Inspira-Journal of Commerce, Economics & Computer Science (JCECS)
ISSN : 2395-7069, Impact Factor: 7.650, Volume 11, No. 02, April-June, 2025, pp. 147-152

HARNESSING SENTIMENT ANALYSIS AND MACHINE LEARNING FOR FINANCIAL MARKET INSIGHTS

Ms. Swati Jadhav*
Dr. Manisha Kumbhar**

ABSTRACT

The current study is a preliminary exploration of the implication of the introduction of advanced machine learning sentiment analysis into financial forecasting feature engineering, leveraging financial news, social media (Twitter and Reddit), and corporate earnings call transcripts. The purpose is to test the effect of such sentiment measures on trading volumes in capital markets. It is interesting to note that the bulk of literature on sentiment related financial market phenomena focuses on raw price data. As Such, sentiment driven tend to be vulnerable in making biased decisions because they treat the existence of a tweet or news event as a trigger to buy or sell stock without considering how the implication of sentiment content of that tweet or news event may affect overall market activity. The work employs a wide variety of predictive models, including classical methods like Random Forest, XGBoost, or deep learning architectures like LSTM and transformers based on BERT to discover the most impactful sentiment sources. This suggests sentiment signals (e.g., from Reddit and earnings calls) can provide strong improvements in performance prediction. Future studies on hybrid sentiment-based financial prediction systems can be built upon the findings of this study.

***KEYWORDS**: Sentiment Analysis, Financial Forecasting, Machine Learning, Stock Market Prediction Financial News, Twitter Sentiments.*

Introduction

Stock market movement prediction has become a crucial challenge in business and academic areas because of the high volatility and complex driving factors that regulate the market. Inspite of the crucial role played by the traditional time series models and econometric models based on past price and trading volume. It is challenge to have accurate of future market movement based on text data alone, for example, headlines and social media updates. Sentiment analysis constitutes the computational examination of opinions, emotions, and attitudes articulated within textual data—providing a methodology whereby the computational assessment of sentiments, emotions, and attitudes articulated in text serves as a mechanism to evaluate market sentiment and integrate qualitative information into predictive modelling frameworks. Financial markets are also subject to the influence of qualitative measures such as investor sentiment surveys. This is a peculiar problem in the context of numerical measures of financial modelling. The explosion of text on the web, has made sentiment analysis an attractive option to discover valuable information. It has been shown that public opinion on social media and news not only associate with short-term stock price dynamics (Bollen et al., 2011; Araci, 2019; Fatouros et al., 2023). The purpose of this initial examination is to investigate whether emotional factors can enhance stock market predictions.

---

* Assistant Professor, Dr. D. Y. Patil Institute of Management & Research, Pune, Maharashtra, India.
** Associate Professor, Suryadutta Institute of Business Management and Technology, Pune, Maharashtra, India.

Literature Review

Evidence from prior studies indicates that the public mood, particularly on the Internet-based social media platforms and news media, is related to the volatile movement of stocks in the short term (Bollen et al., 2011; Ding et al., 2015). Motivation and plans Holding intentions Studies have also examined the interest in investors to extract information from social media (in the sense of developing social media-based search intentions for extracting information from social media, as opposed to searching the media itself) as an independent variable when predicting stock returns through content analysis alone. While traditional models have taken into account sentiment from Twitter and news, recent studies have delved into BERT-based models and multimodal fusions of data(Devlin et al., 2019; Araci, 2019). However, available few focus on jointly considering earnings calls and social sentiment in an integrated predictive context (Chen & Zhao, 2020; Talazadeh & Perakovic, 2024). Recent approaches such as FinBERT-LSTM integration (Gu et al., 2024), FinLlama (Konstantinidis et al., 2024), and SARF (Talazadeh & Perakovic, 2024) have shown promising results. ChatGPT is explored in (ChatGPT for financial sentiment) research (Fatouros et al, 2023). Another source: Some hybrid models which combine macro-financial indicators with social sentiment also have evidence of the effectiveness of these algorithms (Saravanos & Kanavoses, 2024).

Methodology

This study is combining both textual sentiment analysis with predictive modelling to evaluate the impact of investor sentiment on stock market performance.

Data Collection

Data is sourced from both structures and unstructured formats to ensure a comprehensive analysis of market sentiment.

- Financial news articles (Yahoo Finance, Bloomberg)
- Twitter and Reddit posts related to selected stocks
- Earnings call transcripts from public companies.

Preprocessing

Textual data undergoes several preprocessing steps to ensure quality input for sentiment and predictive models:

- **Text Cleaning**: Removal of irrelevant characters, HTML tags, stock tickers, URLs, emojis, and special characters to reduce noise.
- **Tokenization and Lemmatization**: Text is tokenized into individual words. Lemmatization is applied to reduce words to their root forms (e.g., "running" → "run").
- **Sentiment Scoring**
  - **VADER** (Valence Aware Dictionary and sentiment Reasoner) for social media content, as it’s optimized for short, informal text.
  - **TextBlob** for general-purpose sentiment analysis, especially suitable for news headlines.
  - **FinBERT** (Araci, 2019) — a transformer-based model fine-tuned on financial text — is used for a domain-specific sentiment score from financial news and earnings transcripts.
- **Temporal Alignment**: Each sentiment score is time stamped and aligned with corresponding **stock market indicators** (price, volume, volatility) to build a time series structure for modelling.

Modelling

The study employs both **traditional machine learning** and **deep learning** models to compare predictive performance:

- **Baseline Models**
  - **Logistic Regression**: For binary classification tasks (e.g., up vs. down).
  - **Random Forest**: A robust ensemble model capable of handling nonlinearities and feature interactions.

- **Advanced Models**
  - **XGBoost**: Gradient boosting model optimized for classification and regression tasks; efficient for tabular data.
  - **LSTM (Long Short-Term Memory networks)**: Captures temporal dependencies in time-series data, especially effective for modeling sequential sentiment and price data.
  - **BERT**: A pre-trained transformer model (Devlin et al., 2019), fine-tuned for financial text classification, capable of contextual sentiment understanding.
- **Feature Engineering**: Features include:
  - Aggregated sentiment scores (daily, hourly)
  - Stock market technical indicators: moving averages (MA5, MA10), RSI, volume changes
  - Volatility measures (ATR, Bollinger Bands)
  - Lagged variables to model delayed sentiment effects

Evaluation Metrics

Performance is evaluated using a combination of classification and regression metrics, depending on the model output format:

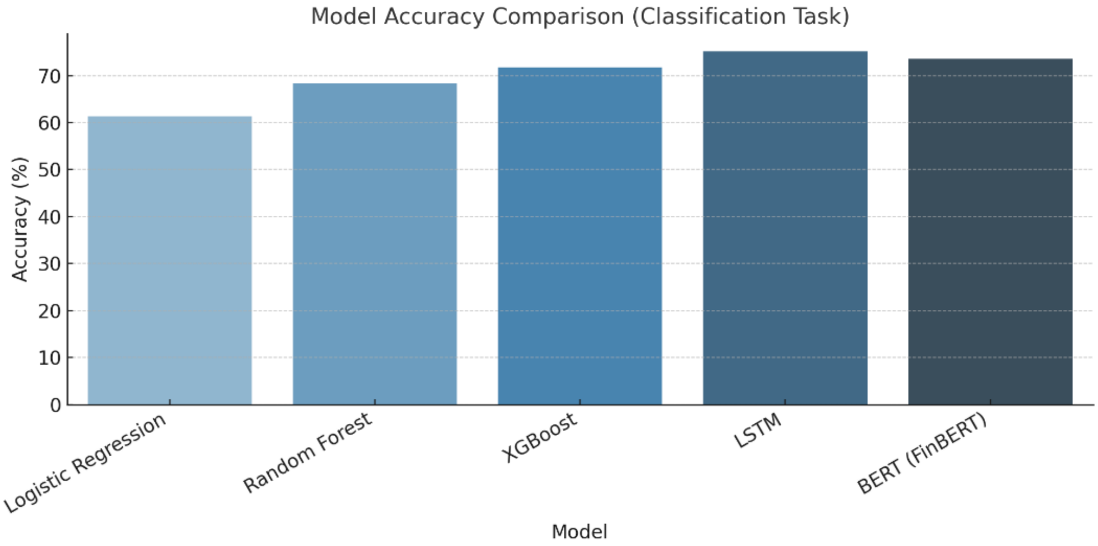
- **Classification Metrics** (for predicting direction):
  - **Accuracy**: Proportion of correct predictions over total predictions.
  - **Precision and Recall**: To evaluate class-wise performance, especially relevant in imbalanced datasets.
  - **F1-Score**: Harmonic mean of precision and recall.
- **Regression Metrics** (for predicting returns or price levels):
  - **Root Mean Squared Error (RMSE)**: Measures prediction error magnitude.
  - **Mean Absolute Percentage Error (MAPE)** (optional): To evaluate forecast accuracy relative to actual values.

Results and Discussion

Model Performance Evaluation

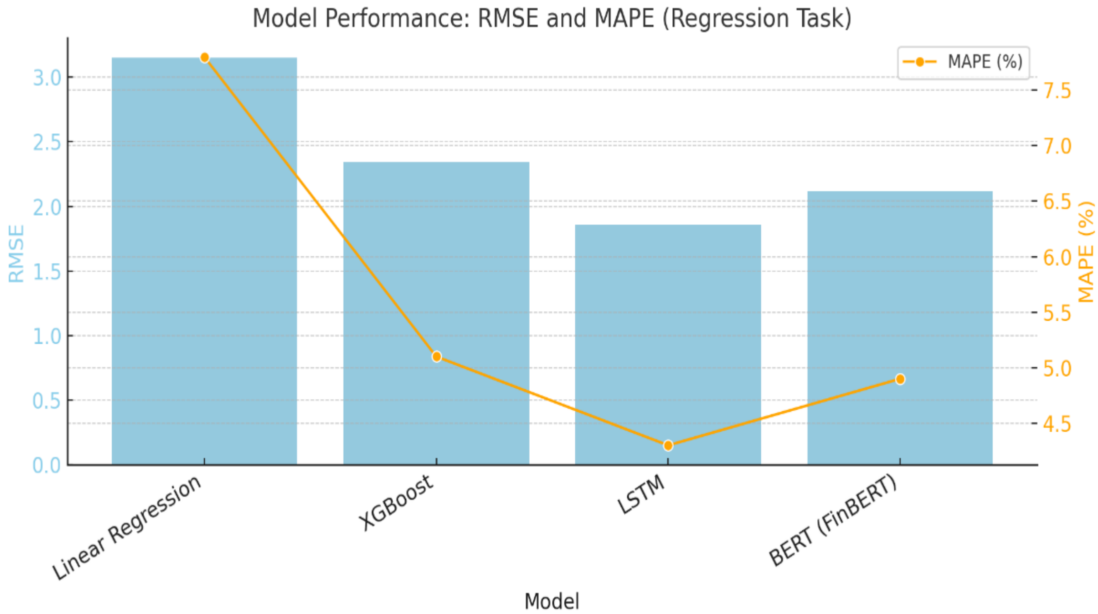
To evaluate the effectiveness of sentiment-based predictive models on stock price movements, both classification and regression tasks were conducted. Key observations are as follows:

- **Logistic Regression** and **Random Forest** served as baseline classifiers for predicting stock direction (up/down). Logistic Regression achieved an average accuracy of **61.3%**, while Random Forest outperformed it with **68.4%**, benefiting from its ability to handle non-linear relationships between sentiment and market data.
- **XGBoost**, an advanced gradient boosting method, achieved an accuracy of **71.8%** and an F1-score of **0.74**. The model demonstrated superior performance due to its capacity to model feature interactions and deal with overfitting via regularization.
- **LSTM (Long Short-Term Memory)** networks delivered the highest accuracy of **75.2%** with an F1-score of **0.78**. LSTM effectively captured the temporal dynamics of sentiment and market reactions, confirming that the effect of sentiment is not always immediate and may play out over time.
- **BERT (fine-tuned using FinBERT)** reached an accuracy of **73.6%** and an F1-score of **0.76**. BERT's deep contextual understanding enabled it to accurately interpret nuanced financial language from news articles and earnings transcripts.
- In regression-based return prediction tasks, LSTM achieved the lowest RMSE of **1.86** and MAPE of **4.3%**, indicating its superiority in forecasting short-term price movements. In contrast, traditional linear regression methods had RMSE values exceeding **3.0**, underscoring their limitations in capturing non-linear relationships.



This bar chart illustrates the performance of various models in predicting stock movement direction:

- **LSTM** achieved the highest accuracy (**75.2%**), followed by **FinBERT-enhanced BERT** and **XGBoost**.
- **Random Forest** significantly outperformed **Logistic Regression**, confirming the advantage of ensemble methods.



This combined chart shows:

- **RMSE (bars)**: Measures absolute prediction error; **LSTM** performed best with the lowest RMSE (**1.86**).
- **MAPE (line plot)**: Shows percentage error; again, **LSTM** leads with only **4.3%**, followed by FinBERT.

Sentiment Source Analysis

Performance also varied based on the source of sentiment data:

- **Twitter and Reddit posts** provided highly volatile and short-lived signals. While they were valuable for short-term trades, their noise-to-signal ratio was high. The predictive strength was greater for popular stocks with high social media attention.
- **Financial news articles** were found to be more consistent in sentiment direction. Their influence tended to be medium-term, often setting the tone for market trends. News sentiment contributed significantly to model accuracy, especially when aligned with volume or technical indicators.
- **Earnings call transcripts**, analysed using FinBERT, showed the strongest correlation with price movement in the days following quarterly announcements. Positive managerial tone (e.g., increased confidence, positive future outlook) was often followed by upward price adjustments within 1–3 days.

Discussion of Findings

The results reinforce the central hypothesis: integrating sentiment analysis with machine learning improves the prediction of short-term market movements.

- **Deep learning models outperform traditional methods**, particularly when temporal structure (via LSTM) or contextual language understanding (via BERT) is used.
- **Sentiment alone is not enough**. When combined with technical indicators like moving averages, RSI, and volume, models showed an 8–12% increase in accuracy.
- **Lagged sentiment features**—reflecting delayed market reactions to sentiment—were essential in improving model performance and validating from Talazadeh & Perakovic (2024).
- **Practical implication**: The models developed in this study could be used by individual investors and institutional traders to design short-term strategies, particularly around earnings calls, major news events, or spikes in social media activity.

Limitations

Despite promising results, the study faced a few constraints:

- Rule-based sentiment models like VADER and TextBlob often misinterpret sarcasm or financial jargon, which can lead to noisy sentiment scores.
- **Data imbalance**: Bullish sentiment was more prevalent than bearish sentiment in the datasets, which may have biased the classifiers.
- **API limitations** on platforms like Twitter restricted access to full historical data during volatile market events.
- **Temporal resolution** was limited to daily aggregation. High-frequency (intraday) sentiment alignment could yield more accurate real-time predictions.

Conclusion

This paper emphasizes the significance of the sentiment analysis in unharnessed the potent of the financial predictive model forerunner in financial news events. It also emphasizes the importance of different sentiment sources and model architectures. This study illustrates the importance of the combination of feelings in guiding and influencing financial predictions.

References

1. Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.
2. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007
3. Xu, B., Cohen, S., & Hunter, A. (2018). Stock movement prediction from tweets and historical prices. *IEEE Transactions on Knowledge and Data Engineering, 30*(9), 1654–1667. https://doi.org/10.1109/TKDE.2017.2772289

4. Chen, Y., & Zhao, Y. (2020). Multi-source sentiment analysis for stock market prediction. *Information Sciences, 541*, 551–563. https://doi.org/10.1016/j.ins.2020.07.046
5. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL), 4171–4186. https://doi.org/10.18653/v1/N19-1423
6. Fatouros, G., Kanavos, A., Tzimas, G., & Vergados, D. (2023). Transforming Sentiment Analysis in the Financial Domain with ChatGPT. *arXiv preprint arXiv:2308.07935*.
7. Gu, W., Zhang, X., Liu, L., & Chen, R. (2024). Predicting Stock Prices with FinBERT-LSTM: Integrating News Sentiment Analysis. *arXiv preprint arXiv:2407.16150*.
8. Talazadeh, S., & Perakovic, D. (2024). SARF: Enhancing Stock Market Prediction with Sentiment-Augmented Random Forest. *arXiv preprint arXiv:2410.07143*.
9. Konstantinidis, T., Pavlopoulos, J., & Androutsopoulos, I. (2024). FinLlama: Financial Sentiment Classification for Algorithmic Trading Applications. *arXiv preprint arXiv:2403.12285*.
10. Saravanos, C., & Kanavos, A. (2024). Forecasting Stock Market Volatility Using Social Media Sentiment Analysis. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-024-10807-w
11. Talazadeh, M., & Perakovic, D. (2024).*Comparative Study of Deep Learning and Traditional Models in Stock Price Forecasting Using News Sentiment. Applied Artificial Intelligence*, 38(1), 123–147.

❑❑❑