and V. Srikumar,”Importance of semantic representation: Dataless classification,” In AAAI, pp. 830—835, 2008.

[2] A. Das, W. Kong, R. Sen, and Y. Zhou, ”A decoder-only foundation model for time-series forecasting,”, International Conference on Machine Learning, 2024, https://arxiv.org/abs/2310.10688

[3] G. EP Box and G. M. Jenkins,”Some recent advances in forecasting and control,” Journal of the Royal Statistical Society. Series C (Applied Statistics), 17(2):91–109, 1968.

[4] E. Fama,“Efficient Capital Markets: A Review of Theory and Empirical Work,” 1970.

[5] D. Jurafsky, and J. H. Martin, Speech and Language Processing (3rd ed.). http://web.stanford.edu/ jurafsky/slp3

[6] https://huggingface.co/facebook/bart-large-mnli

[7] https://huggingface.co/datasets/nyu-mll/multi_nli

[8] X. Guo, and J. Li, “A Novel Twitter Sentiment Analysis Model with Baseline Correlation for Financial Market Prediction with Improved Efficiency,” 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, pp. 472–477, 2019

[9] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, ”Bart: Denoising Sequence-To-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension,” arXiv preprint, arXiv:1910.13461, 2019.

[10] L. Thripuranthakam, S. B. Gohad, H. N. S. Singh, G. Charan, and R. Amaradeep, “Stock Market Prediction Using Machine Learning and Twitter Sentiment Analysis: A Survey,” IJRESM, vol. 5, no. 4, pp. 144—149, Apr. 2022.

[11] V. S. Pagolu, K. N. Reddy, G. Panda, and B. Majhi,“Sentiment Analysis of Twitter Data for Predicting Stock Market Movements,” SCOPES, 2016.

[12] P. Malo, A. Sinha, P. Takala, P. Korhonen, and J. Wallenius, “Good debt or bad debt: Detecting semantic orientations in economic texts,” Journal of the Association for Information Science and Technology 65, 4 (2014), pp.782—796. https://doi.org/10.1002/asi.23062 arXiv:arXiv:1307.5336v2

[13] U.S. economic performance based on news articles, Source: https://www.crowdflower.com/data-for-everyone/

[14] V. A Kharde, and S. S. Sonawane,“Sentiment analysis of twitter data: A survey of techniques,” International Journal of Computer Applications (0975 – 8887), Volume 139 – No.11, April 2016.

[15] Z. Wang, Q. Xie, Y. Feng, Z. Ding, Z. Yang, and R. Xia ,“Is ChatGPT a good sentiment analyzer? A preliminary study,” arXiv:2304.04339v2 [cs.CL] 17 Feb 2024.

[16] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[17] V. A Kharde, and S. S. Sonawane,“Sentiment Analysis of Twitter Data: A Survey of Techniques,” International Journal of Computer Applications (0975 – 8887), Volume 139 – No.11, April 2016.

[18] W. Yin, J. Hay, and R. Dan, ”Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach,” arXiv preprint arXiv:1909.00161, 2019.

[19] M. Kesavan, J. Karthiraman, T. E. Rajadurai, and S. Adhithyan, “Stock Movement Prediction with Historical Time Series Data and Sentimental Analysis of Social Media Data”, 2020, International Conference Intelligent Computing and Control Systems, pp. 477-482, 2020.

[20] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, ”N-beats: Neural basis expansion analysis for interpretable time series forecasting,” In International Conference on Learning Representations, 2019.

[21] S. Priyank, V. Bajpai, and A. Bansal. ”Stock price prediction using BERT and GAN,” *arXiv preprint arXiv:2107.09055*, 2021

[22] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, ”Deepar: Probabilistic forecasting with autoregressive recurrent networks,” International Journal of Forecasting, 36(3): pp. 1181–1191, 2020.

[23] M. G. Sousa, K. Sakiyama, L. S. Rodrigues, P. H. Moraes, E. R. Fernandes, and E. T. Matsubara, ”BERT for stock market sentiment analysis,” In *2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI)* pp. 1597-1601

[24] M.L. Thormann, J. Farchmin, C. Weisser, R. M. Kruse, B. Safken, and A. Silbersdorff, ”Stock price predictions with LSTM neural networks and twitter sentiment,” Stat Optim Inf Comput 9 (2): pp. 268–287, 2021.

[25] M. Saloni, M. Sahitya, S. Sudheer, V. Parag, and D. C. Anastasiu, ”Stock Price Prediction Using News Sentiment Analysis,” 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService) pp. 205-208, 2019

[26] https://databricks-industry-solutions.github.io/transformer_forecasting/