ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Use of Heterogeneous Data for Forecasting of Stock Market Movement Using Social Media and Financial News Headlines

Uphar Singh[1], Dhanush Vasa[2], Lekhana Reddy Mitta[3], Kumar Saurabh[4], Ranjana Vyas[5] and O. P. Vyas[6]

ABSTRACT

The stock market holds a significant role in shaping a nation's economic landscape and impacting the prosperity of businesses. It goes hand in hand with the country's development, closely mirroring market behaviour. While global and societal dynamics exert their influence, technological advancements stand as the primary driving force behind market trends. The fusion of social media and financial news headlines provides a valuable lens into the collective sentiment of the public. This study aims to craft an LSTM model for predicting stock volatility movements. This research delves into various facets of the stock market domain, introducing diverse inputs for comparison against the original market trends. These inputs comprise a rich tapestry of data, including information from social media networks and news headlines sourced from official publications. The goal is to gauge public sentiment as a potent factor for more precise predictions, diverging from the conventional transaction-based stock price forecasts. The outcomes have been promising. By leveraging public sentiment analysis, overall errors have been reduced, with a 43.86% drop in Mean Absolute Error (MAE) and an impressive 50.93% decrease in Root Mean Square Error (RMSE) compared to without sentimental analysis. These results signal a step forward in enhancing the accuracy of stock market predictions.

Article information:

Keywords: Stock Market, Social Media, Financial News Headlines, Long Short Term Memory Model, Twitter and Reddit data

Article history:
Received: June 22, 2023
Revised: August 24, 2023
Accepted: October 26, 2023
Published: November 25, 2023 (Online)

DOI: 10.37936/ecti-cit.2023174.253579

1. INTRODUCTION

The stock market [1] plays a crucial role in determining the Gross Domestic Product (GDP) of a country and the financial prosperity of numerous companies. According to the stipulation, a rising trend in the stock market is correlated with a commensurate rise in the country's GDP, indicating a positive trajectory in the nation's economic development. It is evident from historical data that the stock market lacks predictability about individual investments, resulting in the possibility of substantial financial gains, neutral returns, or substantial losses. Numerous factors influencing stock prices have contributed to these results, with historical transaction data playing a significant role. However, these historical transaction data are insufficient to provide an accurate forecast. Thus, factors such as daily financial news and social media have significantly influenced these assets' positive or negative market values. As a result of the arrival of the digital age, social media platforms and financial news outlets have generated an unprecedented amount of data and have become the dominant force. They have the potential to revolutionise the way market dynamics is understood. For accurate stock market forecasting, it is necessary to consider each factor. Due to the high consequences of investing in the stock market, an exact and automated analytic system is required to process massive amounts of social networks and financial news.

Since these data differ in format, structure, content, and source, they are collectively called "heterogeneous data". Combining and analysing such disparate data can result in a more comprehensive and in-depth understanding of market-moving factors. Differences in formats, content, and sources make heterogeneous data analysis challenging. Data integration and custom preprocessing techniques are vital in facilitating exhaustive research. When these issues are resolved, it will be possible to gain a deeper

[1,2,3,4,5,6] The authors are with Indian Institute of Information Technology - Allahabad, Prayagraj, India, E-mail: pse2017003@iiita.ac.in, vasadhanush99@gmail.com, lekhanamitta13@gmail.com, blup0212@gmail.com, ranjana@iiita.ac.in and opvyas@iiita.ac.in

understanding of how the market operates and how different categories of data interact.

Artificial intelligence (AI) is a promising remedy to all these problems. Deep learning models, a subset of AI, are an excellent means of addressing the abovementioned issues. It has demonstrated exceptional skills in analysing heterogeneous data due to its extraordinary ability to identify complex patterns independently across different data types [1]–[6]. This method facilitates comprehensive comprehension and improves the accuracy of future predictions. The long Short-Term Memory (LSTM) model has garnered widespread recognition as a leading deep-learning model for stock price forecasting [1]. Capturing and analysing the long-term relationships inherent in time series data is the most critical factor in their effectiveness. This skill is essential for fully comprehending the stock market patterns. Moreover, it is important to note that LSTM networks have been developed to address the prevalent vanishing gradient problems commonly encountered in deep learning models. Using LSTM networks, which can retain extensive contextual information from limited temporal data, enables the development of a robust framework for time series prediction.

This research will employ the LSTM model, a supervised deep learning technique, to develop a well-structured and well- considered model. Most available data will be used to train the model, while the remainder will be used for testing and evaluation. The current investigation uses social media platforms and financial news headlines as primary sources to analyse public sentiment. The goal is to increase the precision of forecasts by integrating these pertinent components with historical transactional data instead of relying solely on fundamental historical data. Consequently, this phenomenon is expected to expedite the decision-making process for corporations and investors while increasing investor interest in the stock market, thereby contributing to the economic growth of the respective nations.

The remaining part of this paper is organised such that Section II, “Related Work”, investigates all the work related to the topic. The “Proposed Methodology” section III describes different dataset descriptions and the methodology used in this paper, Preprocessing methods, and Deep Learning Models. The “Performance Evaluation” section IV explains the Metrics used in this research. The “Results and Discussion” section V deals with a brief discussion of experimental findings. The “Conclusion” section VI discusses the summary of this work. Finally, the “Future Research” section VII discusses the future directions.

2. RELATED WORK

In this Digital era, entities considered here affect the stock values as social, psychological, political, and economic factors. In 2010, Bjoern Krollner, B. Vanstone, G. [7] presented a paper that overviews machine learning techniques, particularly artificial neural networks, for financial time series forecasting. Highlights the trend of enhancing existing ANN models and combining them with other technologies to improve forecast accuracy. However, it also points out the need for further research on how these techniques can be practically applied to enhance an investor’s risk-return tradeoff in real-world scenarios. Later on, Polamuri Subba Rao [8] analyses stock market prediction using data from Indian stock market websites using various models like ARIMA, Holt-Winters, AI, Hidden Markov Model, and RNN. Introduce a novel approach that combines multiple methods to improve predictive outcomes. However, it acknowledges limitations, such as declining prediction accuracy with noise variation and specific models being suitable for short-term predictions. Derbentsev et al. [9] discussed the effectiveness of various machine learning models for forecasting short-term daily quotes of financial time series data, with a particular focus on Nasdaq and S&P 500 indices. Stochastic Gradient Boosting Machine (SGBM) demonstrated strong performance, outperforming other models, while Random Forest (RF) served as a baseline. The study also highlights the potential for combining ensemble methods with powerful machine learning models and exploring deep learning approaches for feature selection and prediction in financial forecasting. Dingli and Fournier [10] explore using Convolutional Neural Networks (CNN) for forecasting economic time series using datasets like S&P500, DJIA, and NASDAQ-100.The study aimed to predict next month’s and next week’s price direction with 65% and 60% accuracy, respectively. While other techniques like Logistic Regression and Support Vector Machines slightly outperformed CNN, the paper emphasises the configurability of deep learning models for financial forecasting. To overcome the above limitation, Zhang et al. [11] present a novel approach for predicting and analysing economic time series using CEEMD (Complete Ensemble Empirical Mode Decomposition) and LSTM (Long Short-Term Memory) neural networks. The study applied this framework to six representative stock indices in different markets and found that it outperforms benchmark models regarding predictive accuracy, achieving a lower test error and higher directional symmetry. Trading simulations confirmed superior profitability and risk-adjusted performance. Liapis et al. [12] surveyed sentiment analysis in multivariate financial time series forecasting. The study involves testing 22 different input setups using data extracted from social networks over 16 other datasets under the schemes of 27 algorithms. The results show that the use of sentiment analysis improves the models when used for long-term predictions. Specifically, FinBERT and TextBlob performed best in terms of

Table 1: *Comparison of Related Work and Proposed Approach.*

Serial No.	Author Name	Proposed Work	Methodology Used	Advantages	Research Gap
1	Bjoern Krollner, B. Vanstone, G.	Machine Learning for Financial Time Series Forecasting	Artificial Neural Networks (ANN)	Enhancing forecast accuracy, Combining ANN models with other technologies	Need for practical application in real-world scenarios
2	Polamuri Subba Rao	Stock Market Prediction Using Various Models	ARIMA, Holt-Winters, AI, Hidden Markov Model, RNN	Improved predictive outcomes through model combination	Declining prediction accuracy with noise variation
3	Derbentsev et al.	Machine Learning Models for Daily Quotes Forecasting	Stochastic Gradient Boosting Machine (SGBM), Random Forest (RF), Ensemble Methods	Strong performance of SGBM, Potential of ensemble methods	Limited exploration of deep learning approaches
4	Dingli and Fournier	CNN for Forecasting Financial Time Series	Convolutional Neural Networks (CNN), Logistic Regression, Support Vector Machines	Configurability of deep learning models for financial forecasting	Slightly lower accuracy compared to other techniques
5	Zhang et al.	Predicting Financial Time Series using CEEMD and LSTM	Complete Ensemble Empirical Mode Decomposition (CEEMD), Long Short-Term Memory (LSTM)	Outperforms benchmark models in predictive accuracy, Superior profitability	Lack of benchmark models' performance
6	Liapis et.al.	Sentiment Analysis in Multivariate Financial Time Series Forecasting	Sentiment Analysis, Various Algorithms, LSTM	Improved long-term predictions with sentiment analysis	Specific focus on long-term predictions
7	Zulfadzli Drus and Haliyana Khalid	Social Media Sentiment Analysis	Lexicon-based SentiWordnet, TF-IDF, Naïve Bayes, SVM	Both methods obtain similar accuracy	Emphasis on text structure, temporal aspects, and data volume
8	Mohan et.al.	Stock Price Prediction based on News Sentiment Analysis	RNN-LSTM, RNN-pp	Effective prediction with RNN-LSTM	Limited performance with highly or less volatile market data
9	Gupta et al.	Sentiment Analysis with StockTwits Data	Logistic Regression, TF-IDF	High accuracy with hybrid model	Potential for improved results with an extensive dataset
10	**Proposed Work**	**Sentiment Analysis with StockTwits Data**	**LSTM**	**High accuracy with a hybrid model, Overall error was reduced by an average of 43.86 per cent for MAE and 50.93 per cent for RMSE**	**Potential for improved results with an extensive dataset**

accuracy. The study also compared the performance of different forecast algorithms, with LSTM variations showing a clear predominance. Zulfadzli Drus and Haliyana Khalid [13] comprehensively review social media sentiment analysis literature. Social media includes content communities, social networking sites, blogs, and microblogs. The study compares two approaches: lexicon-based SentiWordnet and TF-IDF and machine learning Naïve Bayes and SVM. Interestingly, both methods obtained similar accuracy. The paper emphasises text structure, temporal aspects, and data volume in sentiment analysis. It also emphasises choosing the best sentiment analysis method based on the data's unique characteristics. Mohan et al. [14] explore using various models for predicting stock prices based on news sentiment analysis. RNN-LSTM, with the RNN-pp variant, was the most effective combination. However, the model's performance is limited when faced with highly or less volatile market data. For stock price prediction, Gupta et al. [15] explore using sentiment analysis, particularly with StockTwits data. The hybrid model of logistic regression and TF-IDF demonstrated high accuracy. Still, the authors acknowledge the potential for even better results with a more extensive dataset for training and testing.

The literature review discusses financial forecasting and sentiment analysis in stock markets, but a significant gap exists. Despite extensive research on machine learning models, deep learning techniques, and sentiment analysis, there needs to be more studies that integrate these elements into a unified framework. This requires a holistic approach considering social, psychological, political, and economic factors in stock market predictions. This approach could provide a more accurate understanding of market dynamics and lead to more robust and interpretable forecasting models. Future research should develop comprehensive frameworks that bridge the gap between quantitative models and real-world market influences.

3. PROPOSED METHODOLOGY

When it comes to data collection and machine learning, this consists of the procedure of analysing, calculating, and precisely collecting information for research using the validated reader. Next, when discussing stock market forecasting, the critical step is data collection, which is mandatory and crucial for future analysis. It is also vital that data are extracted from more than one source, as it may consist of biased and misleading information.

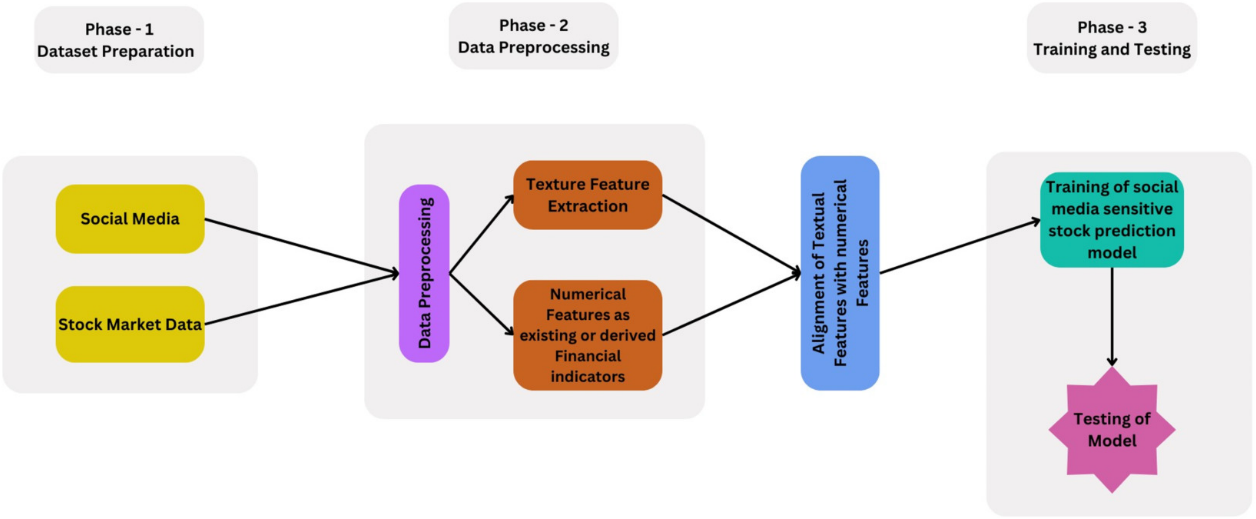
Phase - 1
Dataset Preparation
Phase - 2
Data Preprocessing
Phase - 3
Training and Testing
Social Media
Stock Market Data
Data Preprocessing
Texture Feature Extraction
Numerical Features as existing or derived Financial indicators
Alignment of Textual Features with numerical Features
Training of social media sensitive stock prediction model
Testing of Model

Fig.1: *Proposed Methodology Flow Chart.*

A. DataSet Description

The datasets used for this work are transaction historical data, financial news headlines data, and social media data

1) Historical Transaction Data: The BSE (Bombay Stock Exchange), one of India's major stock markets, is used for historical transaction data. For this part, two different datasets are used.

1) Client Categorywise Turnover- The acquisition of Client Categorywise Turnover data was facilitated using the official BSE website, bseindia.com, which the government administers. This website proved to be a dependable and precise source, furnishing us with a comprehensive dataset spanning from January 2011 to October 2021. The dataset had four distinct types:
   - Client: These individual investors trade in the stock market for personal investment purposes.
   - NRI (Non-Residential Indian): These are foreign entities or investors who invest in the stock market of a particular country.
   - Proprietary: These are trading firms or individuals who trade for their accounts rather than on behalf of clients.
   - DII (Domestic Institutional Investor): DII refers to Indian institutional investors who invest in the financial market in India.
2) Open, High, Low, and Close (OHLC) are four essential prices for analysing a stock's daily price movements or any financial instrument. Here's what each of these terms represents:
   - Open: The open price refers to the first trading price of a stock or any other financial instrument at the commencement of the trading day. The opening price refers to the first value at which a transaction occurs after the beginning of trading activities in the morning.
   - High: The high price represents the peak price at which the instrument or stock was exchanged throughout the trading day. It signifies the most elevated value point achieved throughout the trading session of that particular day.
   - Low: During the trading day, the low price denotes the lowest price at which the stock or instrument was transacted. It signifies the minimum price at- tained throughout the trading session of that particular day.
   - Closed: The closure price denotes the ultimate trading price of the instrument or stock after the trading day. It represents the final price at which a transaction is executed before the daily market closure.

Each category was divided into three subsections: BUY, SALES, and NET. The subsequent dataset consisted of the daily stock values for Open, High, Low, and Closed. This data was obtained from the official BSE website besindia.com, operated by the government. The website proved to be a dependable and precise source, allowing us to extract the dataset from January 2011 to October 2021 quickly. This study integrated both datasets to create a unified historical transaction dataset.

2) Sentiment Data: Sentiment data is information about individuals' emotions, opinions, or attitudes towards a subject, used in sentiment analysis to evaluate the public's perception of products, services, events, or financial markets. For this part, two different datasets are used.

1) Social Media: In social media analysis, data extraction involves retrieving information from two social media networks.
   - Twitter- For Twitter, the snscraper library in Python with a module dedicated to Twitter is used and collected every tweet related to BSE (Bombay Stock Exchange) from January 2011 to October 2021.
   - Reddit- For Reddit, an API from Reddit for app development is used with the help of the request

library in Python and the access tokens from Reddit API and collected posts related to BSE (Bombay Stock Exchange) from January 2011 to October 2021. Still, due to the limitations of the API, only six months of data from April 2021 to October 2021 was extracted.

2) Financial News Headlines For Financial News Head-lines, every headline from a website, such as Business Insider, Times of India, Bloomberg and Financial Express, was web-scarped and extracted for BSE's most recent and historical news.

B. Data Preparation

Analysing heterogeneous data requires data preparation to avoid bias and maintain consistency. It requires standardising and cleaning data to eliminate discrepancies, integrating data types into a single format, handling missing data to avoid bias, and ensuring data unit, scale, and representation compatibility. Quality assurance finds and fixes duplicates, outliers, and errors. Data preparation provides accurate insights and conclusions from reliable and meaningful analysis. It bridges data source gaps and improves analysis reliability and validity. Thus, heterogeneous data analysis requires data preparation to ensure accurate conclusions.

The data preparation for this research consists of two parts:

1) Dataset Preparation for Sentiment Analysis
2) Dataset Preparation for Transactional Data

1) Dataset Preparation for Sentiment Analysis: Sentiment analysis is increasingly used to analyse public sentiment towards news, product information, and other textual data. Its importance has increased due to the widespread use of social media platforms and prominent websites. Sentiment analysis is crucial in the stock market as it helps investors understand public sentiment and investor emotions, providing valuable in-sights into market trends and potential price fluctuations. This understanding helps investors make informed decisions and manage risks, as positive or negative sentiment can influence stock prices and overall market behaviour. Given the inherent volatility of the general public's perception of stocks, investors often rely on sentiment research as a valuable tool to inform the formulation of their long-term investment strategy.

The study focused on preparing social media posts and financial news headlines to standardise and clean data, manage large volumes, and enable practical sentiment analysis. To analyse public sentiment, duplicate posts or news headlines were removed from the same day. All social media posts and financial news headlines were concatenated into one column according to the dates. A new textual data set was created with an index as a date. Python's regex and replace commands were used to remove non-alphabet and space characters, and null or empty text was replaced with neutral. This approach enables future sentimental analysis of public mood by identifying words in posts and headlines.

Analysis of Text Data: The VaderSentiment module, a Python-based Natural Language Processing (NLP) tool, was used to analyse sentiments in social media posts. VADER, also known as Sentiment Reasoner and Valence Aware Dictionary, is a sentiment analysis tool that evaluates the text and determines the polarity of sentiment. The tool is used in various fields, including social media monitoring, consumer feedback assessment, and decision-making based on sentiment. The VADER system categorises feelings into three categories: positive, negative, and neutral, collectively known as "sentiments." The tool was used to determine each written piece's emotional tones and dispositions, enabling the categorisation of messages based on their corresponding emotions. The data was successfully classified into various sentiment categories, with values closer to 1 indicating a strong positive attitude, more comparable to -1 indicating a strong negative attitude, and values around 0 indicating neutrality.

The present study used the VADER sentiment analysis module to perform sentiment analysis on a well-curated and processed data set. The data set exhibited a temporal arrangement. The data set showed a temporal arrangement. Implementing this approach has given us a deeper understanding of the temporal dynamics of public sentiment. This application obtained noteworthy discoveries about the fluctuating patterns of public sentiment within a specific timeframe.

2) Dataset Preparation for Transactional Data: Data acquisition was conducted from multiple sources during the designated phase of data collection, after which the data was systematically categorised into files according to their respective dates. Utilising delimiters within the Microsoft Excel software has proven instrumental in expediting data unification. This functionality dramatically simplifies the consolidation and standardisation of dates, even when they are initially presented in varying formats. The gathered data was compiled from various textual groups following the initial data collection phase. Regarding the transaction data, it was observed that no null or empty locations were present, thereby preventing the need for any preprocessing steps.

Ultimately, the outcomes from transactional and textual data were mixed. To accomplish this goal, the strategy of "merging" was implemented. The left side of the dataset contains the transactional data, while the right side contains the sentiment polarity data. A decision was made to use a left join technique to ensure that all relevant transaction information would be included in the final record. All relevant textual data was incorporated into the organisation's archives, providing supplementary information. The material that has been compiled here is extensive, and the appropriate application of it should be considered. The comprehensive report encompasses a combination of quantitative trade data and qualitative analysis. It's time to examine the collected data and determine its meaning.

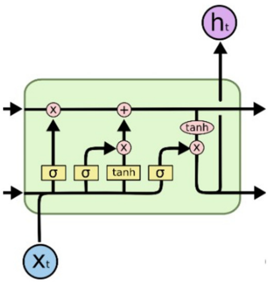

Fig.2: *An LSTM cell [1].*

C. Long Short Term Memory(LSTM)

Neural networks, known for their ability to handle nonlinear tasks, have wide parameter dimensions and nonlinear activation functions in each layer. Recurrent Neural Networks (RNNs) were developed to address the inability of simple neural networks to understand input sequences depending on the context. However, they still need to improve, like Vanishing Gradients and long-term dependency. To overcome these issues, Hocreiter and Schmidhuber presented the Long Short-Term Memory (LSTM) method, which maintains a cell state in the memory line by implementing gates controlling the information flow. Long Short-Term Memory (LSTM) is a specific variation of the Recurrent Neural Network (RNN) network with a modified hidden layer module structure. RNNs can handle nonlinear time series but face challenges in managing extreme delays and training with precalculated delay window lengths. The LSTM model evolved to achieve long-term memory by replacing RNN cells in the hidden layer with LSTM cells, allowing it to learn long-term relationships making it suitable for time series analysis. The three gates that make up a long short-term memory are the input gate, the neglect gate, and the output gate, which operate between 0 and 1 and are based on the sigmoid function. These gates are described as follows:

1) Forget Gate

The forget Gate controls the flow of information from the previous time step. It considers the previous hidden state and the current input to decide how to filter information from the previous cell state. It is formulated as in equation 1.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

where W_f is the weight matrix, b_f is the bias term of the forgetting Gate, x_t and h_{t-1} are the current input and the previous output of the LSTM cell, respectively, and σ is the sigma function.

2) Input Gate

It decides upon the new information to be added to the cell state. It does so by filtering how much information has to be kept from the current input and the previous hidden state. It is defined by equation 2 and equation 3.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t + b_c) \tag{3}$$

The current cell state can be computed by forgetting the useless information and adding new information. Hence, by using equation 1, 2 and 3, the current cell state can be formulated as given in equation 4.

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \tag{4}$$

3) Output Gate

Since now the LSTM cell has the current cell state, the output gate is used to produce the current hidden state of the cell. It is produced using the current input, the previous hidden state, and the current cell state. It is given by the equation 6.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t \cdot \tanh(c_t) \tag{6}$$

1) LSTM Hyperparameters:

1) Hidden layers and Number of nodes- Between the input and output layers are hidden layers with adjustable node counts. Too few nodes may result in under-fitting, while too many nodes can improve accuracy. Units considered for the search space: [16, 32, 64].
2) Dropout Rate - Dropout layers placed between LSTM layers randomly exclude neurons during training to reduce sensitivity to specific weights and minimise over-fitting. Considered dropout rates for search space: [0.1, 0.2, 0.3].
3) Optimiser - Optimisers modify weights and learning rates during DL model training, reducing loss and improving precision. Considered optimisers for search space: ['RMSprop', 'Adam'].
4) Number of Epochs - This hyperparameter determines the number of complete iterations over the dataset during training. The number of epochs considered for search space: [10, 20, 30].
5) Batch Size - The number of samples analysed before changing the model's internal parameters is determined by batch size. Greater batch sizes yield larger gradient steps for the same number of samples. Considered batch sizes for search space: [50, 100]

Table 2: *Experimental Parameters LSTM.*

Parameter	Value
Hidden Layers	7, 15, 30
Dropout Rate	0.2
Optimizer	adam
Epochs	500
Batch Size	1

In evaluating the LSTM (Long Short-Term Memory) model, a systematic methodology was used to divide the dataset into distinct training and testing subsets. The dataset was organised in a time series format, where the first 70% of the dataset was allocated for training purposes in all of the experiments described in this paper. By using this approach, it was ensured that the model acquired relevant knowledge from previous instances.

In contrast, the remaining 30% of the time series was allocated for experimentation and evaluation, including the final segment. These assessments evaluated the model's ability to extrapolate to novel data and provide precise prognostications. This experiment series aims to recreate circumstances in which the model is presented with unfamiliar data.

The continuous division strategy was used in all studies to preserve chronological integrity across the training and testing datasets. Through the use of this methodology, temporal components in the dataset were successfully included, hence enabling comprehensive evaluations of the LSTM model's performance.

4. PERFORMANCE EVALUATION

The evaluation of the efficacy of a Long Short-Term Memory (LSTM) model often relies on the use of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) as the prevailing metrics. The factors above play a crucial role in assessing the LSTM model's ability to forecast fluctuations in the stock market. The root mean square error (RMSE) is a valuable metric for assessing the degree of forecast inaccuracies. The objective above is accomplished by computing the square root of the mean deviations, which have been squared between the anticipated and observed values. The Mean Absolute Error (MAE) is a statistical metric used to assess the average size of mistakes. It is calculated by taking the average absolute disparities between the predicted and actual results. The present study uses several measures to assess the efficacy of the Long Short-Term Memory (LSTM) model in predicting stock market trends. Additionally, this research thoroughly examines the model's predictive capacities.

1) Mean Absolute Error (MAE): To calculate the Mean Absolute Error, the absolute difference between the actual and forecasted values is calculated, and finally, the average of the differences of the testing set is mentioned.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7}$$

where n is the total number of data points, y_i represents the actual observed value for the ith data point, and $\hat{y}_i$ represents the predicted value by the model for the ith data point.

2) Root Mean Square Error (RMSE): The root-mean-square error (RMSE) is a commonly used measure of the differences between predicted and observed values (sample or population values). The RMSE is the square root of the second sample moment of discrepancies between predicted and observed values or the quadratic mean of these differences. When calculated over the data sample used for estimation, these deviations are called residuals, and when computed out-of-sample, they are called errors (or prediction errors). The RMSE aggregates the magnitudes of prediction errors for various data points into a single measure of predictive power. Because it is scale-dependent, RMSE is a measure of accuracy used to compare forecasting errors of different models for a specific dataset rather than between datasets.

$$RMSE = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - x_i)^2} \tag{8}$$

where n is the total number of data points, y_i is ith actual value and x_i is i$^{\text{th}}$ actual value.

5. RESULTS AND DISCUSSION

A. Client

The client is when the trades are executed on client accounts. Regarding the implementation part, the client trading happening in the BSE stock exchange is analysed and forecasted with reasonable accuracy. This is achieved with various methods of testing with directly extracted transaction data and also with sentimental analysis, which is an influential method from social media and financial news headlines for different lookback days:

1) Forecasting of Client Buy Results:

- **Look Back 7 days and without sentimental analysis and with sentimental analysis:** Root Mean Square Error and Mean Absolute Error approximately 39.11 per cent reduction in RMSE value also 60.97 per cent reduction in MAE value by introducing Sentiment analysis with look back as 7 days.
- **Look Back 15 days and without sentimental analysis and with sentimental analysis:** There is almost 26.77 per cent reduction in MAE value and a 16.71 per cent reduction in RMSE value when sentiment analysis is included

- **Look Back 30 days without sentimental analysis and with sentimental analysis:** About 28.59 per cent and 43.79 per cent reductions in MAE and RMSE values were observed when evaluated using sentiment analysis.

2) Forecasting of Client Sales Results:

- **Look Back 7 days and without sentimental analysis and with sentimental analysis:** Root Mean Square Error and Mean Absolute Error, approximately 32.64 per cent reduction in RMSE value also 63.65 per cent reduction in MAE value by introducing Sentiment analysis with look back as 7 days.
- **Look Back 15 days and without sentimental analysis and with sentimental analysis:** There is almost 35.61 per cent reduction in MAE value and 72.94 per cent reduction in RMSE value when sentiment analysis is included .
- **Look Back 30 days without sentimental analysis and with sentimental analysis:** About 57.52%, 87.22% reduction in MAE, RMSE values when evaluated using sentiment analysis.

B. Proprietary

For implementation, the proprietary trading analysis that occurred in the BSE stock market was analysed and forecasted with greater accuracy. This was achieved with various scenar- ios of testing with plain transaction data and also with social media and financial news headlines sentimental analysis as influence for varying days for look back:

1) Forecasting of Proprietary Buys Results:

- **Look Back 7 days and without sentimental analysis and with sentimental analysis:** On comparing error results, there is a change of 64.47 per cent in RMSE error also 66.42 per cent in MAE error with the introduction of sentimental analysis with having a look back as 7 days.
- **Look Back 15 days and without sentimental analysis and with sentimental analysis:** On comparing error results, there is a change of 50.70 per cent in RMSE error also 42.48 per cent in MAE error with the introduction of sentimental analysis with having a look back as 15 days.
- **Look Back 30 days and without sentimental analysis and with sentimental analysis:** On comparing error results, there is a change of 59.86 per cent in RMSE error also 56.26 per cent in MAE error with the introduction of sentimental analysis with having a look back as 30 days.

2) Forecasting of Proprietary Sales Results:

- **Look Back 7 days without sentimental analysis and with sentimental analysis:** On comparing error results, there is a change of 64.47 per cent in RMSE error and 59.17 per cent in MAE error with the introduction of sentimental analysis with having the look back as 7 days.
- **Look Back 15 days and without sentimental analysis and with sentimental analysis:** On comparing error results, there is a change of 50.15 per cent in RMSE error also 46.97 per cent in MAE error with the introduction of sentimental analysis with having a look back as 15 days.
- **Look Back 30 days and without sentimental analysis and with sentimental analysis:** On comparing error results, there is a change of 56 per cent in RMSE error also 52.61 per cent in MAE error with the introduction of sentimental analysis with having a look back as 30 days.

C. NRI - Non-Residential Indian

Regarding the implementation part, the NRI trading hap- pening in the BSE stock exchange is analysed and forecasted with good accuracy. This is achieved with various methods of testing with directly extracted transaction data and also with sentimental analysis, which is an influential method from social media and financial news headlines for different lookback days :

1) Forecasting of NRI Buy Results:

- **Look Back 7 days and without sentimental analysis and with sentimental analysis:** For Root Mean Square Error and Mean Absolute Error, approximately 10.10 per cent reduction in MAE value also 62.91 per cent reduction in RMSE value by introducing Sentiment analysis with look back as 7 days.
- **Look Back 15 days without sentimental analysis and with sentimental analysis:** There is almost a 12.80 per cent reduction in MAE value and 58.46 per cent reduction in RMSE value while including sentiment analysis.
- **Look Back 30 days without sentimental analysis and with sentimental analysis:** Observing 19.49 per cent and 56.99 per cent reduction in MAE and RMSE values when evaluated using sentiment analysis.

2) Forecasting of NRI Sales Results:

- **Look Back 7 days and without sentimental analysis and with sentimental analysis:** For Root Mean Square Error and Mean Absolute Error, approximately 3.33 per cent reduction in MAE value also 57.65 per cent reduction in RMSE value by introducing Sentiment analysis with look back as 7 days.
- **Look Back 15 days without sentimental analysis and with sentimental analysis:** There is almost a 63.39 per cent reduction in MAE value and 80.62 per cent reduction in RMSE value while including sentiment analysis.
- **Look Back 30 days without sentimental analysis and with sentimental analysis:** Observing 46.97 per cent, 68.45 per cent reduction in MAE, RMSE values when evaluated using sentiment analysis.

Table 3: *With and Without Sentimental Analysis.*

Categories Name	Buy/Sales	Days	Without Sentiment		Without Sentiment	
			RMSE	***MAE***	***RMSE***	***MAE***
Client	Buy	7	0.0115	0.0033	0.0296	0.0054
		15	0.0129	0.0089	0.0155	0.0065
		30	0.0086	0.0049	0.0154	0.0068
	Sales	7	0.0115	0.0065	0.0316	0.0097
		15	0.0096	0.0070	0.0356	0.0109
		30	0.0059	0.0048	0.0343	0.0114
Proprietary	Buy	7	0.0953	0.0633	0.0320	0.0227
		15	0.0760	0.0033	0.0374	0.0292
		30	0.0795	0.0526	0.0319	0.0230
	Sales	7	0.0978	0.0650	0.0347	0.0265
		15	0.0672	0.0474	0.0335	0.0251
		30	0.0703	0.0462	0.0309	0.0219
NRI	Buy	7	0.0175	0.0088	0.0474	0.0098
		15	0.0200	0.0149	0.0483	0.0170
		30	0.0210	0.0141	0.0490	0.0175
	Sales	7	0.0108	0.0106	0.0280	0.0109
		15	0.0054	0.0036	0.0282	0.0100
		30	0.0089	0.0103	0.0283	0.0110
DII	Buy	7	0.0956	0.0647	0.0391	0.0331
		15	0.0931	0.0635	0.0400	0.0341
		30	0.0880	0.0630	0.0291	0.0227
	Sales	7	0.1260	0.0878	0.0472	0.0365
		15	0.1163	0.0826	0.0444	0.0336
		30	0.1319	0.0928	0.0625	0.0553
OHLC		7	0.0156	0.0120	0.0109	0.0085
		15	0.0487	0.0364	0.0107	0.0077
		30	0.0259	0.0179	0.0116	0.0082

D. Domestic Institutional Investors

- **Look Back 7 days without sentimental analysis and with sentimental analysis:** On comparing error results, there is a change of 56 per cent in RMSE error and 31 per cent in MAE error with the introduction of sentimental analysis and looking back as 7 days.
- **Look Back 15 days and without sentimental analysis and with sentimental analysis:** On comparing error results, there is a change of 53 per cent in RMSE error and 29 per cent in MAE error with the introduction of sentimental analysis with looking back 15 days.
- **Look Back 30 days and without sentimental analysis and with sentimental analysis:** On comparing error results, there is a change of 59 per cent in RMSE error and 40 per cent in MAE error with the introduction of sentimental analysis with having a look back as 30 days.

2) Forecasting of Domestic Institutional Investors sales Results:

- **Look Back 7 days and without sentimental analysis and with sentimental analysis:** On comparing error results, there is a change of 80 per cent in RMSE error also 50 per cent in MAE error with the introduction of sentimental analysis with having a look back as 7 days.
- **Look Back 15 days and without sentimental analysis and with sentimental analysis:** On comparing error results, there is a change of 70 per cent in RMSE error also 50 per cent in MAE error with the introduction of sentimental analysis with having a look back as 15 days.
- **Look Back 30 days and without sentimental analysis and with sentimental analysis:** On comparing error results, there is a change of 70 per cent in RMSE error also 40 per cent in MAE error with the introduction of sentimental analysis with having a look back as 30 days.

E. Forecasting the BSE Stock Market Prices

In this section, a variable called OHLC, which stores the average of the 4 Open, High, Low and Close values, is calculated, and later, it is used to forecast the BSE stock market.

1) Forecasting of OHLC BSE Stock Market Prices:

- **Look Back 7 days and without sentimental analysis and with sentimental analysis:** On comparing error results, there is a change of 30.23 per cent in RMSE error also 28.94 per cent in MAE error with the introduction of sentimental analysis with having a look back as 7 days.
- **Look Back 15 days and without sentimental analysis and with sentimental analysis:** On comparing error results, there is a change of 78.02 per cent in RMSE error also 78.65 per cent in MAE error with the introduction of sentimental analysis with having a look back as 15 days.
- **Look Back 30 days and without sentimental analysis and with sentimental analysis:**

On comparing error results, there is a change of 55.10 per cent in RMSE error also 53.92 per cent in MAE error with the introduction of sentimental analysis with having a look back as 30 days.

F. Discussion

In conclusion, the research on Stock price prediction using heterogeneous data, it can be said that the deviations observed in every category by taking a different look back days:

1) Client: Using sentiment analysis in Clients' Buy data has significantly reduced error for both MAE and RMSE for varying lookback days. The average reduction was 31.49 per cent for MAE and 33.20 per cent for RMSE. The moderate number of lookback days and sentiment analysis resulted in higher accuracy.

Clients Sales showed reduced error changes in MAE and RMSE values for various lookback days with sentiment analysis, compared to Clients Buy. The LSTM model better- predicted transactions with sentimental analysis, resulting in an average reduced error of 52.29 per cent for MAE and 62.73 per cent for RMSE.

2) Proprietary: In Proprietary Buy, there is considerable change in MAE and RMSE values while using sentiment analysis and without. That being said, there has been an average improvement of 56.05 per cent in MAE and 58.34 per cent in RMSE using sentimental analysis than without.

Compared to Proprietary Buy, Proprietary Sales have also reduced MAE and RMSE values equally well when using sentiment analysis for different lookback days. On average, the MAE value has decreased by 52.92 per cent, and the RMSE value has reduced by 56.87 per cent using sentiment analysis to with sentiment analysis.

3) NRI: Regarding NRI Buy, there has been a change in the MAE and RMSE values being reduced with the sentimental analysis as an extra feature and various lookback days. On average, the MAE error was decreased by 14.13 per cent, and the RMSE error was reduced by 59.45 using sentimental analysis for different lookback days.

On the contrary, with NRI Sale, it has been seen that the MAE and RMSE have performed better with sentimental analysis on varying lookback dates. On average, the MAE value decreased by 37.90 per cent, and the RMSE value decreased by 68.90 per cent with sentimental analysis.

4) DII: In DII buy, a decrement is observed in the MAE and RMSE values changes with sentiment analysis for different lookback days. The decreases may be primarily observed in the RMSE values but not much in the MAE values. On average, a 40% reduction in MAE value and a 29.67% reduction with RMSE are observed.

In DII sales, a decrement is observed in the MAE and RMSE values changes with sentiment analysis for different lookback days. Perhaps decrements are primarily observed in RMSE values but only slightly in MAE values. On average, a 55.61 per cent reduction in MAE value and a 69.67 per cent reduction with RMSE is observed.

5) OHLC: As the most popular feature in predicting the stock market, there is a substantial increase in MAE and RMSE values change while using sentiment analysis and without. It is observed that there has been an average improvement of 53.76 per cent in MAE value and 55.98 per cent in RMSE value using sentiment analysis than without.

6. CONCLUSIONS

In summary, the results of this study unambiguously indicate that the enhancement of stock market prediction may be achieved by a planned amalgamation of diverse data sources rather than only depending on conventional transactional data. Compared to traditional approaches, the LSTM-based model consistently demonstrates more excellent performance. The intervention showed significant efficacy in reducing error rates, with an average decrease of 43.66% for Mean Absolute Error (MAE) and 50.93% for Root Mean Square Error (RMSE) across diverse experimental circumstances. This supports the notion that incorporating various data, such as sentiment analysis research, into the field of stock market prediction is very advantageous.

However, it is crucial to comprehensively comprehend the constraints evident throughout this empirical investigation's progression. Notably, there was a lack of social media or internet news content primarily about the Indian market. This observation is particularly noteworthy since it can potentially restrict the comprehensiveness of sentiment analysis. Implementing a strategic diversification strategy is proposed for these data sources to enhance the precision of predictions about the stock market. This will enable us to use a broader range of data points. Including a wider range of social media and news narratives in this research would enhance its overall value and effectiveness. Implementing this strategic expansion is expected to provide more favourable outcomes in sentiment analysis and the prediction of stock market trends.

In summary, the results of this research provide a strong case for including diverse data sources and acknowledging sentiment analysis as a crucial component in predicting stock market trends. This study provides empirical evidence to demonstrate the benefits of using this approach compared to conventional techniques that rely only on transactional data. In subsequent periods, it is advisable to direct research efforts on addressing the constraints associated with the available data and devising innovative approaches to enhance the precision of predictions within the dynamic landscape of financial markets.

7. FUTURE WORK

The potential of this undertaking in the future is highly promising in various significant aspects. To begin with, a note-worthy opportunity exists to automate the current workload, thereby converting it into an efficient pipeline suitable for im- mediate utilisation. This will significantly improve operational effectiveness and promptness, facilitating more expeditious decision-making and flexibility in ever-changing markets. Further, investigating more sophisticated algorithms to enhance precision constitutes a captivating domain. The continuous progression of technology will enable the implementation of cutting-edge algorithms in data analysis and predictive modelling, resulting in improved accuracy and dependability of insights. Lastly, it is a strategic move to extend the implementation of the current algorithm to stock markets with more significant social media traffic. By granting access to a broader and potentially more influential data source via this extension, market sentiment and trends can be comprehended in greater depth. These prospective developments collectively signify an ever-evolving trajectory towards enhanced and all-encompassing stock market analysis.

References

[1] U. Singh, S. Tamrakar, K. Saurabh, R. Vyas and O. P. Vyas, "Hyperparameter Tuning for LSTM and ARIMA Time Series Model: A Comparative Study," *2023 IEEE 4th Annual Flagship India Council International Subsections Conference (INDISCON)*, Mysore, India, pp. 1-6, 2023.

[2] K. Saurabh, S. Sood, P. A. Kumar, U. Singh, R. Vyas, O. P. Vyas, and R. Khondoker, "LBDMIDS: LSTM Based Deep Learning Model for Intrusion Detection Systems for IoT Networks," *2022 IEEE World AI IoT Congress (AIIoT)*, Seattle, WA, USA, pp. 753-759, 2022.

[3] K. Saurabh, A. Singh, U. Singh, O. P. Vyas and R. Khondoker, "GANIBOT: A Network Flow Based Semi Supervised Generative Adversarial Networks Model for IoT Botnets Detection," *2022 IEEE International Conference on Omnilayer Intelligent Systems (COINS)*, Barcelona, Spain, pp. 1-5, 2022.

[4] U. Singh, T. Musale, R. Vyas, and O. Vyas, "Agricultural plantation classification using transfer learning approach based on cnn," *arXiv preprint arXiv:2206.09420*, 2022.

[5] K. Saurabh, T. Kumar, U. Singh, O. P. Vyas and R. Khondoker, "NFDLM: A Lightweight Network Flow based Deep Learning Model for DDoS Attack Detection in IoT Domains," *2022 IEEE World AI IoT Congress (AIIoT)*, Seattle, WA, USA, pp. 736-742, 2022.

[6] U. Singh, K. Saurabh, N. Trehan, R. Vyas and O. P. Vyas, "Terrain Classification using Transfer Learning on Hyperspectral Images: A Comparative study," *2022 IEEE 19th India Council International Conference (INDICON)*, Kochi, India, pp. 1-6, 2022.

[7] B. Krollner, B. J. Vanstone, and G. R. Finnie, "Financial time series forecasting with machine learning techniques: a survey," in *The European Symposium on Artificial Neural Networks*, 2010. [Online]. Available: `https://api.semanticscholar.org/CorpusID:14947431`

[8] P. S. Rao, K. Srinivas, and A. K. Mohan, "A survey on stock market prediction using machine learning techniques," in *ICDSMLA 2019*, A. Kumar, M. Paprzycki, and V. K. Gunjan, Eds. Singapore: Springer Singapore, pp. 923–931, 2020.

[9] V. Derbentsev, A. V. Matviychuk, N. Datsenko, V. Bezkorovainyi, and A. Azaryan, "Machine learning approaches for financial time series forecasting," in *M3E2-MLPEED*, 2020. [Online]. Available: `https://api.semanticscholar.org/CorpusID:229356996`

[10] A. Dingli and K. S. Fournier, "Financial time series forecasting – a deep learning approach," *International Journal of Machine Learning and Computing*, vol. 7, pp. 118–122, 2017. [Online]. Available: `https://api.semanticscholar.org/CorpusID:159007072`

[11] J. Cao, Z. Li, and J. Li, "Financial time series forecasting model based on ceemdan and lstm," *Physica A: Statistical mechanics and its applications*, vol. 519, pp. 127–139, 2019.

[12] C. M. Liapis, A. Karanikola, and S. Kotsiantis, "A multi-method survey on the use of sentiment analysis in multivariate financial time series forecasting," *Entropy*, vol. 23, no. 12, 2021. [Online]. Available: `https://www.mdpi.com/1099-4300/23/12/1603`

[13] Z. Drus and H. Khalid, "Sentiment analysis in social media and its application: Systematic literature review," *Procedia Comput. Sci.*, vol. 161, no. C, pp. 707–714, jan 2019. [Online]. Available: `https://doi.org/10.1016/j.procs.2019.11.174`

[14] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia and D. C. Anastasiu, "Stock Price Prediction Using News Sentiment Analysis," *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*, Newark, CA, USA, pp. 205-208, 2019.

[15] R. Gupta and M. Chen, "Sentiment Analysis for Stock Price Prediction," *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, Shenzhen, China, pp. 213-218, 2020.

Uphar Singh is currently working towards her Ph.D. in the Department of IT at the Indian Institute of Information Technology Allahabad. Her research interests span across heterogeneous data analytics, time series analytics, and process mining. Uphar Singh's academic journey includes the successful completion of her post-graduation (M.Tech.) from the Department of IT at IIIT Allahabad. Prior to her post-graduation, she earned her B.Tech. in Computer Science and Engineering from the United College of Engineering and Research, Allahabad, affiliated with Dr. A.P.J. Abdul Kalam University, Lucknow. Her unwavering passion for research is evident in her dedication to exploring the applications of artificial intelligence and deep learning in the fields of time-series data analysis and process mining. Uphar Singh has made remarkable contributions to these domains, which are reflected in her numerous notable publications. Her commitment to advancing the frontiers of knowledge in these areas is driving her towards becoming a prominent figure in the field of research.

Dhanush Vasa an Indian born-Australian Citizen, embarked on a remarkable educational journey that led him from Australia to the prestigious Indian Institute of Information Technology in Allahabad (IIITA). After completing his early education in Australia, Dhanush's pursuit of a passion for Information Technology brought him to India for his further studies. During his undergraduate studies at the institute, Dhanush was actively engaged in research projects specializing in Time Series Data Forecasting with Deep Learning models. Along the way, he cultivated a valuable network of colleagues and mentors who greatly contributed to his academic and research pursuits.

Lekhana Reddy Mitta earned her Bachelor's degree in the year 2023 from the distinguished Information Technology department of her academic institution. During her undergraduate years, she wholeheartedly delved into the world of Machine Learning and Deep Learning models, with a particular focus on their applications in Forecasting Time Series Data. Her dedication to this field was evident as she actively engaged in various research projects. Lekhana not only excelled in her studies but also cultivated a robust network of mentors and peers who played a crucial role in shaping her academic journey. Her ability to connect with experts in the field and collaborate with like-minded individuals further propelled her towards achieving excellence in the field of data science and machine learning.

Kumar Saurabh is an accomplished researcher with over seven years of industrial experience. Currently, he is pursuing his Ph.D. from the Department of IT at the Indian Institute of Information Technology Allahabad, under joint supervision from Technische Hochschule Mittelhessen (THM), Friedberg, Germany. He completed his post-graduation (M.Tech.) from the Department of IT at IIIT Allahabad and obtained his B.Tech. in Electronics & Communication Engineering from IITT College of Engineering, affiliated with Punjab Technical University. He is passionate about research, with a focus on the application of Artificial Intelligence in the Time-series Data Analysis and cybersecurity of Industrial IoT. He has made significant contributions to these areas, resulting in several notable publications.

Ranjana Vyas is an Assistant Professor at the Indian Institute of Information Technology-Allahabad (IIITA) and the Coordinator of NewGen IEDC. She has a background in Business Informatics and has stayed in Europe for over 18 months. Dr. Ranjana received a prestigious DAAD Scholarship and worked as a Post Doctoral Fellow at Germany's Technical University of Kaiserslautern. She has taught subjects like Software Requirement Engineering, Digtal Marketing, and Business Process Modeling. Dr. Ranjana has academic interests in Data Science and Process Science, particularly Recommender Systems, Churn Prediction, and Business Process Analytics. She collaborated with Prof. Artus KG of University Paderborn-Germany and was invited to Germany's Hof University of Applied Sciences for a talk titled 'Business Process Re-engineering: from BPM to Process Mining'. Dr. Ranjana has organized various events, including IT Entrepreneurship mentoring programs, Hackathons, and Symposiums. She has published approximately 30 articles in peer-reviewed journals and prestigious conferences. Dr. Ranjana has also participated in several overseas assignments, including a visit to Bielefeld University, Germany, and a research collaboration visit to the Department of Business Intelligence in Paderborn, Germany. She also studied in Germany from 1996-97 for 18 months.

O. P. Vyas is currently a Professor in the Department of Technology at the Indian Institute of Information Technology Allahabad (IIITA), India. He also holds the positions of Dean (Technology & Development) and Dean (Institute Works Department) at IIIT Allahabad. Prof. Vyas has contributed in starting a B.Tech. program in Business Informatics course at the IIIT Allahabad as Coordinator, which is the first Business Informatics program in technology domain in India. O. P. Vyas earned his M.Tech. degree in Computer Science from IIT Kharagpur, India, and a Ph.D. research in jointly from IIT Kharagpur, India and the Technical University of Kaiserslautern, Kaiserslautern, Germany. Having worked as a Visiting Professor at the University of Bielefeld, Germany, University of Paderborn, Germany alongwith Visiting Researcher at Inria-Lille, France Prof. Vyas has made significant contributions to the field with over 150 research publications. He successfully completed an Indo-German project under the Department of Science and Technology, Bundesministerium für Bildung und Forschung, an Indo-French Project with Inria-France, an Indo-Norwegian Project with NTNU-Gjovik-Norway. His academic journey includes being a recipient of the Deutscher Akademischer Austauschdienst (German Academic Exchange Service) Fellowship from the Technical University of Kaiserslautern, and the Association for Overseas Technical Scholarship Fellowship from the Center of International Cooperation for Computerization, Japan. Previously, he has also served as a Professor and Dean (Academic) at the International Institute of Information Technology, Naya Raipur.