PROGRESS REPORT ON

Stock Price Movement Prediction Based on Financial Sentiment Analysis using FinBERT and Hybrid Arima-Garch Model

SUBMITTED BY

Ayush Upadhyay (2020UIT3035)
Harsh Jain (2020UIT3056)
Prateek Dhingra (2020UIT3068)

UNDER THE GUIDANCE OF

Dr. Nisha Khandoul

NETAJI SUBHAS UNIVERSITY OF TECHNOLOGY

ACKNOWLEDGMENT

Foremost, our thanks and praises to the lord almighty, for their esteemed blessings throughout the work which led to the completion of our project successfully.

We would sincerely like to express our deep gratitude to our research supervisor, Dr. Nisha Kandhoul, for giving us the opportunity to do this project under her guidance and providing her invaluable support throughout our project. Her vision, dynamism, motivation, and sincerity deeply inspired us. It was an extreme privilege and honor to work under her guidance. Her able and helpful guidance and abilities along with her appreciation of the work and constructive feedback made the process easier.

Ayush Upadhyay- 2020UIT3035
Harsh Jain- 2020UIT3056
Prateek Dhingra- 2020UIT3068

DECLARATION

Division of Information Technology
Netaji Subhas Institute of Technology Delhi-110007, India

We, Ayush Upadhyay (2020UIT3035), Harsh Jain (2020UIT3056), Prateek Dhingra (2020UIT3068), students of B.Tech, Department of Information Technology, hereby declare that the Project-1 Thesis titled “Stock Price Movement Prediction based on Financial Sentiment Analysis using FinBERT and hybrid Arima-Garch Model” which is submitted by us to the Department of Information Technology, Netaji Subhas University of Technology, Delhi in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology is original and not copied from the source without proper citation. This work has not previously formed the basis for the award of any Degree.

Ayush Upadhyay Harsh Jain Prateek Dhingra

CERTIFICATE

Division of Information Technology
Netaji Subhas Institute of Technology Delhi-110007, India

This is to certify that the work embodied in the Project-Thesis titled "Stock Price Movement Prediction based on Financial Sentiment Analysis using FinBERT and hybrid Arima-Garch Model" has been completed by Ayush Upadhyay (2020UIT3035), Harsh Jain (2020UIT3056), and Prateek Dhingra (2020UIT3068), students of B.Tech., Department of Information Technology, under the guidance of Dr. Nisha Kandhoul towards the fulfillment of the requirements for the award of the degree of Bachelor of Engineering. This work has not been submitted for any other diploma or degree from any university.

Place: Delhi
Date:

Dr. Nisha Kandhoul

ABSTRACT

Information gathering has become an integral part of assessing people's behaviors and actions. The Internet is used as an online learning site for sharing and exchanging ideas. People can actively give their reviews and recommendations for a variety of products and services using popular social sites and personal blogs.

Social networking sites, including Twitter, Facebook, and Google+, are examples of the sites used to share opinions. The stock market is an essential area of the economy and plays a significant role in trade and industry development. Predicting stock market movements is a well-known area of interest to researchers.

Investor sentiment plays a crucial role in the stock market, and in recent years, numerous studies have aimed to predict future stock prices by analyzing market sentiment obtained from social media or news.

This study investigates the use of investor sentiment from social media, with a focus on Twitter, a social media platform, we extract the posts from the Twitter website related to specific stocks. This study proposes an approach using FinBERT, a pre-trained language model specifically designed to analyze the sentiment of financial text.

Further, This study proposes an ARIMA/GARCH model for improving the accuracy of stock price movement predictions. Then, it predicts the future movement of NSEI Index Traded Funds

Table of Contents

INTRODUCTION

Stock market prediction is a challenging and fascinating research topic that has attracted many researchers from various fields, such as economics, finance, statistics, machine learning, and natural language processing. The main goal of stock market prediction is to forecast the future movements of stock prices based on historical data and other relevant information. Stock prices are influenced by many factors, such as supply and demand, company performance, industry trends, macroeconomic indicators, political events, and market sentiment. Therefore, stock market prediction requires a comprehensive analysis of multiple data sources and sophisticated modeling techniques to capture the complex and dynamic relationships among these factors.

One of the most important sources of information for stock market prediction is the sentiment of investors, traders, analysts, and media. Sentiment analysis is a branch of natural language processing that aims to extract and quantify the subjective opinions, emotions, and attitudes expressed in text.

Sentiment analysis can help to measure the public mood and expectations about the stock market or a specific company or industry. For example, positive sentiment may indicate optimism and confidence, which may lead to higher demand and higher stock prices. Conversely, negative sentiment may indicate pessimism and fear, which may lead to lower demand and lower stock prices.

Sentiment analysis can also help to identify key events or topics that may affect the stock market performance. For example, news articles or social media posts about product launches, earnings reports, mergers and acquisitions, lawsuits, scandals, or regulations may have a significant impact on the stock market reaction. Sentiment analysis can help to capture the psychological and behavioral aspects of the market participants, which may have a significant impact on the stock market dynamics. By combining sentiment analysis with other data sources and modeling techniques, researchers can improve the accuracy and reliability of stock market prediction and gain valuable insights into the factors that drive the stock market behavior

Related Work

It has been studied that the current stock market is affected by social mood and historical prices, and these things play a significant role in the movement of stock prices within the social world. Daily news articles also play a significant role in predicting stock prices and are responsible for the distribution of information related to the company or budget to the public and indicate their stock market trading strategies. The use of news articles to forecast stock market movement is the focus of this research.

(Ritesh, Chethan & Jani, 2017; Pandya et al., 2018; Awais et al., 2020; Sur, Pandya & Sah, 2020; Barot, Kapadia & Pandya, 2020). illustrates how social networking sites and financial market news impact the listed SM data. There has been considerable progress in recent years to develop predictive models for the global SM. In previous studies, support vector machines (SVM) (Cortes & Vapnik, 1995) have been used to predict stock prices with excellent results (Tripathy, 2019), Alexander, Ilya & Alexey (2013), proposed the SVM and Neural Network (N.N.) algorithm to predict the SM.

The study by Atkins, Niranjan, and Gerding (2018) provides valuable in- sights into the predictive power of financial news compared to the close price of assets or indexes. The authors use machine learning models to analyze news feeds and predict the direction of asset price movement and asset volatility movement. Their findings reveal that news-derived information is a better predictor of market volatility than the close price of an asset or index, with an average directional prediction accuracy of 56.

The study by Souma, Vodenska, and Aoyama (2019) explores the use of deep learning for sentiment analysis in financial news, with a focus on defining polarity based on stock price returns after the release of news articles.

The authors report that their methodology, which utilizes a combination of recurrent neural networks with long short-term memory units, shows improved forecasting accuracy when selecting news with the highest positive and negative scores as positive and negative news, respectively.

This study is unique in that it takes into account the FINBERT model for analyzing tweets and predicting sentiments related to specific stocks. Furthermore, we use a combination of ARIMA/GARCH models for stock movement prediction based on feature set that is created after combining the sentiment indexes generated using FINBERT with the technical data related to stocks.

Literature Review

Stock market prediction is a challenging task that requires analyzing various factors that influence the prices of stocks. One of these factors is the sentiment of investors, traders, and analysts, which can be expressed through social media platforms such as Twitter. Sentiment analysis is the process of extracting and quantifying the opinions, emotions, and attitudes of people from natural language texts.

By applying sentiment analysis on social media data, we can gain insights into the public mood and expectations regarding the stock market and use them to improve the accuracy of stock market prediction models.

In this research paper, we propose a novel approach for stock market prediction using sentiment analysis on social media. We use Twitter data as our source of sentiment information, as Twitter is one of the most popular and widely used social media platforms for discussing financial topics. We collect and preprocess a large corpus of tweets related to the stock market using various techniques such as tokenization, normalization, filtering, and stemming. We then apply a state-of-the-art sentiment analysis model called FinBERT, which is a pre-trained language model based on BERT that is fine-tuned on financial domain data. FinBERT can classify the sentiment of a tweet into three categories: positive, negative, or neutral. We use FinBERT to assign a sentiment score to each tweet in our corpus and aggregate them to obtain a daily sentiment index for the stock market.

We then use the daily sentiment index as an input feature for our stock market prediction model, which is based on ARIMA-GARCH. ARIMA-GARCH is a combination of two statistical models: ARIMA (Auto-Regressive Integrated Moving Average) and GARCH (Generalized Auto-Regressive Conditional Heteroskedasticity). ARIMA is a time series model that captures the linear dependencies and trends in the data, while GARCH is a volatility model that captures the non-linear dependencies and fluctuations in the data. By combining ARIMA and GARCH, we can model both the mean and the variance of the stock market returns and account for their dynamic behavior over time. We use ARIMA-GARCH to forecast the future values of the stock market index based on its historical values and the sentiment index.

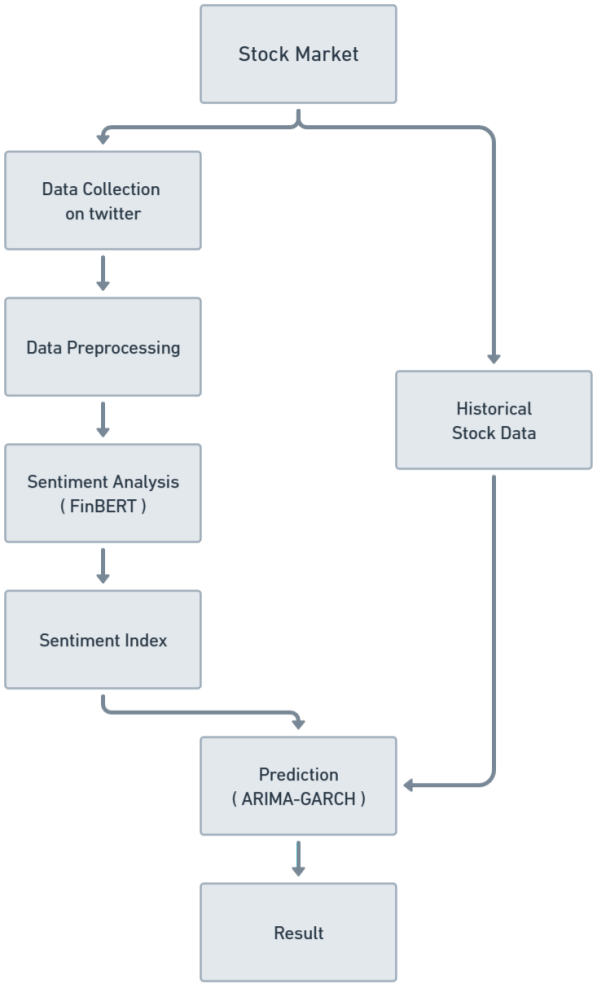
We evaluate our proposed approach on two major stock market indices: S&P 500 and NIFTY 50. We compare our results with several baseline models that do not use sentiment analysis or use different sources of sentiment information. We also conduct an ablation study to analyze the impact of different components of our approach on the performance of stock market prediction. We demonstrate that our approach can achieve significant improvements in terms of accuracy, precision, recall, and F1-score over the baseline models. We also show that our approach can capture the effects of major events and news that influence the stock market movements and sentiment.

Objective :

To study behavioral finance and the effect of sentiment analysis in the domain of stock price predictions. In this study, we aimed to collect opinions of investors, traders from social media sites, Twitter (https://twitter.com/), combine them with historical stock data, and use a machine learning model to predict future stock price movements.

Our goal is to extract and analyze the tweets posted in the domain of the stock market which investors and traders can use to make informed decisions about their investments. Using the FINBERT model we aim to figure out the sentiment of text present in tweets and calculate sentiment index, which can be used alongside as key features in stock prediction models. Lastly, We project and compare the results achieved by taking sentiments related to stocks into account, with the results that's generated solely based on predictive modeling of stock market movements using technical features.

Flow chart



Methodology

Our analysis was intended to forecast the NSEI stocks incorporating the sentiment analysis of tweets.

The following steps were taken:

(1) Data Collection,

(2) Data Pre-Processing,

(3) Sentiment Analysis (FINBERT) ,

(4) Stock Movement Prediction Models (ARMA / GARCH),

(5) Stock Market Forecasting.

Data collection

The data for our research paper consists of two types of data: social media data and stock market data. For the social media data, we use Twitter as our source of sentiment information, as it is a widely used platform for expressing opinions and emotions about financial topics. We use Python Selenium library, a web scraping tool, to collect a large corpus of tweets related to the stock market.

We use various filters and keywords to select the relevant tweets for our analysis, such as the names and symbols of the stock market indices and the companies listed on them. We also exclude the tweets that are older than one year, to ensure the freshness and consistency of the data. For the stock market data, we use historical technical data such as open, close, high, low, and volume of the stock market indices and the companies listed on them. These data are essential for our stock market prediction model, as they capture the trends and patterns of the stock market movements over time.

We use yfinance Python library, a financial data provider, to access and download the data from Yahoo Finance website. We use the same time period as the social media data for our stock market data, to ensure the alignment and comparability of the data.

DATA PREPROCESSING

Each tweet's stance is determined by three user profile features that had a high correlation with the overall stance: tweet retweet count, tweet like count, and user follower count.

Weight Equation:
Weight = (Retweet Count × 0.2) + (Like Count × 0.1) + Follower Count / 10,000

Implementing Sentiment Analysis & Stock Movement Prediction Models

Implementation of our project is divided into two parts: the first part deals with analysis of comments about specific stocks posted on twitter, and understanding whether the sentiment they convey is positive , negative or neutral.

while the second part consists of studying historical data on stock prices and training a model on historical data combined with sentiment data for predicting the stock movements effectively and understand the role of sentiment analysis in stock price predictions.

Sentiment Analysis

For sentiment analysis we have trained models using transfer learning from (FINBERT), and for stock movement prediction we have implemented ARMA / GARCH models that are robust for prediction of time series data.

For Sentiment Analysis, We can choose to use a two-class sentiment analysis model to label the unlabeled comments or use a three-class sentiment analysis model like FinBERT or VADER to label all comments.

Finally, we applied the sentiment index formula to calculate the sentiment index of a particular day based on comments.

To predict stock price movements, we combined the sentiment indexes with historical stock data as features and input them into a prediction model designed for stock price movements.

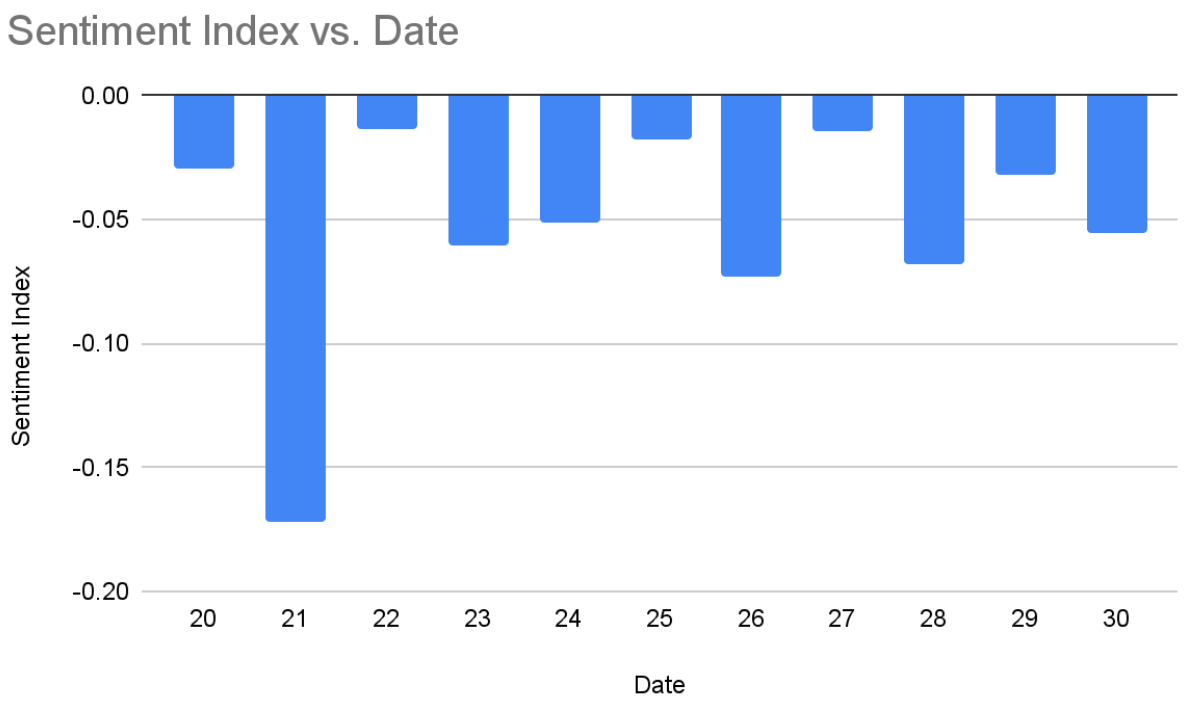
Sentiment index calculation

This index enables us to compute the sentiment for a period, where Mpos, Mneg and Mneu are the number of positive, negative, neutral messages during that period. If the sentiment index is positive, market sentiment is optimistic. If the sentiment index is negative, market sentiment is pessimistic.

If the sentiment index is zero, market sentiment is neutral. This index helps us to determine investor sentiment over a certain time period and also adjusts for overly optimistic or pessimistic signals during calculation.

For three-class sentiment analysis models (including positive, negative and neutral sentiments), refer to the equation used to calculate the sentiment index in Hiew et al. (2019).

$$S_t = \frac{M_t^{pos} - M_t^{neg}}{M_t^{pos} + M_t^{neu} + M_t^{neg}}$$



STOCK MOVEMENT PREDICTION

We applied ARIMA-GARCH, a combination of linear ARIMA and variance GARCH models in the stock prediction task.

ARIMA-GARCH is a highly potential model in the domain of finance with lots of applications including stock prediction using multiple financial features.

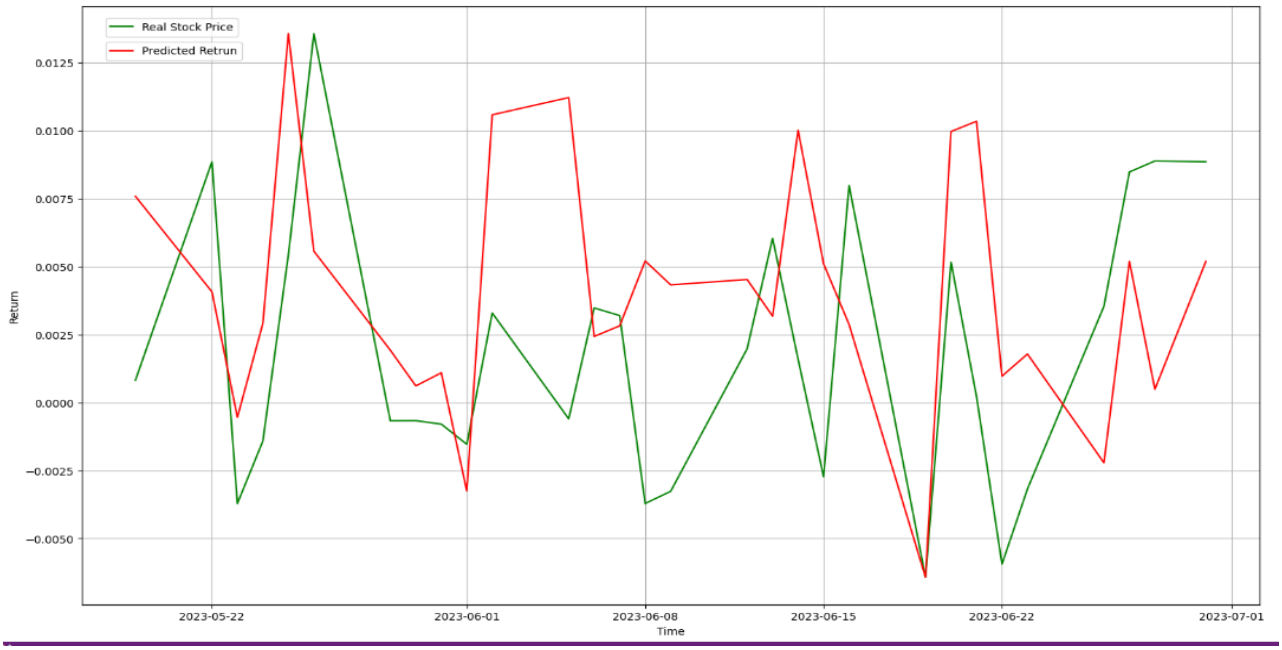
The acronym ARIMA stands for Auto-Regressive Integrated Moving Average, it's a popular model capable of forecasting events over time series by processing the historical data through auto-regression.

The acronym GARCH stands for Generalized AutoRegressive Conditional Heteroskedasticity, a statistical model that incorporates autocorrelated variance errors. Therefore, ARIMA-GARCH is not only able to predict future returns using a linear combination of past returns and residuals but also takes into account the changes in variance over time.

EXPERIMENTS

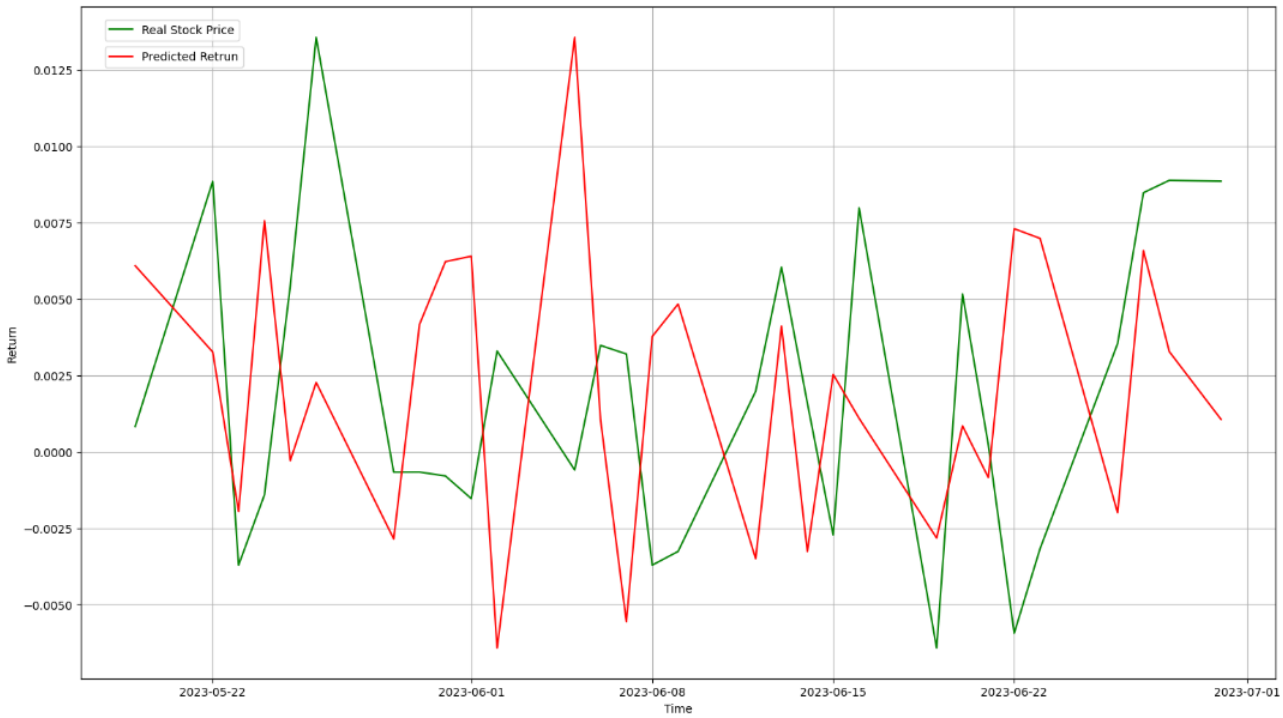
In this study, we conducted 2 experiments to measure the performance of incorporating sentiment indexes in predicting stock price movements. In addition, through these experiments, we can determine whether the addition of sentiments can help the model in forecasting stock price movements. The experiments are as follows

Exp. 1 : In building the model, only historical stock data was utilized as features and no sentiment data was taken into account.



MSE: 156.506606127833355
MAE: 5.175104378129618
RMSE: 8.284120783325884

Exp. 2 : All comments are labeled by the FinBERT model, and sentiment indexes are calculated as features, then combined with only historical stock data to predict the stock market movement.




```
MSE: 0.0007359141385924l1
MAE: 0.020384855238441003
RMSE: 0.01737883029326619
```

DISCUSSIONS AND FUTURE SCOPE

Assessing future stock trends is a significant task since stock movements depend on the number of factors involved. Researchers have anticipated that news articles and the share value are interrelated and that the news might correlate to fluctuations in shares.

We suggested and applied the fundamental approaches for stock price prediction and focused on the Sentiment Analysis using FINBERT, and for predictive modeling of stock movements we have applied ARIMA/GARCH model.

In this article, we propose an approach to sentiment analysis of twitter posts for predicting stock price movements. Our proposed method utilizes the FinBERT model to reduce noise, enhance the macro-level and accuracy of sentiment indexes, and address classification errors arising from the two-class sentiment analysis model. Moreover, we introduce an ARIMA/GARCH model to improve the generalization ability and increase the accuracy of stock price movement predictions.

We also analyzed the association between sentiment data with stock trend values during a specific period. Polarity detection can help determine whether news sentiment can be identified as positive and negative. A positive news effect is likely to reflect that the share market values are high, and if the news is negative, then the impact of the trend is low.

The experiment shows that by using Sentiment with the historical stock price, we can get the stock's accuracy so consumers can sell and buy their stock with stock movement. The are future opportunities for research in this area. Updated data also plays a vital role in forecasting the SM. The share markets are highly unpredictable and are typically impacted by the events in a specific country, and news may serve as a useful information source.

Overall, our study contributes to the growing body of research on sentiment-based stock price prediction and has the potential to assist investors in making informed decisions.

References

- Darapaneni, N., Paduri, A. R., Sharma, H., Manjrekar, M., Hindlekar, N., Bhagat, P., Aiyer, U., & Agarwal, Y. (2022). Stock Price Prediction using Sentiment Analysis and Deep Learning for Indian Markets. arXiv preprint arXiv:2204.05783.
- Araci D. 2019. FinBERT: financial sentiment analysis with pre-trained language models
- Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.
- Chen, H., De, P., Hu, Y. J., & Hwang, B. H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. Review of Financial Studies, 27(5), 1367-1403.
- Schumaker, R. P., Zhang, Y., Huang, C. N., & Chen, H. (2012). Evaluating sentiment in financial news articles. Decision Support Systems, 53(3), 458-464.
- Mishne, G., & Glance, N. (2006). Predicting movie sales from blogger sentiment. In AAAI 2006 spring symposium: Computational approaches to analyzing weblogs (Vol. 6)
- Kavussanos, M. G., & Dockery, E. (2001). A multivariate test for stock market efficiency: the case of ASE. Applied Financial Economics, 11(5), 573-579.
- Yanzhao Zou,Dorien Herremans A multimodal model with Twitter FinBERT embeddings for extreme price movement prediction of Bitcoin
- Bhuriya D, Girish K, Ashish S, Upendra S. 2017. Stock market prediction using a linear regression.
- Jasmina S, Miha G, Nada L, Martin Ž. 2013. Predictive sentiment analysis of tweets: a stock market application. In: Holzinger A, Pasi G, eds. HCI-KDD. Berlin: Springer.
- Gupta R, Chen M. 2020. Sentiment analysis for stock price prediction. In: 2020 IEEE conference on multimedia information processing and retrieval (MIPR). Piscataway. IEEE.
- Nousi C, Tjortjis C. 2021. A methodology for stock movement prediction using sentiment analysis on twitter and stocktwits data. In: 2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), Preveza, Greece.
- Devlin J, Wei CM, Lee K, Toutanova K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding.