Analyzing the Effect of Real Time Public Opinion on Stock Prices

Sentiment Analysis on Twitter Opinions and its Correlation with Financial Data

FINAL PROJECT REPORT

Pallav SAHU
pallav.sahu@student-cs.fr

Akshay SHASTRI
akshay.shastri@student-cs.fr

Anmol KATIYAR
anmol.katiyar@student-cs.fr

Arun JEGATHESH
arun.jegathesh@student-cs.fr

1. ABSTRACT

There has been a growing interest in the use of social media data, specifically Twitter, for sentiment analysis and its potential impact on financial markets. The idea is that the collective sentiment of Twitter users may be able to provide insight into the high-level sentiment of the market as a whole and potentially even predict stock price movements.

There have been several studies that have attempted to investigate the relationship between Twitter sentiment and stock prices. Some of these studies have found a positive correlation between the two, while others have found no significant relationship.

One potential explanation for the inconsistent findings in previous research may be attributed to the utilization of various methodologies for measuring sentiment and different time frames for analyzing the data. Additionally, it is possible that the correlation between social media sentiment and stock prices may vary based on the specific company or industry being studied. Another potential limitation could be the presence of bias in the available data, which may provide a limited perspective on true sentiments, and potentially act as a confounding variable in predicting stock price movements.

Overall, the literature suggests that while the use of social media sentiment analysis as a predictor of stock prices has shown some promise, there is still a need for further research to fully understand the relationship between social media sentiment and stock prices. This includes understanding the impact of different methodologies used for measuring sentiment, analyzing data over different time frames, and studying the relationship between social media sentiment and stock prices for different companies and industries. Additionally, more research is needed to determine the practical value of this approach and to determine the limitations and potential biases of the available data in predicting stock prices. Furthermore, it would be important to validate these methods with real-world scenarios before making any investment decisions based on them.

2. KEYWORDS

Sentiment Analysis; Twitter; Stock Price; Classification; Financial Markets; Natural Language Processing; Ensemble Techniques

3. INTRODUCTION / MOTIVATION

This project aims to examine the relationship between real-time public opinion on Twitter and stock prices. To do so, we use sentiment analysis techniques to explore text-based Twitter opinions and compare them to financial data for a given company or sector. We aim to leverage and learn the state-of-art text processing techniques that can be utilized for generating quantifiable sentiments which, further, can be used as a predictor for our Machine learning models to predict the upward or downward movements of the value of stocks in a financial market.

Our goal is to determine whether there is a significant correlation between Twitter sentiment and stock prices and, if so, to what extent Twitter sentiment can be used as a predictor of stock price movements. The primary question we will try to answer through our analyses and models are whether the price movements for a particular stock in previous timeframes like trend of opening and closing prices, once without and then with Twitter sentiment, is predictive of the trajectory for the subsequent day's price movement, and how much value is being added by an additional variable of Twitter sentiment.

Stock price prediction is important for various reasons. For investors, accurate stock price predictions can help to make informed investment decisions and potentially lead to better financial returns. For companies, stock prices can affect their ability to raise capital and can also be a key indicator of the company's financial health and performance.

Additionally, stock prices can have a broader economic impact as they can affect consumer confidence and spending patterns and can also influence the allocation of financial resources. As a result, understanding the factors that influence stock prices and developing effective methods for predicting stock prices is an important area of study. Given the importance of stock prices, it is not surprising that predicting stock prices

has long been a topic of interest in finance and economics.

Thus, this project aims to provide some insight into the role of public opinion in shaping the financial market and to inform investment strategies. Although, a caveat is that there are multitude of factors influencing an investment decision and the financial markets, and the project is likely to be an introduction to understand one of such predictors with a potential scope of improvements in our methodology and analysis.

4. RELATED WORK

There has been a significant amount of previous research that has sought to investigate the relationship between social media sentiment and stock prices. Some studies have found a positive correlation between sentiment on social media platforms and stock prices (Bollen et al., 2011 [2]), while other studies have found little or no correlation between social media sentiment and stock prices (Smith et al., 2021 [15]).

One approach that has gained acceptance in the academic community is the use of natural language processing (NLP) techniques to extract sentiment from social media data (Xiang et al., 2018 [16]; Nguyen et al., 2015 [17]). These studies have found that NLP techniques can effectively extract sentiment from social media data, and that the sentiment extracted from social media data is strongly correlated with stock prices.

Another approach that has been proposed is the use of social media-based stock prediction, where the information available in social media platforms is used to predict stock prices. For example, "Wang et al., 2017 [18]" proposed a multi-task learning model that simultaneously predict the stock price and the sentiments of the tweets, and they found that the predictions were more accurate than using traditional models only based on sentiments or stock prices.

Moreover, state-of-the-art methods currently employed to study the relationship between social media sentiment and stock prices include the use of deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) (Jiang et al., 2021 [19]; Othan et al. 2019 [20]). These studies have found that deep learning techniques can effectively capture the complex relationships between social media sentiment and stock prices, and achieve better performance compared to traditional methods.

One potential reason for the conflicting results in previous research is the use of different methods to analyze the relationship between social media sentiment and stock prices. Some studies have employed machine learning techniques such as support vector machines (SVMs) and decision trees (Bollen et al., 2011 [2]), while others have used more traditional statistical methods such as linear regression.

Despite the conflicting results from previous research, recent studies have found that advanced machine learning and NLP techniques can effectively capture the complex relationships between social media sentiment and stock prices and provide accurate predictions. However, it is important to note that more research is needed to fully understand the relationship between these variables and to develop more robust models for stock price prediction.

In our project, we will be limited to using machine learning, and not use further advanced techniques as Deep neural networks, to analyze the relationship between financial indicators, Twitter sentiment data and stock price trends. Specifically, we will be using Ensemble techniques in terms of Gradient Boosting and Random Forest Algorithms to build predictive models for stock price movements as classifications of upwards or downwards trend versus the previous day prices. We aim to build multiple models with variations in input variables in terms of using sentiment data with technical indicators for stocks as referred from the articles and research mentioned ahead. Further, we aim to test multiple models with hyperparameter tuning across various evaluation metrics like precision, recall and F1-score to justify whether we find the sentiment variables as important feature or not.

In their article, "Sabyasachi Mohapatra et al., 2022 [6]" in the Journal of Risk and Financial Management, the authors have conducted a comprehensive investigation of the use of technical indicators as input variables for stock price prediction using various machine learning techniques. Specifically, the authors have sought to examine the relevance of technical indicators as inputs to predict short-term stock prices using regression models. We aim to utilize their input variables to determine the usefulness of these indicators in predicting the stock trend (classification) for a specific use-case.

The authors have reviewed the literature on the use of technical indicators in stock market prediction and have identified several commonly used indicators that have been found to be relevant to stock price prediction. These indicators include, but are not limited to, moving averages, relative strength index, and Bollinger Bands. The authors then proceed to test the effectiveness of these indicators as input variables in various machine learning techniques such as Random Forest, Support

Vector Machines, and Neural Networks for short-term stock price prediction.

Additionally, the authors have examined by implementing these indicators as input variables in different machine learning algorithms such as Random Forest and SVM. The authors have used various performance metrics such as accuracy, precision, and recall evaluating the performance of the models and the results show the effectiveness of the technical indicators in predicting the stock trend.

In summary, the authors of the research work "Sabyasachi Mohapatra et al., 2022 [6]" have conducted a thorough investigation of the use of technical indicators as input variables in stock trend prediction using machine learning techniques. The authors have reviewed relevant literature, identified commonly used indicators, and evaluated their effectiveness in short-term stock price prediction, providing valuable insights and contributions to the field of stock market prediction.

The above article was primarily based on the research work by "Neely et al., 2014 [7]" and "Dai et al., 2020 [5]", on forecasting stock market returns with technical indicators and macro-economic constraints. "Neely et al. 2014 [7]" and "Dai et al. 2020 [5]" are two research papers that focus on using technical analysis in ensemble models to forecast stock market returns. Technical analysis is a method of evaluating securities by analyzing statistics generated by market activity, such as past prices and volume. Ensemble models, on the other hand, involve using multiple models to make a prediction, with the idea that the combined predictions will be more accurate than any individual model. In these papers, the authors likely explore different technical indicators and ensemble techniques and evaluate their effectiveness in predicting stock market returns.

In "Neely et al. 2014 [7]", the authors propose an ensemble model that combines multiple technical indicators, such as moving averages and relative strength index, with different machine learning algorithms, such as neural networks and decision trees, to predict stock market returns. The paper evaluates the performance of the ensemble model using historical data and compares it to traditional technical analysis methods and single-model techniques.

Similarly, "Dai et al., 2020 [5]" also uses an ensemble approach to forecast stock market returns using technical analysis. The authors propose a technique that combines multiple technical indicators, such as Bollinger Bands and momentum, with different machine learning algorithms, such as Random Forest and Support Vector Machine. The paper evaluates the performance of the ensemble model using historical data and compares it to traditional technical analysis methods and single-model techniques.

Both papers demonstrate that the use of technical analysis in combination with machine learning algorithms in an ensemble model can improve the accuracy of stock market return predictions. These papers propose new approach in stock market prediction by combining technical analysis with machine learning which could be used to improve the performance of trading strategies in the future.

The usefulness of the ensemble models proposed in "Neely et al., 2014 [7]" and "Dai et al., 2020 [5]"in predicting stock market trends is likely to depend on the specific data and market conditions used in the evaluation. Both papers demonstrate improved performance compared to traditional technical analysis methods and single-model techniques when evaluated on historical data. However, it's important to note that the performance of any model, including those proposed in these papers, can vary greatly depending on the specific data and market conditions it is applied to. Additionally, it's also important to note that the stock market is highly dynamic and unpredictable, and it's difficult to accurately predict the stock market trend with a high level of certainty.

It's also worth noting that these papers are based on a specific point in time, and the stock market and technology have evolved since their publication date, which may affect their effectiveness in current market conditions. Therefore, while these papers provide valuable insights into the potential usefulness of ensemble models combining technical analysis and machine learning for stock market predictions, it's important to consider the limitations and to test the models in real-world scenarios before making any decisions.

Overall, our project aims to replicate and extend upon previous research by using a combination of machine learning and traditional statistical techniques to investigate the relationship between Twitter sentiment and stock prices. Our goal is to learn and try to generate a more comprehensive understanding of the extent to which Twitter sentiment can be used as a predictor of stock price movements.

5. METHODOLOGY

5.1 TWITTER SENTIMENT DATA

To prepare and process the input for the model, we used the readily available Twitter libraries, i.e., Twint, since Tweepy gives access to only the tweets from the last 7-days. Even though support for Twint is no longer active and the library is known to have bugs due to Twitter's efforts to block data scraping via external APIs, it was deemed suitable for the task at hand. Using this, we scraped the tweets related to the company name "Tesla" for the last 5 years, from areas within a 5000km range from New York. Due to the challenges, the quality and quantity of the obtained data was suboptimal but sufficient for the purpose of the project. The second step involved cleaning the data, where we used various python libraries to remove all the filler words, punctuation, emojis, and the stop-words.

VADER (Valence Aware Dictionary and sentiment Reasoner) is a lexicon-based tool for sentiment analysis in natural language processing (NLP) developed in 2014. VADER is specifically designed to analyze sentiments from social media text that usually contains mix of text format such as punctuation, emoticons and slang usage. It also considers the use of capitalization, punctuation, and the presence of words associated with positive or negative sentiment in the same sentence (sarcasm). It also considers the context in which words are used, such as negations, and the use of intensifiers and diminishers.

It is available in the NLTK package and can be applied directly to unlabeled text data. VADER sentimental analysis relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores. The sentiment score of a text can be obtained by summing up the intensity of each word in the text.

Each tweet is assigned a sentiment score between -1 and 1, with negative sign denoting more sentiments associated with negative emotions and a positive sign denoting positive sentiments; magnitude of the sentiment score denotes the extent of that emotion. The final data set was created by aggregating the sentiment score by the date, so that we can check its impact on each day's adjusted closing price of the stock.

Since this analysis was lexical and not based on machine learning, we implemented a second mechanism to calculate the sentiments. This time we used a pretrained Bert model. BERT is a method of pre-trained transformer based neural network model for Natural Language Processing tasks developed by Google in 2018. These tasks include entity recognition, question answering and language inferences. One of the unique features of BERT is that it is capable of understanding the context of a word in a sentence by analyzing it from both directions of the word. That is, the model identifies context based on the left and right side of the word rather than just the regular left-to-right interpretation of context from sentences. For our use case we used multilingual BERT model as we expect the tweets to be registered across different languages. BERT's pre-training process allows the model to be fine-tuned on a specific task using a smaller dataset, making it highly

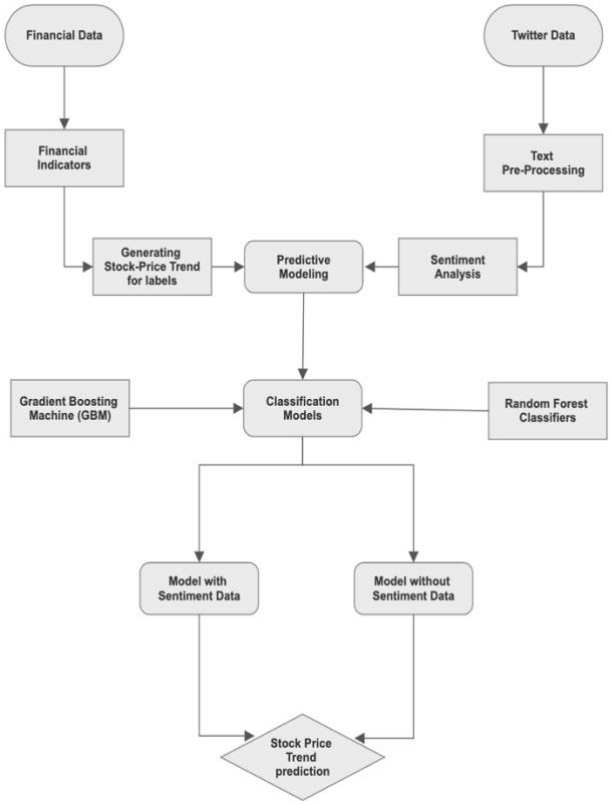

Fig.1 Project Workflow and Methodology

adaptable to different NLP tasks. This has led to significant improvements in many NLP benchmarks, and BERT has been used in a wide range of applications, including language translation, sentiment analysis, and text summarization.

BERT model is used from pre-built packages of AutoTokenizer from the transformers package. As mentioned above, the multilingual BERT tokenizer named 'bert-base-multilingual-uncased-sentiment' is used for getting the sentiments.

5.2 FINANCIAL DATA

In order to incorporate financial market data into our models, we utilized the Yahoo finance python API

"yfinance" to import historical stock market data for the past five years (since 2017) for the stock of Tesla Motors, represented by the ticker symbol "TSLA". By referencing the research done by "Sabyasachi Mohapatra et al., 2022 [6]", "Neely et al., 2014 [7]" and "Dai et al., 2020 [5]", we aimed to replicate and create relevant technical indicators from the Yahoo financial data for our stock. To do this, we utilized the widely used Python library "TA" which specializes in creating datasets on trading indices and technical analysis.

By using the "TA" library, we were able to create variables based on the principles of value, volatility, and momentum across stock price and volume data. Additionally, we tried to incorporate buying and selling pressures, conjugated volume-price trends, and trend-breaking average directional movements as variables. These indicators, such as moving averages, and stochastic oscillator, provide a quantitative measure of the stock's historical performance and are commonly used to identify patterns and trends in the data. By creating these indicators, we aimed to extract valuable information from the financial market data and use it to make predictions about the future performance of the stock.

These variables were selected as they were particularly relevant to the modeling process due to similarities in choice of Machine Learning techniques and excluded macroeconomic data, allowing the models to be based solely on financial market data without the influence of other factors that may have affected the prediction accuracy and other metrics. The list and description of the technical indicators used are as follows:

Variables	Description	Computation
pvo_MbyN	The Percentage Volume Oscillator (PVO) is a momentum oscillator for volume. The PVO measures the difference between two volume-based moving averages as a percentage of the larger moving average. This has been modified w.r.t. to variables 10by20 and 10by40 mentioned in the article.	$\frac{MA(vol,m)}{MA(vol,n)}$
ppo_MbyN	The Percentage Price Oscillator (PPO) is a price momentum oscillator that measures the difference between two moving averages as a percentage of the larger moving average.	$\frac{MA(P,m)}{MA(P,n)}$
pop_M	This refers to a price-based growth variable as momentum lags and is a ratio of price at a previous instant over current.	$\frac{P_{t-m}}{P_t}$
	This is used to identify price movement trends and is replicated from the article.	
msv_M	This refers to a volume-based deviation and variance as a rolling standard deviation over last M days. This is used in tracking volatility and variations in stock traded volume over a time window	MSD(vol)
volume_fi	This refers to the Force Index (FI) based on the combination of volume and price. It illustrates how strong the actual buying or selling pressure is. High positive values mean there is a strong rising trend, and low values signify a strong downward trend.	$EMA_{13}[(P_t - P_{t-1}) * Vol]$
volume_vpt	Volume-price trend (VPT) is based on a running cumulative volume that adds or subtracts a multiple of the percentage change in share price trend and current volume, depending upon the investment's upward or downward movements. This identifies buying and selling pressure in the stock market. A rising VPT indicates buying pressure, while a falling VPT indicates selling pressure.	$VPT_{t-1} + [(P_t - P_{t-1}) * Vol]$
trend_adx_pos	These refer to the Average Directional Movement Index (ADX). The Plus Directional Indicator (+DI) and Minus Directional Indicator (-DI) are derived from smoothed averages of these differences, and measure trend direction over time. These two indicators are often referred to collectively as the Directional Movement Indicator (DMI). The Average Directional Index (ADX) is in turn derived from the smoothed averages of the difference between +DI and -DI and measures the strength of the trend (regardless of direction) over time. Using these three indicators together, traders can determine both the direction and strength of the trend.	$+DI : 100 * (\frac{+DM}{ATR})$
trend_adx_neg		$-DI : 100 * (\frac{-DM}{ATR})$
trend_adx		$100 * (\frac{Abs(+DI - -DI)}{Abs(+DI + -DI)})$

5.3 PREDICTIVE MODEL PREPARATION

The financial data and variables for the last five years were then merged with the sentiment data by joining on the dates, as to get the data on a daily to predict the values of next day's price trend. The data was then encoded to fit our modeling process. We created a dependent variable 'Trend' which was encoded as 1, if the day's closing price was greater than previous day's

MA: *Moving average*; **MSD**: *Moving Standard Deviation*; **m, n**: *Holding period (in days)*; **EMA**: *Exponential Moving average*; **Vol**: *Traded Volume*
P: *Adjusted Closing Price*; **+DM**: *High_t - High_t-1*; **-DM**: *Low_t – Low_t-1*; **ATR**: *Average True Range*; **Abs(x):** *Absolute value*

closing price, and 0 if day's closing prices is lower than the previous day's closing price. Since the objective of our project is to assess the impact of Twitter sentiments on predicting the trend of next day's closing prices, we feel that this encoding will help us achieve that better.
The sentiment scores were also binary encoded after categorizing them into three nominal categorical variables, i.e., Positive, Neutral, and Negative. For Vader sentiments, the output of the sentiments range from -1 to 1 as a continuous numeric. So, if the daily average sentiment score was greater than +0.05, then it was categorized as 'Positive' variable and was encoded as '1', else it was '0'. Similarly, when the score was less than -0.05, then it was categorized as 'Negative' sentiment and the variable was encoded as '1', else for any other value, 'Neutral' sentiment and the variable was encoded as '1' else '0'.

The BERT module outputs values of the sentiments ranging from 1 to 5. So, for Bert sentiments, the sentiment values of 1 and 2 were categorized as 'Negative', 3 as 'Neutral', and 4 and 5 as 'Positive'. These 3 categorical variables, like the VADER sentiments were, then binary encoded for feeding them into our classifiers.

Overall, we were able to create a total of 25 relevant variables for our machine learning models.

5.4 CLASSIFICATION MODELS

We used tree-based ensemble models, namely Gradient Boosting Machine Classifier model (GBM) and Random Forest Classifier model (RF) to check if the inclusion of twitter sentiment alongside financial indicators improves the performance of our prediction across desired model metrics.

Our initial approach was to use the SVM methodology as mentioned in the "Bollen et al., 2011 [2]", but as per the new research and online articles, we found that decision trees-based classifiers work better on this type of classification problem.

5.4.1 GRADIENT BOOSTING MACHINE CLASSIFER

Gradient Boosting Machine (GBM) is a machine learning technique for classification and regression problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalize them by allowing optimization of an arbitrary differentiable loss function.

We ran three models using this methodology. The first model used VADER variables created, the second model consisted of BERT variables, and the last model was run only using the financial data. We passed all of the features in the model and tuned the hyperparameters to the best fit on the training data. We then extrapolated the model to the test data and the results are presented in the evaluation section.

5.4.2 RANDOM FOREST CLASSIFIER

Random Forest is an ensemble learning method for classification and regression problems. It creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision

Fig.2 Historical Financial Data for Tesla's stock

trees to decide the final class of the test object. This approach improves the stability and accuracy of the model since it reduces overfitting by averaging the results. Random Forest algorithm can be used for both classification and regression problems.

Similar to the gradient booster method, we created three models with the same feature specification and generated the best fit models and extrapolated to the test data the results are presented in the evaluation section.

6. EVALUATION

The model evaluation was based on various classification metrics like precision score, accuracy score, recall score, F1-score, and auc-roc score as we are dealing with binary classification. We chose F1-score as our basis for hyperparameter tuning using GridSearchCV with 5-fold cross validation, as the F1-score is a measure of a model's accuracy that considers both precision and recall. It is commonly used in machine learning models to evaluate the performance of binary classification problems. The F1-score is calculated as the harmonic mean of precision and recall.

Precision is the number of true positives divided by the number of true positives plus false positives. It is a measure of how many of the positive predictions made by the model are actually correct.

Recall is the number of true positives divided by the number of true positives plus false negatives. It is a measure of how many of the actual positive instances were correctly predicted by the model.

The F1-score is a balance between precision and recall. A high precision means that the model has a low false positive rate, while a high recall means that the model has a low false negative rate.

A high F1-score indicates that the model has a good balance of precision and recall. The F1-score ranges from 0 to 1, where 1 is the best possible score.

However, as accuracy is not always a good indicator of performance, it is often more informative to use the F1-score as a metric for model performance instead of accuracy. Fortunately, our dataset is not imbalanced, so using any of F1-score or accuracy along with auc-roc would give consistent evaluation of the model's performance. Additionally, when the objective is to balance the trade-off between precision and recall, F1-score is a better metric compared to accuracy as it takes both into account, so we feel F1-score is a better metric for parameter tuning.

In our case, "GBM with BERT" model performed the best with F1-score of 77.23% and auc-roc score of 80.57% on test data, for classifying the next day's stock trend as binary label. The model evaluation metrics during training and testing have been mentioned below along with the features importance for each of the 6 models.

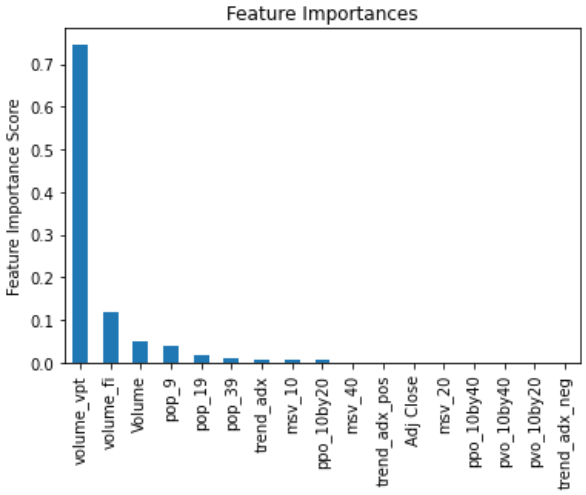

Fig.3.1 Feature Importance: GBM – Financial Data only

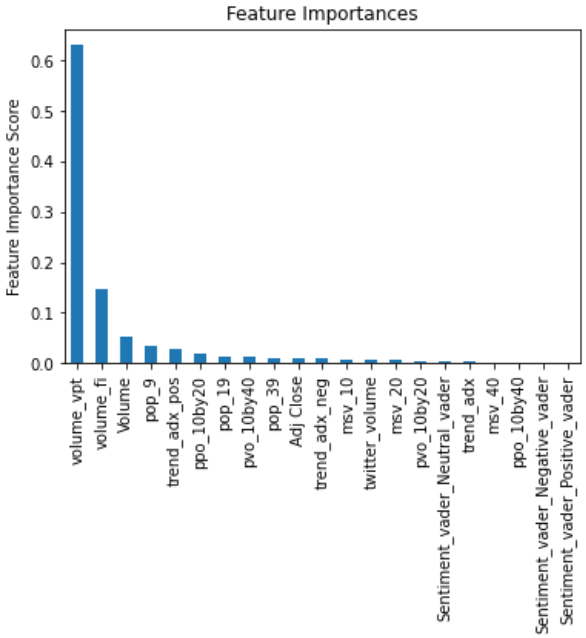

Fig.3.2 Feature Importance: GBM – With VADER

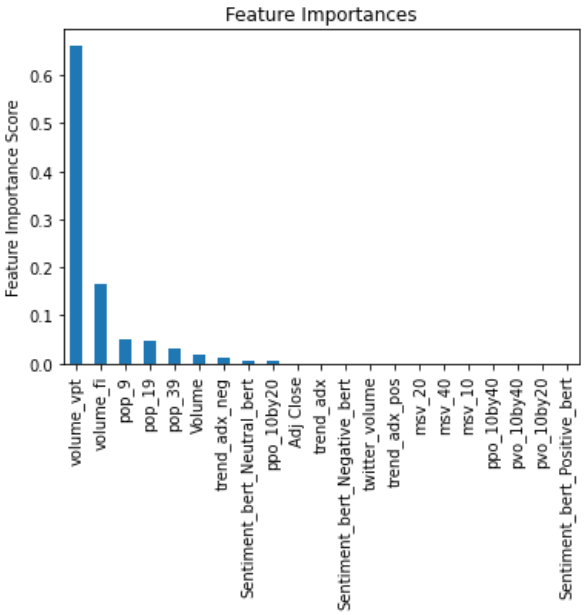

Fig.3.3 Feature Importance: GBM – With BERT

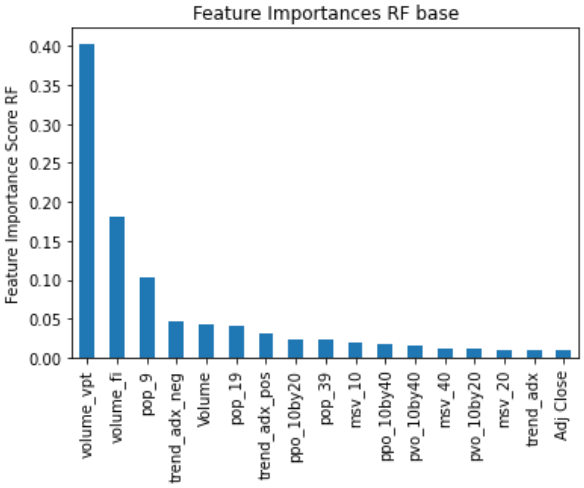

Fig.3.4 Feature Importance: RF – Financial Data only

As we can see from the above graphs, that technical indicators like "volume_vpt", "volume_fi", "pop_9" and "trend_adx" have importance covering roughly 60-80% of the information, even when used with sentiment data. The best model "GBM with BERT" has a sentiment variable as the 8th most important with importance values as low as 3%.

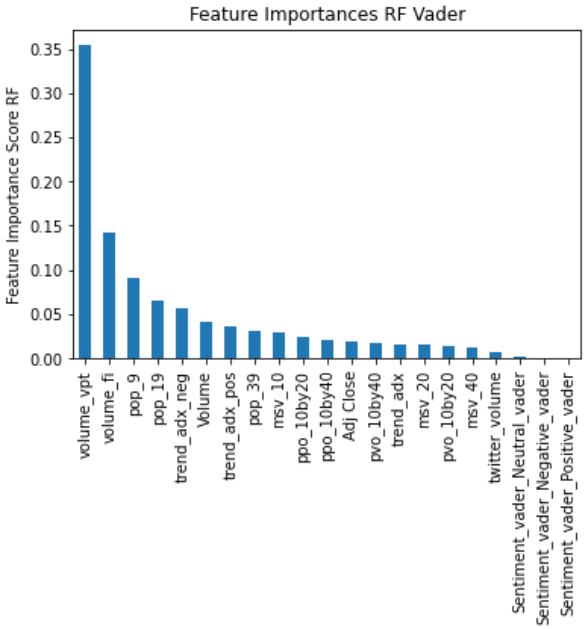

Fig.3.5 Feature Importance: RF – With VADER

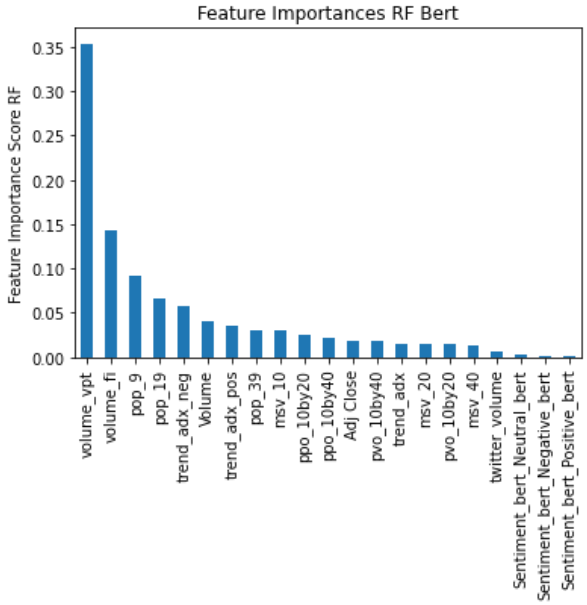

Fig.3.6 Feature Importance: RF – With BERT

GRADIENT BOOSTING MACHINE CLASSIFER	MODELS					
	VADER		BERT		FINANCIAL DATA ONLY	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
Accuracy	73.52%	75.40%	73.52%	**75.66%**	73.18%	73.81%
Precision	70.89%	77.04%	71.13%	**76.10%**	70.49%	77.47%
Recall	82.60%	75.88%	81.94%	**78.39%**	82.60%	70.85%
F1-score	**76.30%**	**76.46%**	**76.15%**	**77.23%**	**76.06%**	**74.02%**
AUC Score	79.28%	79.17%	78.59%	**80.57%**	78.95%	79.62%

RANDOM FOREST CLASSIFER	MODELS					
	VADER		BERT		FINANCIAL DATA ONLY	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
Accuracy	74.55%	71.69%	74.55%	71.69%	74.32%	72.22%
Precision	73.09%	76.74%	73.09%	76.74%	72.89%	77.33%
Recall	80.18%	66.33%	80.18%	66.33%	79.96%	66.83%
F1-score	**76.47%**	**71.16%**	**76.47%**	**71.16%**	**76.26%**	**71.70%**
AUC Score	80.83%	77.91%	80.83%	77.91%	80.85%	78.67%

7. CONCLUSIONS & IMPROVEMENTS

The study aimed to analyze the effect of real-time public opinion on stock prices by conducting sentiment analysis on Twitter opinions and its correlation with financial data. The results of the study show that the inclusion of sentiment as a feature in the models led to an approximate 2-3% improvement in the F1-score when compared to the models that were solely based on financial indicators. This suggests that incorporating public opinion, as represented by sentiment on Twitter, can provide valuable insights in predicting stock prices. It's worth mentioning that the data source used in this study had some limitations, such as only scraping the tweets from a certain area, quality/quantity issues, and the limitation of only scraping tweets from the last 5 years. Thus, we conclude that with improved tweet data quality and quantity, or any other source of public sentiment such as news articles along with tweets, in addition to the financial indicators would be very good predictors for stock market prices. However, it's important to note that this is just one of many factors that can influence stock prices and more research is needed to fully understand the relationship between public opinion and stock prices. Additionally, the study also highlights the importance of using appropriate methods for data collection and analysis when working with social media data.

8. REFERENCES

[1] V. S. Pagolu, K. N. Reddy, G. Panda and B. Majhi, "Sentiment analysis of Twitter data for predicting stock market movements," *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, 2016, pp. 1345-1350, DOI: 10.1109/SCOPES.2016.7955659.

[2] Bollen J., Mao H., Zeng X.: Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8 (2011). DOI: https://doi.org/10.1016/j.jocs.2010.12.007

[3] Khan, W., Ghazanfar, M.A., Azam, M.A. *et al.* Stock market prediction using machine learning classifiers and social media, news. *J Ambient Intell Human Comput* **13**, 3433–3456 (2022). DOI: https://doi.org/10.1007/s12652-020-01839-w

[4] Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, Arun Kumar, "Stock Closing Price Prediction using Machine Learning Techniques", *Procedia Computer Science*, *Volume 167, 2020, Pages 599-606,* ISSN 1877-0509, DOI: https://doi.org/10.1016/j.procs.2020.03.326

[5] Zhifeng Dai, Xiaodi Dong, Jie Kang, Lianying Hong, Forecasting stock market returns: New technical indicators and two-step economic constraint method The North American Journal of Economics and Finance, Volume 53, 2020, 101216, ISSN 1062-9408, https://doi.org/10.1016/j.najef.2020.101216

[6] Mohapatra, S.; Mukherjee, R.; Roy, A.; Sengupta, A.; Puniyani, A. Can Ensemble Machine Learning Methods Predict Stock Returns for Indian Banks Using Technical Indicators? J. Risk Financial Manag. 2022, 15, 350. https://doi.org/10.3390/jrfm15080350

[7] Christopher J. Neely, David E. Rapach, Jun Tu, Guofu Zhou (2014) Forecasting the Equity Risk Premium: The Role of Technical Indicators. Management Science 60(7):1772-1791. https://doi.org/10.1287/mnsc.2013.1838

[8] https://github.com/google-research/bert

[9] https://huggingface.co/bert-base-multilingual-cased

[10] https://pypi.org/project/twint/

[11] https://pypi.org/project/yfinance/

[12] https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664

[13] https://www.red-gate.com/simple-talk/development/data-science-development/sentiment-analysis-python/

[14] R. K. Bakshi, N. Kaur, R. Kaur and G. Kaur, "Opinion mining and sentiment analysis," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2016, pp. 452-455.

URL:https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7724305&isnumber=7724213

[15] Smith, S., O'Hare, A. Comparing traditional news and social media with stock price movements; which comes first, the news or the price change? J Big Data 9, 47 (2022). https://doi.org/10.1186/s40537-022-00591-6

[16] Yuan, S., Wu, X. & Xiang, Y. Incorporating pre-training in long short-term memory networks for tweet classification. Soc. Netw. Anal. Min. 8, 52 (2018). https://doi.org/10.1007/s13278-018-0530-1

[17] Thien Hai Nguyen, Kiyoaki Shirai, Julien Velcin, Sentiment analysis on social media for stock movement prediction, Expert Systems with Applications, Volume 42, Issue 24, 2015, Pages 9603-9611, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2015.07.052

[18] Wang, Y. Stock market forecasting with financial micro-blog based on sentiment and time series analysis. J. Shanghai Jiaotong Univ. (Sci.) 22, 173–179 (2017). https://doi.org/10.1007/s12204-017-1818-4

[19] Weiwei Jiang, Applications of deep learning in stock market prediction: Recent progress, Expert Systems with Applications, Volume 184, 2021, 115537, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2021.115537.

[20] Othan D, Kilimci ZH, Uysal M. Financial sentiment analysis for predicting direction of stocks using bidirectional encoder representations from transformers (BERT) and deep learning models. InProc. Int. Conf. Innov. Intell. Technol. 2019 Dec (Vol. 2019, pp. 30-35). https://doi.org/10.17758/URUAE8.UL12191013