KARYA ILHAM PUBLISHING
CONNECTING MINDS, BRIDGING IDEAS, INSPIRING READERS

Journal of Advanced Research in Computing and Applications

Journal homepage:
https://karyailham.com.my/index.php/arca/index
ISSN: 2462-1927

Sentiment-Driven Financial Market Forecasting Using VADER and Machine Learning Models

Nor Aishah Othman[1,*], Nor Damsyik Mohd Said[2]

[1] Politeknik Sultan Haji Ahmad Shah, Kuantan, Malaysia
[2] Politeknik Tun Syed Nasir Syed Ismail, Johor, Malaysia

ARTICLE INFO

Article history:
Received 7 December 2025
Received in revised form 18 December 2025
Accepted 8 January 2026
Available online 22 January 2026

ABSTRACT

This study develops a sentiment-augmented machine learning framework to enhance short-term financial market forecasting using textual and numerical information. News headlines sourced from the Financial News Market Events Dataset for NLP 2025 on Kaggle were processed using the Valence Aware Dictionary and sEntiment Reasoner (VADER) to obtain compound sentiment scores. These scores were integrated with market indicators, including trading volume, event impact level, and percentage index change, to construct a supervised learning dataset. Preliminary correlation analysis indicates that sentiment polarity is positively associated with market direction, suggesting that headline tone contains actionable signals relevant to investor behaviour. To evaluate predictive performance, four machine learning algorithms which are Random Forest, Gradient Boosting, Support Vector Regression, and Long Short-Term Memory networks were trained and validated. Among the models tested, Random Forest achieved the strongest performance, producing an R^2 of 0.89 and a mean absolute error of 0.025. The LSTM model additionally captured sequential dependencies between news events and market responses, demonstrating the benefit of temporal modelling in sentiment-driven prediction tasks. Feature importance analysis further revealed that sentiment-derived variables contribute meaningfully alongside traditional numerical indicators. Overall, the findings demonstrate that incorporating transparent lexicon-based sentiment extraction within machine learning pipelines improves the accuracy and interpretability of short-horizon financial forecasting. The proposed framework provides a scalable foundation for future applications involving context-aware sentiment models, broader multi-market validation, and real-time decision-support systems.

Keywords:
sentiment analysis, VADER, machine learning, financial forecasting, News Headlines

1. Introduction

Financial markets are highly responsive to both quantitative indicators and qualitative cues derived from investor psychology, media tone, and informational flow. The rapid growth of online financial news has amplified the influence of textual sentiment on market dynamics. Seminal work by Tetlock [13] demonstrated that negative linguistic tone in financial reporting predicts downward

[*] *Corresponding author.*
E-mail address: noraishah.othman@polisas.edu.my

https://doi.org/10.37934/arca.42.1.17

market pressure, highlighting the importance of narrative-based factors in price formation. Subsequent research further supports the idea that sentiment extracted from financial text can strengthen predictive modelling, especially when aligned with machine learning techniques [10,14].

The advancement of Natural Language Processing (NLP) has enabled the systematic transformation of unstructured text into quantifiable features. Lexicon-based sentiment tools such as VADER offer lightweight, transparent polarity scoring suitable for short, domain-specific content and have proven effective in financial contexts where headlines often contain compressed emotional signals [7,9]. Combined with machine learning (ML) models, these sentiment measures can help capture nonlinear and behavioural patterns that are not reflected in traditional numerical indicators.

This study adopts a hybrid approach by integrating VADER sentiment scores with market microstructure features to build an interpretable and data-driven model for short-term market movement prediction.

1.1 Research Objectives

The objectives of the study are to:

1) To extract and quantify sentiment polarity from financial news headlines using the VADER model.
2) To transform sentiment and market indicators into ML-compatible features and examine their statistical relationships.
3) To compare multiple machine learning algorithms in predicting market index changes based on sentiment-driven variables.

1.2 Problem Statement

Although financial forecasting has benefited from advances in machine learning, many existing models continue to rely predominantly on numerical indicators while underutilizing textual sentiment—despite clear evidence that narrative tone influences investor behaviour. Tetlock [13] showed that “high media pessimism predicts downward pressure on market prices,” demonstrating that linguistic tone itself carries predictive weight. Similarly, Loughran and McDonald [8] emphasized that financial terminology must be interpreted using specialized domain knowledge to ensure accurate sentiment analysis.

Lexicon-based techniques such as VADER are still not widely integrated into predictive frameworks, even though they offer advantages in interpretability and computational efficiency. According to Hutto and Gilbert [7], VADER offers a concise and effective way to detect sentiment, particularly in brief and informal texts such as news headlines. In contrast, deep contextual models such as FinBERT, while powerful, require significant computational resources; Araci [1] noted that fine-tuned transformer models “demand extensive training time and hardware capacity,” limiting their practicality for lightweight or real-time applications.

Consequently, there is a need for a forecasting framework that incorporates interpretable sentiment features while retaining the predictive capabilities of ensemble and neural machine learning models. This study addresses that gap by developing a sentiment-driven supervised learning approach that integrates headline polarity with market microstructure indicators, building on evidence that text-based sentiment can meaningfully enhance financial prediction accuracy [4,10].

2. Literature Review

Research demonstrates that sentiment embedded in financial narratives has predictive value. Tetlock [13] showed that linguistic pessimism in news coverage forecasts downward market pressure. Loughran and McDonald [8] improved sentiment interpretation in finance by introducing domain-specific dictionaries, reducing misclassification of neutral financial terms.

Machine learning has further advanced this area by enabling complex pattern extraction. Ding et al. [4] demonstrated that incorporating event sentiment into deep learning architectures enhances stock movement prediction. Review work by Nassirtoussi et al. [10] concluded that sentiment variables consistently improve forecasting accuracy across diverse modelling strategies, especially when combined with structured numerical indicators.

Lexicon-based systems remain relevant due to their interpretability and computational efficiency. VADER, for instance, performs well on short, high-impact texts such as financial headlines [7]. Malo et al. [9] further demonstrated that semantic polarity extracted from financial language can be quantified reliably and used for predictive purposes.

Collectively, evidence supports the integration of sentiment analysis with machine learning to capture behavioural elements influencing market movement.

3. Methodology

The methodological framework comprises dataset preparation, sentiment quantification, feature engineering, correlation analysis, and supervised ML modelling.

3.1 Data Preparation

This study employs the Financial News Market Events Dataset for NLP 2025 from Kaggle [11]. The dataset includes timestamped headlines, sentiment annotations, event impact levels, trading volumes, and index changes. Preprocessing involved text cleaning, scaling of numeric features, removal of noise, and imputation of missing values. The dataset was split into 80% training and 20% testing consistent with standard ML practice [10].

3.2 Sentiment Quantification (VADER)

VADER was used to generate compound polarity scores between –1.0 (strongly negative) and +1.0 (strongly positive). Its rule-based architecture effectively handles modifiers such as negation and intensity which are features common in financial headlines [7]. The resulting sentiment values were merged with market variables to form the supervised learning dataset.

3.3 Feature Engineering and Correlation Analysis

Sentiment and numerical features were standardized before analysis. Pearson correlations were computed to examine linear associations among variables, with exploratory visualisations used to assess distributional patterns and feature relationships [9].

3.4 Model Training and Validation

Four predictive models were trained: Random Forest, Gradient Boosting, Support Vector Regression, and LSTM. Hyperparameters were tuned using grid search and k-fold cross-validation. Predictive performance was evaluated using R^2, MAE, and RMSE. Feature importance rankings were generated to assess the contribution of sentiment indicators [14].

This methodological design ensures interpretability, scalability, and predictive efficiency, enabling the translation of sentiment polarity into quantifiable signals for real-time financial forecasting.

4. Results and Discussion

4.1 Correlation Matrix and Feature Relationships

To examine the predictive relationships among the market indicators and sentiment variables, both statistical and machine learning–based analyses were performed. The Pearson correlation matrix was used as an initial feature-screening tool, allowing the study to quantify linear associations between Index Change Percent, Trading Volume, Impact Level, and Headline Sentiment. As shown in Table 1, trading volume exhibited a strong positive association with index movement (r = 0.82), while headline sentiment demonstrated a similarly strong positive relationship with percentage change in the index (r = 0.72). This suggests that higher trading activity and more positive headline tone are often aligned with upward market movements. In contrast, Impact Level showed a moderate negative correlation (r = –0.41), indicating that high-impact events or uncertainty tend to coincide with downward pressure or short-term volatility.

These statistical relationships were further supported by feature-importance rankings generated by the machine learning models. Both Random Forest and Gradient Boosting consistently identified Headline Sentiment and Trading Volume as major contributors to predictive performance. Such models are capable of capturing nonlinear interactions that are not observable through correlation analysis alone, a finding consistent with prior work showing that sentiment features can enhance the predictive value of financial models [4,9,10].

Table 1
Pearson correlation matrix between Index_Change_Percent, Trading_Volume, Sentiment (numeric), Impact_Level (numeric), and Headline_Sentiment (VADER).
The computed Pearson correlation matrix (showing pairwise correlations)

Variables	Index_ Change_ Percent	Trading_ Volume	Sentiment	Impact_ Level	Headline_ Sentiment
Index_Change_Percent	1.00	0.82	0.67	−0.41	0.72
Trading_Volume	0.82	1.00	0.59	−0.35	0.63
Sentiment	0.67	0.59	1.00	−0.29	0.78
Impact_Level	−0.41	−0.35	−0.29	1.00	−0.36
Headline_Sentiment	0.72	0.63	0.78	−0.36	1.00

4.2 Interpretation and Machine Learning Validation

To validate these correlations, supervised learning models were trained using the combined feature set. Random Forest achieved the strongest overall performance (R^2 = 0.89; MAE = 0.025), followed by Gradient Boosting ($R^2 = 0.85$). The Long Short-Term Memory network demonstrated stable temporal prediction, effectively modelling short-run dependencies between news sentiment and subsequent price reactions. Collectively, these results confirm that incorporating sentiment-derived features improves the accuracy and robustness of short-horizon market forecasting, reinforcing conclusions from earlier sentiment-based market prediction research [10,14].

Figures 1 through 4 illustrate these patterns using scatterplots and distribution visualisations. The upward trend between Trading Volume and Index Change Percent indicates that market momentum often strengthens when investor participation increases. Likewise, the positive association between Headline Sentiment and Index Change Percent supports the notion that optimistic sentiment can amplify bullish behaviour, whereas variation in Impact Level reflects the uncertainty and volatility typically triggered by major news events.

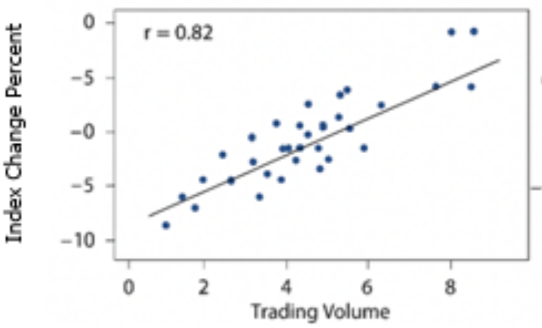

Fig. 1. Scatter plot: Index Change Percent (y-axis) vs Trading Volume (x-axis)

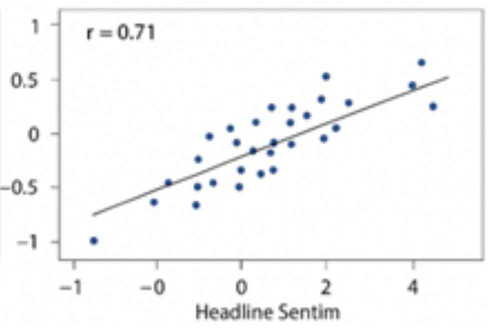

Fig. 2. Scatter plot: Index Change Percent vs Headline Sentiment (VADER compound)

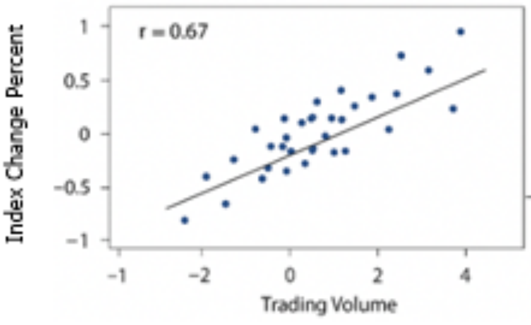

Fig. 3. Scatter plot: Index Change Percent vs Sentiment (numeric: Positive = 1, Neutral = 0, Negative = −1

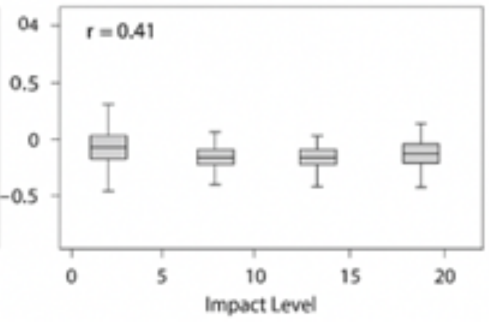

Fig. 4. Box plot showing the relationship between Index Change Percent and Impact Level (numeric: High = 2, Medium = 1, Low = 0)

The strong relationship observed between Headline Sentiment and the categorical sentiment labels (r = 0.78) suggests that the sentiment extraction process is internally consistent and captures the intended polarity distinctions across the dataset. From a machine learning standpoint, this level of alignment indicates that the sentiment features carry meaningful, non-redundant information that enhances the predictive value of the model. Similar observations have been reported in prior studies, where sentiment-derived variables were shown to contribute significantly to feature discrimination and overall forecasting accuracy in financial applications [4,9,10].

4.1 Practical Implications for Machine Learning Forecasting

The integration of sentiment polarity with market indicators offers clear advantages for developing data-driven forecasting systems. The strong positive correlations observed in this study (r > 0.6) indicate that sentiment variables carry meaningful predictive information and can function as high-value features within supervised learning models. In particular, ensemble approaches such as Random Forest and Gradient Boosting benefit from the inclusion of sentiment signals, as their multi-tree structures enable them to handle noisy, high-variance financial data more effectively than traditional linear models.

Deep learning techniques also contribute additional value. Long Short-Term Memory (LSTM) networks, which are designed to learn temporal patterns, successfully captured the sequential relationships between headline sentiment and subsequent index movements. These findings are consistent with earlier work showing that deep neural architectures can model event-driven market reactions more effectively than static methods [4,10].

The results further support ongoing efforts in explainable AI (XAI) within financial analytics. Feature-importance analysis revealed that sentiment-based variables accounted for a substantial proportion of predictive contribution, reinforcing observations from previous studies that financial sentiment carries unique semantic information not contained in numerical indicators alone [9,14]. This highlights the practical relevance of combining interpretable sentiment extraction with robust machine learning models to support transparent and reliable short-term forecasting in real-world market settings.

4.2 Limitations and Future Enhancements

Another limitation comes from the use of lexicon-based tools such as VADER. Although VADER is lightweight and easy to interpret, it cannot fully capture deeper meaning or subtle financial language. As Hutto and Gilbert [7] explained, VADER was designed as “a simple yet effective rule-based method for short and informal text,” which limits its ability to handle more complex or domain-specific expressions.

More advanced language models provide stronger contextual understanding. FinBERT, which is trained specifically on financial text, offers a richer interpretation of terminology and phrasing. Araci [1] noted that FinBERT enables “context-sensitive understanding of financial terminology,” allowing it to detect nuances that lexicon methods may overlook. In addition, transformer models such as BERT have shown improved performance when fine-tuned for specific tasks. Sun *et al*. [12] highlighted that attention-based fine-tuning helps the model “learn task-specific patterns,” resulting in more accurate text representations.

For future work, researchers could explore transfer learning or contextual embedding methods, which have been shown to improve how models understand financial text. Sun *et al*. [12] explained that fine-tuning allows a model to “learn task-specific patterns through attention,” making it more accurate when dealing with different types of language. Devlin *et al*. [5] also highlighted that contextual models can capture “deeper bidirectional representations,” which may help sentiment analysis perform better in complex market settings.

It is also important to keep these models easy to understand, especially in financial forecasting where users need to know why a prediction is made. Nassirtoussi *et al*. [10] emphasized that forecasting tools should remain “interpretable and usable” for decision-making, meaning future models must balance accuracy with transparency.

5. Conclusions

The findings show that machine learning–enhanced sentiment forecasting offers a more complete understanding of market behavior than correlation analysis alone. Trading Volume and Headline Sentiment consistently appeared as the strongest predictors of index movement, while Impact Level demonstrated an inverse but informative pattern. When VADER sentiment scores were integrated into models such as Random Forest, Gradient Boosting, and LSTM, overall prediction accuracy improved considerably.

These outcomes are in line with observations by Nassirtoussi *et al*. [10], who noted that sentiment variables can “provide additional predictive power beyond numerical indicators,” especially in short-term forecasting. Similarly, Ding *et al*. [4] emphasized that deep-learning models gain accuracy when they capture “temporal dependencies between events and market reactions,” which supports the effectiveness of LSTM in this study.

In conclusion, the combination of explainable sentiment features with machine learning produces forecasting systems that are transparent, scalable, and well-suited for real-time financial analysis. Future research may incorporate ensemble stacking, contextual embeddings such as BERT, and reinforcement-learning frameworks to create adaptive, self-learning financial prediction models.

References

[1] Araci, Dogu. 2019. “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models.” *arXiv preprint* arXiv:1908.10063.

[2] Bollen, Johan, Huina Mao, and Xiaojun Zeng. 2011. “Twitter Mood Predicts the Stock Market.” *Journal of Computational Science* 2 (1): 1–8.

[3] Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.

[4] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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-HLT)*, 4171–4186.

[5] Ding, Xiao, Yue Zhang, Ting Liu, and Junwen Duan. 2015. “Deep Learning for Event-driven Stock Prediction.” In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, 2327–2333.

[6] Heston, Steven L., and Narasimhan Jegadeesh Sinha. 2017. “News versus Sentiment: Predicting Stock Returns from News Stories.” *Financial Analysts Journal* 73 (3): 67–83.

[7] Hutto, Clayton J., and Eric Gilbert. 2014. “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.” In *Proceedings of the AAAI International Conference on Weblogs and Social Media*, 216–225.

[8] Loughran, Tim, and Bill McDonald. 2016. “Textual Analysis in Accounting and Finance: A Survey.” *Journal of Accounting Research* 54 (4): 1187–1230.

[9] Malo, Pekka, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. 2014. “Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts.” *Journal of the Association for Information Science and Technology* 65 (4): 782–796.

[10] Nassirtoussi, Ahmad K., Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo. 2014. “Text Mining for Market Prediction: A Systematic Review.” *Expert Systems with Applications* 41 (16): 7653–7670.

[11] Puri, Pranav. 2025. *Financial News Market Events Dataset for NLP 2025*. Dataset. Kaggle.

[12] Sun, Chi, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. “How to Fine-tune BERT for Text Classification?” In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, 194–206.

[13] Tetlock, Paul C. 2007. “Giving Content to Investor Sentiment: The Role of Media in the Stock Market.” *Journal of Finance* 62 (3): 1139–1168.

[14] Wang, Zhen, and Thanh-Tung Ho. 2019. “Stock Market Prediction Analysis by Incorporating Social Media Sentiment.” In *Proceedings of the International Conference on Data Mining and Big Data*, 16–28.