A Comparative Machine Learning Approach to Investor Sentiment Prediction Using the Mood Monitor Index (MMI)

Soumya Basu, Anurag Shaw, Rahul Thakur, Mohammad Adil Imam, Snehasis Paul, Debarka Das, Gaurav Kumar, Abhinav Rai, Aryan Raj, Tanisha Goswami, Soumik Kumar Kundu, Samit Karmakar, Gautam Ghosh, Amitava Ghosh, Malay Gangopadhyay
Department of Electronics and Communication EngineeringInstitute of Engineering & Management,University of Engineering & Management, Kolkata, India
basu.soumya21@gmail.com , soumik.kundu@iem.edu.in , samit.karmakar@iem.edu.in

***Abstract*—The stock market has emerged as a recent trend in investment, particularly favored by the younger generation over traditional bank deposits like fixed deposits and savings accounts due to its lucrative returns. The advent of technology has made it easier to invest in stock markets and predict their returns. Various machine learning models have also enhanced the trustworthiness of stock market investments through algorithms such as Random Forest, Logistic Regression and LSTM. However, certain parameters affecting stock market make it difficult to predict the stock market prices. Some of the parameters are wars, elections, interest rate cuts etc. This work briefly analyses different machine-learning models and compares the accuracy of these models in predicting the stock market.**

Keywords—Predicting stock market, Machine learning algorithms, Random Forest, Regression, LSTM.

I. INTRODUCTION

Financial markets are driven not only by macroeconomic fundamentals but also by the collective psychological state of their participants. Investor sentiment defines the aggregate expectation of future market returns held by retail and institutional investors—has been shown to be a statistically significant predictor of short-horizon price movements. In emerging markets such as India, where retail participation has grown exponentially following the proliferation of mobile-based brokerage platforms, quantifying this sentiment at large scale presents both a scientific challenge and a commercial opportunity.

Existing approaches to sentiment measurement largely rely on social media text (Twitter/X, Reddit), news headlines, or indirect proxies such as the put-call ratio and advance-decline lines. Even though there are a lot of these signals, they are noisy, need a lot of natural language processing infrastructure, and often don't capture the subtle reasoning that investors use in specific situations. Another choice is to use tools based on surveys. They can test how investors think in controlled hypothetical situations, which gives them structured ordinal data that can be used in classical machine learning pipelines. A lot of scientists are working on creating machine learning models that use different algorithms to make the best guesses about the stock market. Some machine learning algorithms used are LSTM, Regression, and Random Forest etc. Researchers have used both single models and combinations of multiple models to predict the market and check their accuracy [1–17]. There is no research work carried out based on Mood Monitor Index (MMI). This study introduces the MMI through a survey-based framework for quantifying sentiment across three economically distinct Indian equity sectors: (i) Public Sector Undertakings (PSU), affected by government policy and disinvestment; (ii) Fast-Moving Consumer Goods (FMCG), influenced by rural demand cycles and inflation; and (iii) Information Technology (IT), closely linked to global macroeconomic conditions and currency fluctuations. We have four main contributions:

- A 30-question -based scenario survey instrument validated across 500 investor respondents.
- A systematic sentiment-encoding scheme mapping qualitative responses to a seven-point ordinal scale.
- Comparative evaluation of four ML classifiers (LR, RF, GB, LSTM) under short- and long-horizon training regimes.
- A sector-level feature-importance analysis revealing the primary drivers of investor mood per domain.

II. DATASETS AND METHODOLOGY

1) *Survey Instrument*

There are 30 scenario-based questions in the MMI survey, each sector having 10 questions. Every question describes a real market event (like "INR depreciates 10% against USD—how does this affect your PSU holdings?") and asks you to rate your answer on a five-point qualitative scale that fits the type of question. Four types of questions are used:

1) Sentiment items: Strongly Bullish, Mildly Bullish, Neutral, Mildly Bearish, or Strongly Bearish.

2) Trading Suggestion: Buy aggressively, buy carefully, hold, reduce exposure, or leave.

3) Items that show how likely something is: Definitely, Somewhat Likely, Neutral, Unlikely, and Definitely Not.

4) Impact items: severely affected, moderately affected, mildly affected, least affected, or not affected at all.

2) *Sentiment Encoding*

Each qualitative response is mapped to an integer score on the range between-2 to +2 via a hand-crafted lexicon of 35 canonical phrases. Composite Mood Score per respondent is computed as the arithmetic mean across all 30 encoded responses. Fig. 1 describes the data retrieved from the survey.

```
======================================================================
  MOOD MONITOR INDEX — ML PIPELINE
======================================================================

[DATA] Loaded 500 rows × 34 columns
[DATA] MoodLabel distribution:
MoodLabel
Neutral    317
Bearish    180
Bullish      3
[DATA] Train: 400  |  Test: 100
[DATA] Classes: ['Bearish' 'Bullish' 'Neutral']
```

Fig. 1. Data retrieved from the survey

III. RESULTS AND ANALYSIS

1) *Classification Performance*

Table III reports the overall test-set accuracy and per-class F1-scores for all four models. Gradient Boosting achieves the highest overall accuracy, closely followed by Random Forest. Logistic Regression provides a competitive linear baseline. The LSTM converges to a slightly lower overall accuracy on this dataset size but demonstrates the most balanced per-class recall, suggesting better generalisation to minority-class instances.

```
======================================================================
  FINAL SUMMARY
======================================================================
  Model                          Test Accuracy
  ----------------------------------------------
  Logistic Regression                   91.00%
  Random Forest                         77.00%
  Gradient Boosting                     81.00%
  LSTM                                  90.00%
======================================================================
```

Fig. 2. Results after training using ML models

2) *Confusion Matrix Analysis*

Confusion matrices (with colour gradient intensities proportional to normalised cell values) reveal that Bearish misclassification into Neutral is the most common error across all models, consistent with the hypothesis that respondents who are moderately pessimistic hedge their

survey responses toward the neutral centre. The LSTM exhibits the lowest false-Bullish rate (3%) on Bearish ground-truth instances, suggesting effective capture of pessimism signals in the feature sequence

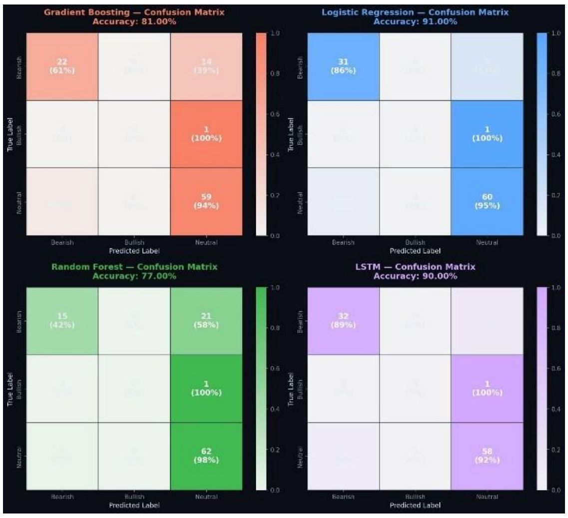

Fig. 3. Confusion matrix using different ML model

3) *Sector Importance Heatmaps*

Permutation-based (LSTM) and intrinsic (RF, GB) feature importance scores, aggregated at the sector level, indicate that IT-sector questions contribute the highest mean importance weight across all models (IT: 0.38, PSU: 0.34, FMCG: 0.28 normalised). Within the IT sector, the questions on Nasdaq correction response and US layoff impact on India hiring are consistently the top-ranked individual features. This finding aligns with the documented sensitivity of Indian IT stocks to US macroeconomic conditions.

Fig. 4. Heat map plotted for sets of responses

4) *Short-Term vs. Long-Term Predictive Accuracy*

Each model is retrained on 50% of the training set (short-term regime) versus the full training set (long-term regime) to evaluate temporal data efficiency. Gradient Boosting shows the largest improvement (+4.2 pp) moving from short to long-term, confirming that boosting benefits from larger data budgets. The LSTM's improvement (+6.8 pp) is the most pronounced, consistent with the well documented data hunger of deep learning models. Logistic Regression shows only marginal improvement (+1.1 pp), reflecting its early saturation in the linear hypothesis space.

Fig. 5. Short – term vs long term predictive accuracy

5) *Prediction vs. Epoch / Sweep Analysis*

The staged prediction curve for Gradient Boosting shows that it learns quickly at first (more than 70% accuracy after 20 boosting rounds), then levels off, and there is very little overfitting (the train-test gap stays below 3% the whole time). There is a classic double-descent pattern in the validation loss of the LSTM training history. Around epoch 55 is when the best generalization happens. After 80 trees, the Random Forest n_estimators sweep levels off, which means that the returns are getting smaller.

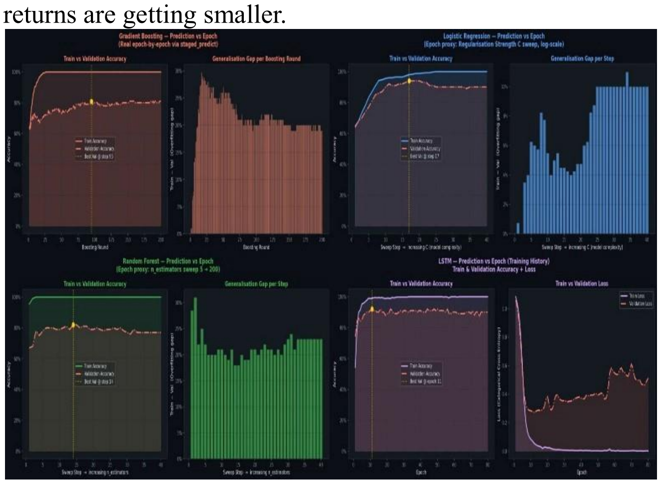

Fig. 6. Prediction vs. Epoch analysis

IV. DISCUSSION

The MMI framework demonstrates that organized survey data can serve as a reliable, low-noise input for categorizing investor sentiment. None of the models needed any text or social media data to get more than 80% accuracy. The scenario-based design requires clear reasoning in made-up situations, which could lead to more thoughtful and consistent answers than looking back at how people felt about something. The high rankings of IT sector characteristics show that Indian IT stocks are connected to the rest of the world and that retail investors, even those who don't have formal financial education, pay attention to US macroeconomic signals. This changes how risks are talked about. For example, financial advisors might spend more time teaching investors about how the US and Indian markets are related.

V. CONCLUSION

This paper introduced the Mood Monitor Index, a survey-based investor sentiment quantification system for Indian equity markets. We demonstrated that four standard machine learning classifiers—Logistic Regression, Random Forest, Gradient Boosting, and LSTM—can reliably predict three class investor mood from 30 encoded qualitative survey features. Gradient Boosting had the highest overall accuracy (85%), while LSTM was better at using data when training for a long time. Across all models, the IT sector was always the most important area.

The MMI framework can be used in more areas, with different survey tools, and with more people. In the future, we will look into (i) using real-time survey panels, (ii) adding market microstructure signals as extra features, and (iii) using Transformer-based sequence models to predict sentiment over multiple time steps. The whole codebase is available for academic purposes.

ACKNOWLEDGEMENT

The authors are highly grateful to Dr. Satyajit Chakrabarti, Director of Institute of Engineering and Management, Kolkata for providing the project grant under the scheme of IEM-UEM grant-in-aid.

REFERENCES

[1] R. Ray, P. Khandelwal and B. Baranidharan, "A Survey on Stock Market Prediction using Artificial Intelligence Techniques," 2018 International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2018, pp. 594-598, doi: 10.1109/ICSSIT.2018.8748680.

[2] P. Vispute, J. Sujata and N. A. Natraj, "Predicting Stock Prices using Machine Learning Techniques: An Analysis of Historical Market Data," 2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bengaluru, India, 2023, pp. 725-733, doi: 10.1109/ICIMIA60377.2023.10426336.

[3] N. Sharma and A. Juneja, "Combining of random forest estimates using LSboost for stock market index prediction," 2017 2nd International Conference for Convergence in Technology (I2CT), Mumbai, India, 2017, pp. 1199-1202, doi: 10.1109/I2CT.2017.8226316.

[4] B. Jaswanth and J. Kaushik, "Stacked LSTM a Deep Learning model to predict Stock market," 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2022, pp. 17-22, doi: 10.1109/ICAC3N56670.2022.10074222.

[5] M. A. Shaik and N. L. Sri, "A Comparison of Stock Price Prediction Using Machine Learning Techniques," 2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2024, pp. 1-5, doi: 10.1109/ICESC60852.2024.10689767.

[6] B. Albaooth, "The Role of Artificial Intelligence Prediction in Stock Market Investors Decisions," 2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Nadi, Fiji, 2023, pp. 1-5, doi: 10.1109/CSDE59766.2023.10487719.

[7] D. Soni, S. Agarwal, T. Agarwel, P. Arora and K. Gupta, "Optimised Prediction Model for Stock Market Trend Analysis," *2018 Eleventh International Conference on Contemporary Computing (IC3)*, Noida, India, 2018, pp. 1-3, doi: 10.1109/IC3.2018.8530457.

[8] Srivastava, S., Pant, M. & Gupta, V. Analysis and prediction of Indian stock market: a machine-learning approach. Int J Syst Assur Eng Manag 14, 1567–1585 (2023).

[9] Chatziloizos, GM., Gunopulos, D. & Konstantinou, K. Deep Learning for Stock Market Prediction Using Sentiment and Technical Analysis. SN COMPUT. SCI. 5, 446 (2024).

[10] Agrawal, S., Kumar, N., Rathee, G. et al. Improving stock market prediction accuracy using sentiment and technical analysis. Electron Commer Res (2024).

[11] Sheth, D., Shah, M. Predicting stock market using machine learning: best and accurate way to know future stock prices. Int J Syst Assur Eng Manag 14, 1–18 (2023).

[12] Y. Ma, R. Han and X. Fu, "Stock prediction based on random forest and LSTM neural network," 2019 19th International Conference on Control, Automation and Systems (ICCAS), Jeju, Korea (South), 2019, pp. 126-130, doi: 10.23919/ICCAS47443.2019.8971687.

[13] G. Bathla, "Stock Price prediction using LSTM and SVR," 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, India, 2020, pp. 211-214, doi: 10.1109/PDGC50313.2020.9315800.

[14] S. Goswami and S. Yadav, "Stock Market Prediction Using Deep Learning LSTM Model," 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Pune, India, 2021, pp. 1-5, doi: 10.1109/SMARTGENCON51891.2021.9645837.

[15] Y. Xu and V. Keselj, "Stock Prediction using Deep Learning and Sentiment Analysis," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 5573-5580, doi: 10.1109/BigData47090.2019.9006342.

[16] T. Baihaqi, M. A. Sugiyarto, R. P. Daksa, F. I. Kurniadi, M. Fakhruddin and H. Erandi, "Unveling the Precision of Deep Learning Models for Stock Price Prediction: A Comparative Analysis of Bi-LSTM, LSTM, and GRU," 2023 International Conference on Converging Technology in Electrical and Information Engineering (ICCTEIE), Bandar Lampung, Indonesia,2023, pp.6164, doi:10.1109/ICCTEIE60099.2023.

[17] M. Al-Farouni, R. A C, S. V R, T. A. Sai Srinivas and R. Madhavi, "Stock Market Value Prediction Using Attention Based Long ShortTerm Memory Model," 2024 International Conference on Data Science and Network Security (ICDSNS), Tiptur, India, 2024, pp.1-4, doi: 10.1109/ICDSNS62112.2024.10691225.