Analyzing Twitter Sentiment for Stock Market Prediction

Narendra Kumar Vasa*[1] and Dr. Abhinesh kaushik[1]

[1]Department of Information Technology, Indian Institute of Information Techology, Lucknow

***Abstract*—This research paper investigates the relationship between public sentiment and stock price movements by leveraging data from Twitter. The importance of this work stems from the growing influence of social media on public opinion and its potential impact on financial markets. Traditional stock price prediction is fraught with challenges due to the multitude of factors involved, including economic conditions, political events, and other external influences. This complexity highlights a critical research gap: the need to understand how specific factors, such as social media sentiment, contribute to market trends.**

To address this gap, we applied sentiment analysis and machine learning techniques to Twitter data, capturing the public's emotional response to news and events related to various companies. Our methodology involved collecting relevant tweets, preprocessing the data to remove noise, and using sentiment analysis tools to classify the sentiment of each tweet. We then employed machine learning models to analyze the relationship between these sentiment scores and subsequent stock price movements.

Our findings indicate a significant correlation between public sentiment on Twitter and stock price trends. Positive sentiment generally precedes upward stock movements, while negative sentiment often predicts declines. These results suggest that social media sentiment can serve as a valuable predictor of stock performance.

The achievement of this research lies in its demonstration of the practical applicability of sentiment analysis and machine learning in financial forecasting. By uncovering the predictive power of public sentiment, this study contributes to a deeper understanding of the factors influencing stock prices and opens new avenues for enhancing market prediction models.

I. Introduction

Predicting stock prices is a complex and multifaceted challenge due to the myriad of factors that influence market movements, including economic conditions, political events, and various environmental influences. Traditional models often rely heavily on historical price data and economic indicators to forecast future stock trends. However, the rise of social media platforms has introduced a new and valuable data source for understanding and predicting market behavior. Social media, particularly Twitter, provides real-time insights into public sentiment and opinion, which can significantly impact investor behavior and, consequently, stock prices.

Our approach is unique in several ways. Unlike previous studies that often use traditional financial indicators or general market sentiment, we specifically target tweets about a specific company. This focused analysis, combined with a comprehensive dataset from January 1, 2016, to August 31, 2019 of Apple Inc. from AAPL, allows for a more precise examination of the impact of public sentiment on stock prices through different prediction machine learning models. Additionally, we employ both classification and regression techniques, using advanced machine learning algorithms like Random Forest, Gradient Boosting, LSTM RNN, and XGBoost to enhance predictive accuracy.

The classification approach categorizes tweets into positive, negative, and neutral sentiments, while the regression approach predicts future stock prices based on these sentiments and historical data. Our results indicate that while sentiment analysis alone may not fully predict stock movements, combining it with historical price data significantly improves the accuracy of predictive models. The XGBoost Regressor, in particular, showed promising results with an RMSE of 0.0477 and an R-squared value of 95.92%.

This project contributes to the field of behavioral economics by highlighting the influence of public emotions and opinions on market behavior. It underscores the potential of integrating social media sentiment with traditional financial models to enhance stock market predictions, paving the way for more comprehensive and effective predictive strategies in the future.

II. Literature Review

The existing literature, articles and journals show the influence of public sentiment on stock market movements has been extensively studied. Bollen et al. (2011) demonstrated that public mood states, as inferred from Twitter feeds[1], could predict the Dow Jones Industrial Average (DJIA) with considerable accuracy. Similarly, Zhang et al. (2011) found that sentiment extracted from Twitter could enhance the prediction accuracy of stock market trends [2]. These studies highlight the potential of social media sentiment as a leading indicator of market movements.

Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) technique used to determine the sentiment expressed in text. Various methods have been employed to perform sentiment analysis on Twitter data, ranging from simple lexicon-based approaches to sophisticated machine learning models. Lexicon-based approaches, such as those used by Pang and Lee (2008), rely on predefined lists of positive and negative words to classify sentiment. However, these methods can be limited by their inability to capture context and nuances in language [3]. More advanced machine learning techniques, such as those used by Pak and Paroubek (2010), and Giachanou and Crestani (2016), have

shown improved performance by leveraging deep learning models, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [4] [5].

In financial forecasting, machine learning algorithms have been widely applied due to their ability to model complex, non-linear relationships. Random forests, an ensemble learning method introduced by Breiman (2001), have been shown to perform well in financial forecasting due to their robustness and ability to handle high-dimensional data [6]. Hsu et al. (2016) demonstrated the effectiveness of random forests in predicting stock price movements based on technical indicators[7].

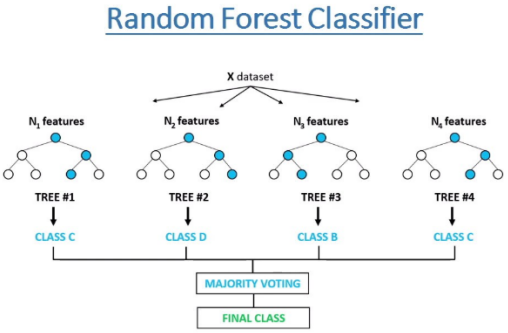

fig-1: Random Forest Classifier

Similarly, gradient boosting, as implemented in the XGBoost algorithm by Chen and Guestrin (2016), has become popular in financial applications due to its high predictive accuracy and computational efficiency [8]. Long short-term memory (LSTM) networks, a type of RNN, have also gained traction in time series forecasting, with Fischer and Krauss (2018) showing that LSTM networks can outperform traditional machine learning models in predicting stock returns by capturing temporal dependencies in sequential data [9].

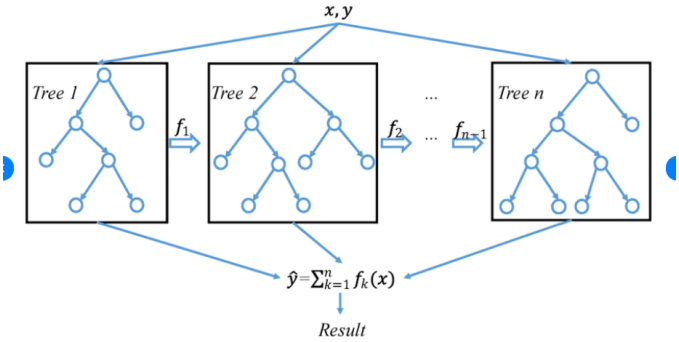

fig-2: XGboosting regressor

Combining sentiment analysis with machine learning for stock prediction has yielded promising results. Oliveira et al. (2017) integrated Twitter sentiment with financial data to predict stock returns using an SVM classifier, demonstrating improved accuracy over models that did not include sentiment data [10]. Similarly, Xu and Keelj (2018) employed LSTM networks to integrate Twitter sentiment and historical stock prices, achieving superior performance in predicting stock price movements [11].

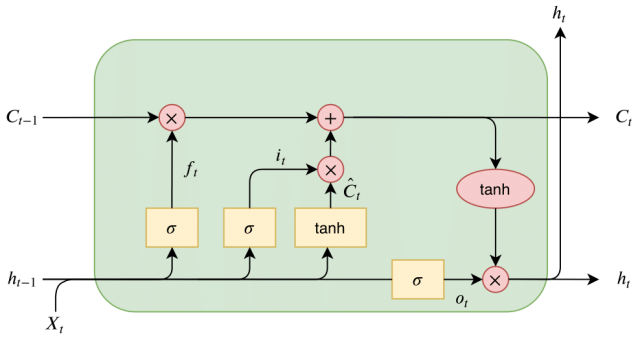

fig-3: Long Short-term Memory(LSTM) cell

Despite the potential of using Twitter sentiment for stock prediction, several challenges remain. The high noise-to-signal ratio in social media data can affect the reliability of sentiment analysis. Additionally, the dynamic nature of language on social media, including slang and abbreviations, poses challenges for NLP models. Data availability and quality are also significant concerns. Twitter's API limitations and the potential biases in publicly available datasets can restrict the comprehensiveness of analyses. Furthermore, integrating sentiment data with traditional financial indicators requires careful consideration of the temporal alignment and the potential lag between sentiment expression and market reaction.

In conclusion, the integration of social media sentiment, particularly from Twitter, with machine learning techniques offers a novel approach to stock market prediction. While traditional financial models rely on historical data and economic indicators, the addition of real-time public sentiment can enhance predictive accuracy. However, addressing the challenges of data quality, noise, and model complexity remains crucial for advancing this field. Future research should focus on improving sentiment analysis techniques, exploring new data sources, and refining machine learning models to fully leverage the potential of social media in financial forecasting.

III. Research Methods

Understanding the intricate relationship between social media sentiment and stock market movements necessitates a multifaceted approach that integrates advanced machine learning techniques with comprehensive data preprocessing and analysis. In this study, we propose a novel methodology that combines sentiment analysis of Twitter data with regression and classification models to predict stock price trends. Our approach entails meticulous data collection, preprocessing, and feature engineering, followed by the training and evaluation of various machine learning algorithms to determine their effectiveness in forecasting stock market behavior.

1) Data Processing:

We begin our methodology by collecting Twitter

data related to Apple inc from Kaggle, covering the period from January 2016 to August 2019. The dataset included tweets, timestamps, and sentiment scores.Simultaneously, we gather historical stock data from Yahoo Finance for the same timeframe, which includes daily adjusted closing prices. Next, we synchronize the datasets to include only trading days, ensuring that the Twitter activity aligns with stock market operations. Following this, we conduct feature engineering by calculating the polarity scores of tweets using sentiment analysis techniques, with the adjusted closing prices serving as the target variable for our predictions.

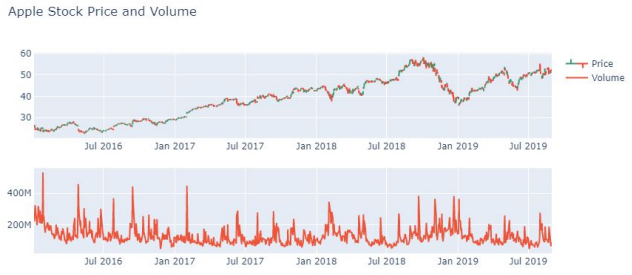

fig-4: Apple Inc stock Price and Volume

Date	Open	High	Low	Close	Adj Close	Volume	ts_polarity	twitter_volume
2016-01-04	25.65	26.34	25.50	26.34	24.44	270597600	0.070389	1133
2016-01-05	26.44	26.46	25.60	25.68	23.83	223164000	0.133635	1430
2016-01-06	25.14	25.59	24.97	25.17	23.36	273829600	0.072042	1949
2016-01-07	24.67	25.03	24.11	24.11	22.38	324377600	0.074369	2289
2016-01-08	24.64	24.78	24.19	24.24	22.50	283192000	0.051595	2235
2016-01-11	24.74	24.76	24.33	24.63	22.86	198957600	0.019443	1222
2016-01-12	25.14	25.17	24.71	24.99	23.19	196616800	0.121364	1293
2016-01-13	25.08	25.30	24.33	24.35	22.60	249758400	0.107714	1292
2016-01-14	24.49	25.12	23.93	24.88	23.09	252680400	0.039248	1264
2016-01-15	24.05	24.43	23.84	24.28	22.54	319335600	0.093784	1336
2016-01-19	24.60	24.66	23.88	24.17	22.43	212350800	0.120585	960

fig-5: Comprehensive Overview of AAPL stock data

2) Feature Engineering:
Feature engineering is a crucial step in our methodology that involves transforming raw data into meaningful features that can be effectively used by machine learning models. In this study, we focus on extracting and engineering features from Twitter data and stock price data to improve the accuracy of our predictive models.
   a) Sentiment Analysis of Tweets:
   We calculate the polarity scores of tweets using sentiment analysis techniques. The sentiment analysis assigns a sentiment score to each tweet, indicating whether the sentiment is positive, negative, or neutral.Sentiment scores are typically calculated using a sentiment analysis library, such as TextBlob or VADER. For example, with TextBlob, the polarity score ranges from -1 (very negative) to +1 (very positive).

$$Polarity\ Score = sentiment(tweet)$$

   b) Aggregating Daily Sentiment Scores:
   We aggregate the polarity scores of tweets for each trading day to create a daily sentiment score. This provides a summary of the overall sentiment on a particular day. Daily sentiment score is calculated as

$$DailySentimentScore = \sum PolarityScore/n$$

   c) Combining Twitter Data with Stock Data:
   We merge the daily sentiment scores and tweet volumes with the historical stock data. This involves aligning the sentiment features with the corresponding trading days. The primary stock feature we use is the adjusted closing price, which serves as the target variable for prediction. We ensure that our final dataset only includes days when the stock market was open, approximately 252 trading days per year.
   d) Creating Lagged Features:
   To capture the temporal dependencies in the data, we create lagged features from the adjusted closing prices. These features represent the stock prices from previous days and help the models understand trends over time.

$$LaggedPrice_{t-n} = AdjustedClosingPrice_{t-n}$$

   e) Binary Trend Labels (for Classification Approach):
   For the classification approach, we convert the adjusted closing prices into binary labels representing the stock price trend (up or down).

$$TrendLabel_t = \begin{cases} 1 \\ if\, AdjclosePrice_t \leq AdjclosePrice_{t-1} \\ 0 \\ if\, AdjClosePrice_t \geq AdjClosePrice_{t-1} \end{cases}$$

By performing these feature engineering steps, we transform raw Twitter data and stock price data into a structured dataset that captures both sentiment and historical price information. This dataset serves as the foundation for training our machine learning models to predict future stock price movements.

3) Model Training and Evaluation:
In this study, we employed both classification and regression approaches to analyze and predict stock price movements based on Twitter sentiment data and historical stock prices. Below is a detailed description of the model training and evaluation process, including the relevant formulas.

- classification approuch:
  The classification approach aimed to predict whether the stock price would go up or down based on the sentiment of tweets.
  a) Random Forest Classifier:
  The Random Forest Classifier is an ensemble learning method that constructs multiple decision trees during training and outputs the class

that is the mode of the classes predicted by individual trees.The model is trained using features derived from Twitter sentiment (daily sentiment score and tweet volume) and the binary trend label (0 for price decrease, 1 for price increase). **Formula:** The prediction for each tree T_i:

$$h_i(x) = T_i(x)$$

The final prediction $H(x)$:

$$H(x) = \text{mode}(h_1(x), h_2(x), \ldots, h_n(x))$$

b) Gradient Bossting Classifier:
Gradient Boosting Classifier builds an ensemble of trees sequentially, where each new tree focuses on correcting errors made by the previous ones. The model uses the features and binary trend labels, adjusting weights iteratively to minimize the classification error. **Formula:** The prediction for each tree T_i:

$$h_i(x) = T_i(x)$$

The final prediction $H(x)$:

$$H(x) = \sum_{i=1}^{n} \gamma_i h_i(x)$$

Where γ_i are the weights adjusted during training.

- Regression Approach:
The regression approach aimed to predict the actual future stock price based on sentiment and historical prices.

a) Random Forest Regressor:
The Random Forest Regressor uses an ensemble of decision trees to predict a continuous target variable, averaging the predictions of individual trees. The model is trained using historical adjusted closing prices and Twitter sentiment features.
Formula: The prediction for each tree T_i:

$$h_i(x) = T_i(x)$$

The final prediction $H(x)$:

$$H(x) = \frac{1}{n} \sum_{i=1}^{n} h_i(x)$$

b) LSTM RNN (Long Short-Term Memory Recurrent Neural Network:
LSTM networks are specialized for capturing long-term dependencies in time-series data. They consist of memory cells that maintain information for long periods. The model is trained on sequences of historical adjusted closing prices and sentiment features.
Formula: LSTM cell equations:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f)$$

$$g_t = \tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

$$h_t = o_t \odot \tanh(c_t)$$

c) XGBoost Regressor:
XGBoost is an optimized gradient boosting algorithm that builds sequential trees, where each tree aims to correct errors made by previous trees. The model is trained using the same features as the other regressors, focusing on minimizing the objective function. **Formula:** Objective function:

$$Obj(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Where l is the loss function, $\hat{y}_i$ is the predicted value, and Ω is the regularization term.

4) Evaluation Metrics:

a) Classification Metrics:

Accuracy: The proportion of correctly predicted instances among the total instances.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Precision, Recall, F1-score: Used to measure the performance on imbalanced datasets.

b) Regression Metrics
Root Mean Square Error (RMSE): Measures the average magnitude of the errors.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}$$

R-squared (R^2): Indicates the proportion of variance in the dependent variable that is predictable from the independent variables.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Where y_i is the actual value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of the actual values.

IV. The Results and Discussion

This section provides an overview of the accuracy rates achieved by the trained classifiers, with all calculations conducted using the Weka tool running on the Java Virtual Machine.

A. Sentiment Analyzer Results

In the preceding sections, we discussed the methodology employed to train the classifier utilized for sentiment analysis of tweets. The classifier, utilizing features such as Word2vec representations of human-annotated tweets, trained on the Random Forest algorithm with a 90% split for training and the remaining for testing, demonstrated an accuracy of 70.2%. Similarly, employing N-gram representations yielded a slightly higher accuracy of 70.5%. Despite the close results, the model trained with Word2vec representations was selected for classifying nonhuman annotated tweets due to its promising accuracy for large datasets and the sustainability in word meaning. Notably, studies have shown that sentiment analyzers above 70% accuracy are considered highly accurate in most cases, aligning with our obtained results. Table I outlines the results of sentiment classification, encompassing accuracy, precision, F-measure, and recall, when trained with different machine learning algorithms. ROC curves are further plotted for detailed analysis.

B. Stock Price and Sentiment Correlation Results

We introduced a classifier trained with aggregate sentiment values for a 3-day period as features and the increase/decrease in stock price represented by 1/0 as the output. The total data was split into two parts, with 80% used for training the model and the remainder for testing operations. The classifier yielded an accuracy value of 69.01% when trained using the Logistic Regression algorithm, with varying accuracy rates observed with different training sets. Notably, training the model with Lib SVM using 90% of the data resulted in an accuracy of 71.82%. These results showcase a significant correlation between stock market movements and the sentiments expressed by the public on Twitter. Moreover, the trend indicates that with increasing dataset size, the models perform better, suggesting potential enhancements through the incorporation of more data in future endeavors.

TABLE I
RESULTS OF SENTIMENT CLASSIFICATION

Algorithm	Accuracy
Random Forest (Word2vec)	70.2%
Random Forest (N-gram)	70.5%

C. Results of Classification and regression approuches

Classification Report

	precision	recall	f1-score	support
0	0.48	0.29	0.36	127
1	0.55	0.73	0.63	150
accuracy			0.53	277
macro avg	0.52	0.51	0.50	277
weighted avg	0.52	0.53	0.51	277

fig-6: Random Forest Classifier Report

Classification Report

	precision	recall	f1-score	support
0	0.53	0.33	0.41	127
1	0.57	0.75	0.65	150
accuracy			0.56	277
macro avg	0.55	0.54	0.53	277
weighted avg	0.55	0.56	0.54	277

fig-7: Gradient Boosting Classification report

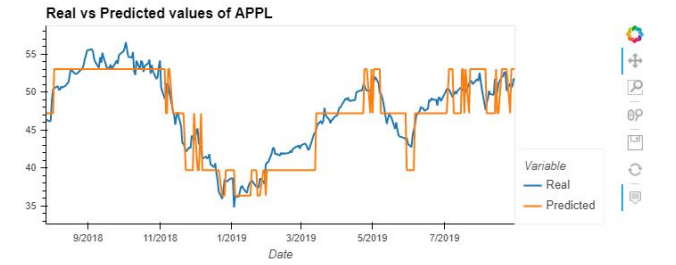

fig-8: Random Forest Regressor real vs Prediction plot

fig-9: LSTM RNN predicted vs real values plot

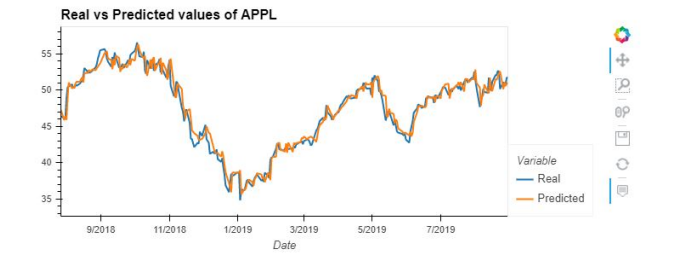

fig-10: XGboosting predicted vs real values plot

V. Conclusion

The culmination of our research project brings forth intriguing insights into the realm of stock market prediction through the lens of social media sentiment analysis. Our endeavor aimed to harness the wealth of real-time information embedded within Twitter data and marry it with robust machine learning methodologies to forecast stock market movements. Through meticulous data integration and preprocessing, we constructed a comprehensive dataset spanning from January 2016 to August 2019, capturing both Twitter sentiment and historical stock prices for our selected company, here it is Apple Inc.

While our exploration encompassed both classification and regression approaches, it became evident that the regression models yielded more promising results. Despite the classification models demonstrating limited efficacy, with accuracy scores hovering around 50-55%, the regression models, particularly the XGBoost Regressor, showcased notable performance metrics. The XGBoost Regressor, with its impressive RMSE of 0.0477 and R-squared value of 95.92%, emerged as the frontrunner in leveraging historical prices and sentiment data to predict stock price movements accurately.

This study bears profound implications for the intersection of social media analytics and financial forecasting. While sentiment analysis alone may not suffice as a sole determinant of stock market behavior, its amalgamation with machine learning models has demonstrated significant potential in augmenting predictive accuracy. However, it is imperative to acknowledge the inherent limitations and caveats embedded within our research. The dataset's temporal and company-specific constraints, coupled with the nuances of sentiment analysis accuracy and the influence of external factors, underscore the need for further refinement and expansion in future studies.

In conclusion, our findings underscore the transformative power of integrating social media sentiment analysis with advanced machine learning techniques in unlocking predictive insights within the dynamic landscape of financial markets. As we traverse into the future, continued exploration and refinement in data analytics methodologies hold the promise of unraveling deeper layers of understanding and foresight in stock market dynamics, thereby empowering investors and stakeholders with enhanced decision-making capabilities.

Acknowledgment

I would like to express my deepest gratitude to my supervisor, Dr. Abhinesh Kaushik, for his expertise, understanding, and patience, which significantly enriched my graduate experience. I appreciate his extensive knowledge and skills in various areas of research design and programming, as well as his assistance in writing reports. His guidance was invaluable throughout my research and the writing of this thesis. I could not have asked for a better advisor and mentor for my studies.

References

[1] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2(1), no. 11, pp. 1–8, 2011.

[2] X. Zhang, H. Fuehres, and P. A. Gloor, "Predicting stock market indicators through twitter "i hope it is not as bad as i fear"," *Procedia-Social and Behavioral Sciences*, vol. 26, pp. 55–62, 2011.

[3] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2(1-2), pp. 1–135, 2008.

[4] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," *In LREc (Vol. 10, No. 2010)*, 2010.

[5] A. Giachanou and F. Crestani, "Like it or not: A survey of twitter sentiment analysis methods," *ACM Computing Surveys (CSUR), 49(2), 1-41.*, 2016.

[6] L. Breiman, "random forests," *Machine Learning, 45(1), 5-32*, 2001.

[7] S.-H. Hsu, S. Lessmann, M.-C. Sung, T.-S. Ma, and J. E. Johnson, "Bridging the divide in financial market forecasting: machine learners vs. financial economists," *Expert Systems with Applications, 61, 215-234*, 2016.

[8] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794)*, 2016.

[9] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *European Journal of Operational Research, 270(2), 654-669*, 2018.

[10] N. Oliveira, P. Cortez, and N. Areal, "The impact of microblogging data for stock market prediction: Using twitter to predict returns, volatility, trading volume and survey sentiment indices," *Expert Systems with Applications, 73, 125-144*, 2017.

[11] Y. Xu and J. Keelj, "Stock price prediction using deep learning models." *arXiv preprint arXiv:1810.09936.*, 2018.

Sign here

Narendra Kumar Vasa

Dr.Abhinesh Kaushik