International Journal of Basic and Applied Science 15 (2) (2026) 86-94

Published by: IOCSCIENCE

International Journal of Basic and Applied Science

Journal homepage: www.ijobas.pelnus.ac.id

Financial News Sentiment Classification: A Comparative Evaluation of Lexicon-Based and Machine Learning Approaches

Hasanul Fahmi[1], Noor Azma Ismail[1] Fazidah Wahit[2], Amalia Amalia[3]

[1] Faculty of Artificial Intelligence and Frontier Technology, UNITAR International University, Selangor, Malaysia

[2] Center For Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia

[3] Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Sumatera Utara, Medan, Indonesia

Article Info

Article history:

Received Feb 09, 2026
Revised Apr 17, 2026
Accepted Sep 03, 2026

***Keywords*:**

Financial news;
Financial PhraseBank;
Machine learning;
Sentiment classification;
Support Vector Machine.

ABSTRACT

Financial news sentiment classification can support the extraction of market-relevant information from unstructured financial text. This study compares a general-purpose lexicon-based method, Valence Aware Dictionary and Sentiment Reasoner (VADER), with three supervised classifiers: Random Forest, Multinomial Naive Bayes, and linear Support Vector Machine (SVM). The experiments use 4,846 labelled statements from the Financial PhraseBank dataset to classify sentiment into positive, neutral, and negative categories. The supervised classifiers use term frequency–inverse document frequency (TF-IDF) features and an 80:20 train-test split, while VADER serves as a rule-based baseline. Based on the reported evaluation, SVM achieves the highest accuracy (0.76) and macro F1-score (0.69), followed by Random Forest, with an accuracy of 0.75 and a macro F1-score of 0.66. VADER yields the lowest reported performance, with an accuracy of 0.55 and a macro F1-score of 0.48. The confusion matrix indicates that positive and negative statements are frequently misclassified as neutral. These findings suggest that supervised classifiers provide stronger sentiment-classification performance than the general-purpose lexicon under the evaluated conditions. However, the results do not establish direct improvements in financial-risk forecasting, and further validation is required to assess generalisability across datasets and market conditions.

This is an open access article under the CC BY-NC license.

Corresponding Author:

Hasanul Fahmi Zuhri,
School of Information Technology,
UNITAR International University,
UNITAR International University Tierra Crest, Jalan SS 6/3, Ss 6, 47301 Petaling Jaya, Selangor, Malaysia
Email: fahmi.zuhri@unitar.my

1. INTRODUCTION

Financial markets respond rapidly to information about earnings, monetary policy, inflation, corporate actions, and geopolitical events. Conventional risk analysis relies mainly on structured variables such as prices, ratios, indices, and macroeconomic indicators. Although these variables remain essential, they do not fully represent the real-time perceptions and behavioural responses expressed in unstructured financial text[1][2].

Natural language processing enables financial news to be converted into sentiment indicators that can support portfolio monitoring, early-warning analysis, and decision-making under uncertainty.

However, financial language is cautious, technical, and domain dependent; expressions that appear neutral in general communication may convey positive or negative implications in a financial context. Recent reviews and economic-news studies therefore emphasise contextual interpretation, domain adaptation, and careful treatment of negation [3]–[5]

Financial sentiment resources have developed from labelled datasets and manually designed dictionaries to entity-aware corpora, automatically generated lexicons, and supervised classification models [6]–[8]. Comparative evidence also shows that machine-learning methods can capture disclosure sentiment more reliably than fixed dictionaries when sufficient labelled data are available [9]. These findings suggest that a general-purpose lexicon should be treated as a transparent baseline rather than assumed to be optimal for specialised financial text.

Domain-specific transformers and large language models have recently improved financial sentiment classification and broadened its use in market analysis[10]–[16]. Related studies have connected textual sentiment with stock reactions, forecasting, sector analysis, and systemic-risk assessment [17]–[23]. Nevertheless, transparent comparisons among computationally lightweight methods remain useful for institutions that require interpretable baselines, modest computing requirements, and reproducible implementation. Despite advances in financial sentiment analysis, a practical evaluation gap remains in the comparison of general-purpose lexicon-based methods and conventional supervised classifiers under a consistent experimental protocol. Differences in datasets, preprocessing procedures, feature representations, and evaluation metrics can make reported performance difficult to compare across studies. Moreover, aggregate accuracy alone may obscure weaknesses in detecting negative financial statements, which can be particularly relevant to risk-monitoring applications. This study addresses this evaluation gap by comparing VADER, Random Forest, Multinomial Naive Bayes, and linear SVM using a common dataset, a shared train-test partition for supervised models, and class-sensitive evaluation metrics. The study focuses on comparative classification performance and error patterns rather than claiming direct improvements in financial-risk prediction.

Accordingly, this study evaluates four sentiment-classification approaches VADER, Random Forest, Multinomial Naive Bayes, and linear SVM using the Financial PhraseBank dataset and a consistent evaluation framework. The study has three objectives: (1) to compare the classification performance of a general-purpose lexicon-based method and three supervised classifiers; (2) to examine class-level performance and identify recurring sentiment-confusion patterns; and (3) to discuss the implications and limitations of these errors for potential financial-news monitoring applications. The contribution lies in a controlled comparison of conventional methods and an explicit analysis of minority-class detection under an imbalanced class distribution. The findings are intended to establish a reproducible baseline for subsequent studies, rather than to demonstrate a direct improvement in market-risk forecasting[24].

2. RESEARCH METHOD

2.1 Dataset and Research Design

A quantitative comparative experimental design was used to evaluate one lexicon-based method and three supervised classifiers for three-class financial news sentiment classification. The experiment used the Financial PhraseBank dataset, originally introduced by Malo et al.[25] and distributed through Kaggle as "Sentiment Analysis for Financial News." The working dataset contained two fields: a financial statement or headline and its positive, neutral, or negative label. The dataset source and experimental setting are summarised in Table 1.

The dataset consisted of 4,846 labelled financial news statements obtained from the Financial PhraseBank dataset, covering three sentiment categories: neutral (2,879 statements), positive (1,363 statements), and negative (604 statements). The dataset was divided into training and test partitions using an 80:20 split, resulting in 3,876 training statements and 970 test statements. The analysis focused on individual financial statements rather than complete news articles. Because the neutral class represented the largest proportion of the dataset, macro-averaged precision, recall, and F1-score

were used alongside accuracy to provide a more balanced evaluation across the three sentiment categories. For the supervised classifiers, the TF-IDF vectoriser was fitted exclusively on the training data and subsequently applied to the test data to prevent information leakage during feature construction. The same train-test partition was maintained across the supervised classifiers to ensure a consistent basis for comparative evaluation. The experiment compared Random Forest, Multinomial Naive Bayes, and linear Support Vector Machine, while VADER was evaluated separately as a rule-based sentiment classification baseline using its standard compound-score thresholds.

Table 1. Dataset and experimental setting

Item	Description
Dataset	Financial PhraseBank public financial news sentiment dataset [25]
Dataset source	Kaggle: Sentiment Analysis for Financial News (https://www.kaggle.com/datasets/ankurzing/sentiment-analysis-for-financial-news); original Financial PhraseBank dataset [25]
Number of records	4,846 labelled financial news statements
Class distribution	Neutral = 2,879; Positive = 1,363; Negative = 604
Sentiment labels	Positive, neutral, and negative
Input field	Financial news statement or headline
Feature representation	TF-IDF fitted only on the training data
Train-test split	80:20 fixed split; test set = 970 statements
Compared methods	VADER, Random Forest, Multinomial Naive Bayes, and linear SVM

2.2 Text Preprocessing and Feature Representation

For the supervised models, the text was lowercased, normalised, tokenised, and processed to remove punctuation and stopwords. A term frequency-inverse document frequency (TF-IDF) vectoriser was fitted only on the training partition and then applied to the test partition, preventing test information from influencing the feature space. The resulting sparse vectors were used by Random Forest, Multinomial Naive Bayes, and Support Vector Machine. For reproducibility, the preprocessing pipeline should specify the tokenisation method, stopword list, punctuation-handling procedure, and treatment of numerical expressions and financial symbols. The TF-IDF vectoriser configuration should also report the n-gram range, minimum and maximum document-frequency thresholds, feature limit, and normalisation setting. These parameters must reflect the actual implementation. The vectoriser was fitted on the training partition only, and the resulting vocabulary and inverse-document-frequency weights were then applied unchanged to the test partition. This procedure was used to prevent information leakage from the test data into feature construction.

The same train-test split and TF-IDF representation were retained for all supervised models to support a fair comparison. VADER was evaluated independently as a rule-based baseline and converted its compound scores into positive, neutral, or negative labels using the standard thresholds defined for the method [26]. All stochastic model operations used the same fixed seed. This design isolated differences among the classification approaches rather than differences in data partitioning.

2.3 Classification Models

Four approaches were evaluated. VADER represented a transparent general-purpose lexicon baseline. Multinomial Naive Bayes represented a lightweight probabilistic text classifier, Random Forest represented an ensemble of decision trees, and linear Support Vector Machine represented a margin-based classifier suited to high-dimensional sparse TF-IDF features. The implementation used Python with pandas, NLTK, and scikit-learn.

The comparison did not assume that one model would be universally superior. Instead, it examined which approach was most suitable for this dataset, class distribution, feature representation, and risk-analysis objective. Table 2 reports the configurations documented in the experiment.

Table 2. Model configuration

Model	Configuration used in this study
VADER	Rule-based compound score mapped to positive, neutral, or negative labels
Random Forest	100 estimators, bootstrap sampling, square-root feature selection, and a fixed seed
Naive Bayes	Multinomial Naive Bayes with additive smoothing
SVM	Linear-kernel Support Vector Machine with C = 1.0

All supervised classifiers were evaluated using the same training and test partitions. Random Forest was configured with 100 decision trees, bootstrap sampling, square-root feature selection, and a fixed random seed. Multinomial Naive Bayes used additive smoothing, with the smoothing parameter reported as [insert actual value]. Linear SVM used a linear kernel with C = 1.0. The TF-IDF representation and all preprocessing settings were kept consistent across the supervised classifiers. Any hyperparameter tuning, class weighting, or resampling procedure should be explicitly described; if none was performed, this should be stated. Software versions and the implementation details necessary to reproduce each model should also be reported.

2.4 Evaluation Metrics

Performance was evaluated using accuracy, macro-averaged precision, macro-averaged recall, and macro-averaged F1-score. Accuracy measured the proportion of correctly classified test statements. Precision measured the reliability of predicted labels, recall measured the proportion of each class detected, and F1-score balanced precision and recall [27].

Macro-averaging was emphasised because it gives equal weight to positive, neutral, and negative classes, preventing strong performance on the dominant neutral class from concealing weak minority-class detection. Confusion-matrix analysis was used to identify whether errors involved positive-neutral, negative-neutral, or positive-negative confusion.

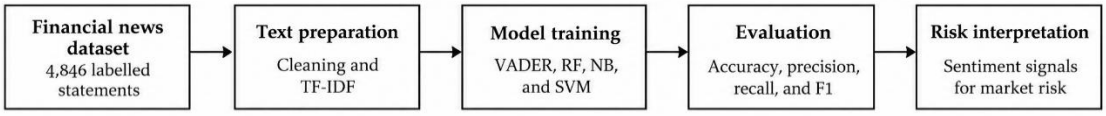

Figure 1. Research workflow for financial news sentiment classification.

Let C denote the set of sentiment classes and $K = |C| = 3$. For each class c, precision, recall, and F1-score were calculated using the corresponding true-positive, false-positive, and false-negative counts. Macro-averaged precision, recall, and F1-score were obtained by calculating the metric separately for each class and then taking the unweighted arithmetic mean across the three classes. Thus, each sentiment class contributed equally to the macro-averaged metrics regardless of its frequency. Accuracy was calculated as the total number of correct predictions divided by the total number of test instances. All reported metrics should be computed from the same test predictions and verified against the confusion matrix.

3. RESULTS AND DISCUSSIONS

3.1 Comparative Model Performance

Table 3 compares the four evaluated approaches. Support Vector Machine achieved the strongest overall performance, with 0.76 accuracy and a 0.69 macro F1-score. Random Forest followed with 0.75 accuracy and a 0.66 macro F1-score. Naive Bayes produced the highest precision value, 0.78, but its

recall and F1-score were substantially lower. VADER recorded the weakest results, with 0.55 accuracy and a 0.48 macro F1-score.

Considering accuracy and macro F1-score together, the ranking was Support Vector Machine, Random Forest, Naive Bayes, and VADER. The advantage of the supervised models indicates that patterns learned from labelled financial statements represented domain-specific sentiment more effectively than the fixed general-purpose lexicon. However, the moderate macro scores also show that minority-class detection remained challenging.

Table 3. Comparative evaluation of the sentiment-classification methods

Model	Accuracy	Precision	Recall	F1-score
VADER	0.55	0.53	0.49	0.48
Random Forest	0.75	0.77	0.62	0.66
SVM	0.76	0.76	0.66	0.69
Naive Bayes	0.68	0.78	0.45	0.43

The comparative results should be interpreted using both aggregate and class-sensitive metrics. Accuracy indicates the overall proportion of correctly classified statements, whereas macro-averaged precision, recall, and F1-score assign equal weight to the positive, neutral, and negative classes. This distinction is important because the dataset contains substantially more neutral statements than negative statements. Accordingly, model selection should not rely on accuracy alone. The class-wise results and confusion matrices provide additional evidence of which sentiment categories are detected reliably and which remain difficult to distinguish.

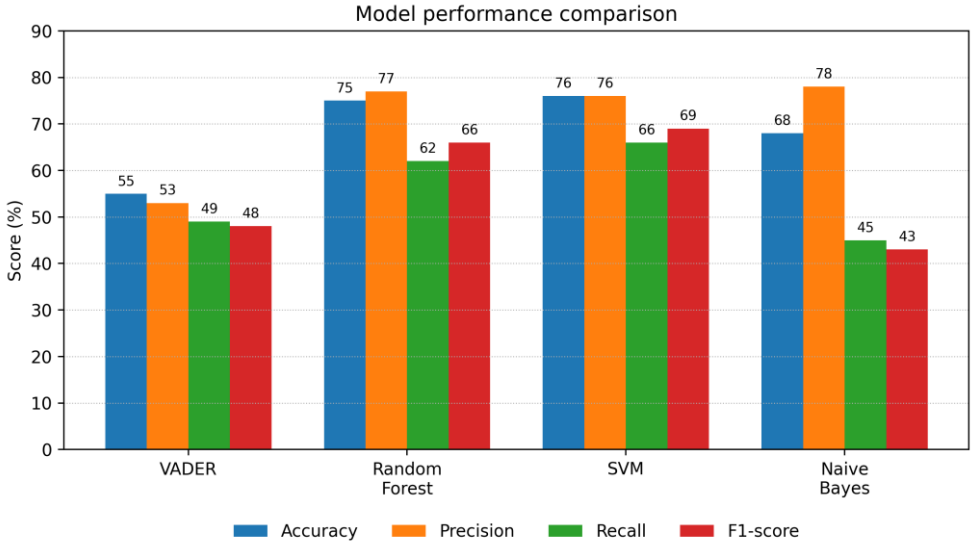

Figure 2. Performance comparison of VADER, Random Forest, SVM, and Naive Bayes.

3.2 Class-wise Interpretation

Neutral sentiment was easier to classify than positive or negative sentiment. Financial news often uses formal and descriptive language, making neutral statements more internally consistent and easier for models to learn. The larger number of neutral training examples also increased the opportunity for the supervised models to learn neutral patterns.

Positive and negative statements were frequently misclassified as neutral. This error is important for risk analysis because negative or cautionary statements may contain early-warning information. A model that converts such statements to neutral may weaken risk signals. Model selection should therefore consider recall and macro F1-score rather than relying on accuracy alone.

3.3 SVM Confusion Matrix Analysis

Because Support Vector Machine achieved the strongest overall performance, its confusion matrix was examined in detail. Figure 3 shows 534 correctly classified neutral statements, 143 correctly classified positive statements, and 60 correctly classified negative statements. The 970-statement test set contained 571 neutral, 289 positive, and 110 negative statements. The most frequent errors were positive-to-neutral and negative-to-neutral predictions: 130 positive and 42 negative statements were assigned to the neutral class.

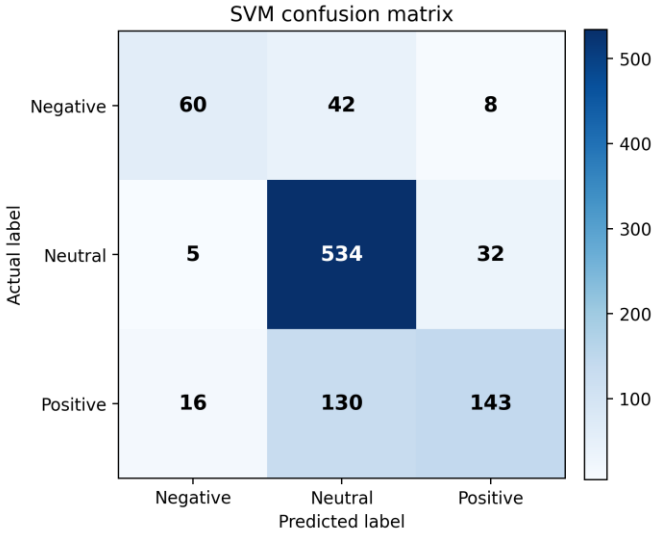

Figure 3. Confusion matrix of the Support Vector Machine classifier.

The confusion matrix reveals substantial differences in class-wise detection. SVM correctly classified 60 of 110 negative statements, 534 of 571 neutral statements, and 143 of 289 positive statements. The corresponding class recall values were approximately 54.5%, 93.5%, and 49.5%, respectively. The model therefore identified neutral statements substantially more reliably than positive or negative statements. In particular, 42 negative statements and 130 positive statements were classified as neutral. These errors indicate that the model's aggregate performance does not fully reflect its limitations in detecting non-neutral financial sentiment. The results support further investigation of class-sensitive training and evaluation strategies, but do not by themselves establish the effectiveness of the classifier as a financial-risk early-warning system.

3.4 Discussion of Individual Models

Random Forest achieved competitive accuracy and precision, indicating that its predictions were generally reliable. Its recall was lower than that of Support Vector Machine, showing that it missed more true sentiment cases. This limitation matters in risk monitoring, where failing to identify negative information may be more costly than issuing an occasional false alert.

Naive Bayes was computationally efficient and produced high precision, but its low recall and F1-score showed that it did not detect all relevant sentiment cases. The conditional independence assumption may be too restrictive for financial expressions whose interpretation depends on combinations of terms and context.

VADER produced the lowest overall performance. A general-purpose lexicon may assign unsuitable polarity to financial words whose meaning changes with context. Terms such as correction, exposure, pressure, volatility, and downgrade may not be represented adequately by fixed scores developed for general social-media language. The observed performance differences suggest that classifier selection depends on more than overall accuracy. Although SVM achieved the strongest aggregate results in this experiment, its class-wise recall indicates that positive and negative statements remain difficult to detect. This limitation is relevant when the intended application assigns

greater importance to negative financial information. Nevertheless, the present experiment did not define application-specific misclassification costs or evaluate downstream risk outcomes. Therefore, SVM should be interpreted as the strongest baseline among the four tested methods under the current experimental conditions, rather than as a validated risk-monitoring solution. Future comparative evaluations should examine whether class weighting, threshold adjustment, or resampling improves minority-class detection without substantially increasing false-positive predictions.

3.5 Comparison with Previous Studies and Practical Implications

The result is consistent with evidence that supervised machine-learning methods can outperform dictionary-based sentiment measures [9] and that finance-adapted language models capture context more effectively than generic approaches [10]-[13]. Recent deep-learning studies also report gains from domain-specific representations and contextual modelling [14]-[16]. The present scores are lower than those reported by several transformer-based studies because this experiment intentionally evaluated transparent conventional baselines rather than complex pretrained models. The findings should be compared with previous studies cautiously because differences in dataset composition, annotation criteria, preprocessing, feature representations, and evaluation protocols can affect reported performance. The present experiment provides a within-dataset comparison under a common evaluation framework, which helps isolate differences among the tested approaches. However, it does not establish that the observed ranking will remain unchanged across financial news sources or market periods. A stronger assessment of generalisability would require evaluation on an independent dataset or a temporally separated test set. Such validation would help determine whether the comparative advantage of SVM reflects transferable sentiment patterns rather than characteristics specific to the selected dataset and partition.

Unlike studies that directly combine sentiment with market forecasting or systemic-risk modelling [17]-[23], this experiment isolated the sentiment-extraction stage under one dataset, split, and evaluation protocol. The controlled design provides an interpretable baseline for deciding whether a lightweight model is adequate before introducing deep or transformer-based architectures. For operational use, Support Vector Machine is the preferred baseline, but its minority-class recall should be monitored before its outputs are used as risk alerts.

3.6 Limitations

This study has four principal limitations. First, it used one public English-language dataset, so the results may not generalise to other news sources, languages, or market conditions. Second, the neutral class dominated the dataset. Third, the study evaluated sentiment classification but did not test whether the predictions improved returns, volatility, or Value-at-Risk forecasting. Fourth, the comparison was limited to conventional machine-learning and lexicon-based baselines. An additional limitation concerns the sensitivity of the reported results to the selected data partition and model configuration. Because the evaluation uses a single train-test split, the reported scores do not quantify variation across alternative partitions or establish statistical significance between competing classifiers. Furthermore, the analysis does not evaluate calibration, application-specific misclassification costs, or the temporal relationship between sentiment signals and market-risk indicators. These limitations restrict the conclusions to comparative sentiment classification on the selected dataset. Future work should use repeated stratified evaluation or an appropriately designed temporal validation procedure, report uncertainty around performance estimates, and assess whether sentiment predictions provide measurable incremental value for downstream financial-risk tasks.

4. CONCLUSION

This study compared VADER, Random Forest, Multinomial Naive Bayes, and linear Support Vector Machine for three-class sentiment classification using 4,846 statements from the Financial PhraseBank dataset. Based on the reported results, SVM achieved the highest overall performance, with an accuracy of 0.76 and a macro F1-score of 0.69, followed by Random Forest, with an accuracy of 0.75 and a macro F1-score of 0.66. VADER produced the lowest reported performance among the evaluated approaches.

The SVM confusion matrix further revealed that neutral statements were detected more reliably than positive and negative statements, highlighting the importance of class-wise evaluation in imbalanced sentiment classification. These findings establish a comparative baseline for the tested methods under the specified experimental conditions. However, the study does not demonstrate that sentiment classification improves financial-risk forecasting or investment outcomes. Further research should verify performance across independent datasets, assess the stability of the model ranking, improve minority-class detection, and evaluate the incremental contribution of sentiment signals to downstream risk models.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge UNITAR International University and Universitas Sumatera Utara for academic support during the preparation of this study.

DECLARATIONS

AI USAGE STATEMENT

AI-assisted tools were used only for language refinement, grammar checking, and formatting support. The authors reviewed and approved the research design, data interpretation, methodological decisions, results, and conclusions and accept full responsibility for the originality, accuracy, and integrity of the manuscript.

AUTHOR CONTRIBUTION

Hasanul Fahmi Zuhri contributed to conceptualization, methodology development, manuscript drafting, supervision, and final manuscript review. Noor Azma Ismail contributed to research validation, financial risk interpretation, and manuscript review. Fazidah Wahit contributed to methodological evaluation, critical review of the research findings, and manuscript revision. Amalia Amalia contributed to data preparation, literature review, and manuscript editing. All authors reviewed and approved the final version of the manuscript and agree to be accountable for the integrity and accuracy of the work.

FUNDING STATEMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CONFLICTING INTERESTS

The authors declare that there are no financial or non-financial conflicts of interest related to this manuscript.

REFERENCES

[1] H. Wasserbacher and M. Spindler, "Machine learning for financial forecasting, planning and analysis: recent developments and pitfalls," *Digit. Financ.*, vol. 4, no. 1, pp. 63–88, 2022, doi: https://doi.org/10.1007/s42521-021-00046-2.

[2] H. Li, H. Gao, C. Wu, and M. A. Vasarhelyi, "Extracting financial data from unstructured sources: Leveraging large language models," *J. Inf. Syst.*, vol. 39, no. 1, pp. 135–156, 2025, doi: https://doi.org/10.2308/ISYS-2023-047.

[3] K. Du, F. Xing, R. Mao, and E. Cambria, "Financial sentiment analysis: Techniques and applications," *ACM Comput. Surv.*, vol. 56, no. 9, pp. 1–42, 2024, doi: https://doi.org/10.1145/3649451.

[4] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-based Syst.*, vol. 226, no. 8, p. 107134, 2021, doi: https://doi.org/10.1016/j.knosys.2021.107134.

[5] A. H. Shapiro, M. Sudhof, and D. J. Wilson, "Measuring news sentiment," *J. Econom.*, vol. 228, no. 2, pp. 221–243, 2022, doi: https://doi.org/10.1016/j.jeconom.2020.07.053.

[6] A. Sinha, S. Kedas, R. Kumar, and P. Malo, "SEntFiN 1.0: Entity-aware sentiment analysis for financial news," *J. Assoc. Inf. Sci. Technol.*, vol. 73, no. 9, pp. 1314–1335, 2022, doi: https://doi.org/10.1002/asi.24634.

[7] S. R. Das, M. Donini, M. B. Zafar, J. He, and K. Kenthapadi, "FinLex: An effective use of word embeddings for financial lexicon generation," *J. Financ. Data Sci.*, vol. 8, no. 11, pp. 1–11, 2022, doi: https://doi.org/10.1016/j.jfds.2021.10.001.

[8] T. Bos and F. Frasincar, "Automatically building financial sentiment lexicons while accounting for negation," *Cognit. Comput.*, vol. 14, no. 1, pp. 442–460, 2022, doi: https://doi.org/10.1007/s12559-021-09833-w.

[9] R. Frankel, J. Jennings, and J. Lee, "Disclosure sentiment: Machine learning vs. dictionary methods," *Manage. Sci.*, vol. 68, no. 7, pp. 5514–5532, 2022, doi: https://doi.org/10.1287/mnsc.2021.4156.

[10] A. H. Huang, H. Wang, and Y. Yang, "FinBERT: A large language model for extracting information from financial text," *Contemp. Account. Res.*, vol. 40, no. 2, pp. 806–841, 2023, doi: https://doi.org/10.1111/1911-3846.12832.

[11] M. Costola, O. Hinz, M. Nofer, and L. Pelizzon, "Machine learning sentiment analysis, COVID-19 news and stock market reactions," *Res. Int. Bus. Financ.*, vol. 64, no. 1, p. 101881, 2023, doi: https://doi.org/10.1016/j.ribaf.2023.101881.

[12] A. M. Ardekani, J. Bertz, C. Bryce, M. Dowling, and S. C. Long, "FinSentGPT: A universal financial sentiment engine?," *Int. Rev. Financ. Anal.*, vol. 94, no. 7, p. 103291, 2024, doi: https://doi.org/10.1016/j.irfa.2024.103291.

[13] G. Fatouros, J. Soldatos, K. Kouroumali, G. Makridis, and D. Kyriazis, "Transforming sentiment analysis in the financial domain with ChatGPT," *Mach. Learn. with Appl.*, vol. 14, no. 15, p. 100508, 2023, doi: https://doi.org/10.1016/j.mlwa.2023.100508.

[14] J. Kim, H.-S. Kim, and S.-Y. Choi, "Forecasting the S&P 500 index using mathematical-based sentiment analysis and deep learning models: A FinBERT transformer model and LSTM," *Axioms*, vol. 12, no. 9, p. 835, 2023, doi: https://doi.org/10.3390/axioms12090835.

[15] A. Groß-Klußmann, "Learning deep news sentiment representations for macro-finance," *Digit. Financ.*, vol. 6, no. 3, pp. 341–377, 2024, doi: https://doi.org/10.1007/s42521-024-00107-2.

[16] I. Almalis, E. Kouloumpris, and I. Vlahavas, "Sector-level sentiment analysis with deep learning," *Knowledge-Based Syst.*, vol. 258, no. 22, p. 109954, 2022, doi: https://doi.org/10.1016/j.knosys.2022.109954.

[17] M. Vicari and M. Gaspari, "Analysis of news sentiments using natural language processing and deep learning," *AI Soc.*, vol. 36, no. 3, pp. 931–937, 2021, doi: https://doi.org/10.1007/s00146-020-01111-x.

[18] F. Khalil and G. Pipa, "Is deep-learning and natural language processing transcending the financial forecasting? Investigation through lens of news analytic process," *Comput. Econ.*, vol. 60, no. 1, pp. 147–171, 2022, doi: https://doi.org/10.1007/s10614-021-10145-2.

[19] W. Khan, M. A. Ghazanfar, M. A. Azam, A. Karami, K. H. Alyoubi, and A. S. Alfakeeh, "Stock market prediction using machine learning classifiers and social media, news: W. Khan et al.," *J. Ambient Intell. Humaniz. Comput.*, vol. 13, no. 7, pp. 3433–3456, 2022, doi: https://doi.org/10.1007/s12652-020-01839-w.

[20] J. Bi, "Stock market prediction based on financial news text mining and investor sentiment recognition," *Math. Probl. Eng.*, vol. 2022, no. 1, p. 2427389, 2022, doi: https://doi.org/10.1155/2022/2427389.

[21] P. Xiao, "Stock market prediction based on financial news, text data mining, and investor sentiment analysis," *Int. J. Inf. Syst. Model. Des.*, vol. 15, no. 1, pp. 1–13, 2024, doi: 10.4018/IJISMD.361593.

[22] Y. Chen, G.-J. Wang, Y. Zhu, C. Xie, and G. Salah Uddin, "Identifying systemic risk drivers of FinTech and traditional financial institutions: machine learning-based prediction and interpretation," *Eur. J. Financ.*, vol. 30, no. 18, pp. 2157–2190, 2024, doi: https://doi.org/10.1080/1351847X.2024.2358940.

[23] J. Li, J. Li, X. Zhu, Y. Yao, and B. Casu, "Risk spillovers between FinTech and traditional financial institutions: Evidence from the US," *Int. Rev. Financ. Anal.*, vol. 71, no. 10, p. 101544, 2020, doi: https://doi.org/10.1016/j.irfa.2020.101544.

[24] A. Yadav, C. K. Jha, A. Sharan, and V. Vaish, "Sentiment analysis of financial news using unsupervised approach," in *Procedia Computer Science*, 2020, vol. 167, pp. 589–598. doi: https://doi.org/10.1016/j.procs.2020.03.325.

[25] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, "Good debt or bad debt: Detecting semantic orientations in economic texts," *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 4, pp. 782–796, 2014, doi: https://doi.org/10.1002/asi.23062.

[26] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, 2014, vol. 8, no. 1, pp. 216–225. doi: https://doi.org/10.1609/icwsm.v8i1.14550.

[27] R. Yacouby and D. Axman, "Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models," in *Proceedings of the first workshop on evaluation and comparison of NLP systems*, 2020, pp. 79–91. doi: https://doi.org/10.18653/v1/2020.eval4nlp-1.9.