Enhancing Financial Sentiment Analysis Using ML Models: A Comparative Study of Balanced vs. Imbalanced Datasets

Mahmudul Haque Shakir
Department of Computer Science
American International University - Bangladesh
Dhaka, Bangladesh
21-45016-2@student.aiub.edu

Sunipun Seemanta
Department of Computer Science
American International university - Bangladesh
Dhaka, Bangladesh
22-47547-2@student.aiub.edu

Riya Das
Department of Computer Science
American International University - Bangladesh
Dhaka, Bangladesh
22-46592-1@student.aiub.edu

Md. Saef Ullah Miah
Department of Computer Science
American International University - Bangladesh
Dhaka, Bangladesh
saef@aiub.edu

Junaida Sulaiman*
Faculty of Computing
University Malaysia Pahang Al-Sultan Abdullah
26600 Pekan, Pahang, Malaysia
junaida@umpsa.edu.my

2025 IEEE 9th International Conference on Software Engineering & Computer Systems (ICSECS) | 979-8-3315-4441-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICSECS65227.2025.11279229

***Abstract*—Financial sentiment analysis plays a crucial role in interpreting market sentiment from textual data, significantly influencing financial decision-making and forecasting. A common challenge in such analyses is the issue of class imbalance, where certain sentiments, such as negative or neutral sentiments, are underrepresented. The objective of this study is to assess the impact of dataset balancing on the performance of machine learning models in the context of financial sentiment classification. To achieve this, a comparative analysis was conducted using a diverse set of machine learning models, including Logistic Regression, Random Forest, AdaBoost, Gradient Boosting, XGBoost, SVM, KNN, Decision Tree, BERT, RoBERTa, and DeBERTa. These models were trained on both balanced and imbalanced datasets, employing resampling techniques to address class imbalance. The models were evaluated based on precision, recall, F1-score, and overall accuracy. The results revealed that balancing the dataset significantly improved the accuracy and reliability of the models, with the RoBERTa model achieving the highest accuracy on balanced datasets among all models considered. Interestingly, the SVM model demonstrated competitive performance, suggesting that non-transformer models can offer a time and resource efficient alternative while maintaining high accuracy. Furthermore, the findings suggest that exploring more advanced balancing techniques and deep learning methods may further improve the accuracy and robustness of financial sentiment classification.**

Keywords—Financial Sentiments Analysis, Machine Learning, Deep Learning, Balanced, Imbalanced, Textual Data

I. INTRODUCTION

The rapid growth of digital financial activities has consequently produced a lot of text, including news articles, social media posts, and financial reports that qualify for assessment of market sentiments. Financial sentiment analysis (FSA) has critical tools for extracting information from text, such as investor text. However, FSA is not only a challenge; many factors make it such a difficult task. Robust machine learning (ML) models that are capable of performing such tasks would be necessary to get reliability in sentiment classification [1]. The basic principle behind sentiment analysis involves employing the techniques of NLP, text analysis, and computational linguistics with biometrics to identify, extract, quantify, and study the regular patterns of affective states and subjective information [2]. FSA reads between the lines of unstructured, jargon-filled financial texts to predict market trends, separates subjective opinions from facts by handling subtleties like sarcasm, and integrates qualitative and quantitative data for accurate insights on market sentiments [3]. Financial sentiment analysis also faces complex terminologies, implicit opinions, limited annotated data, and several risks of misinterpretation; hence, it requires very specialized methods to ensure reliability [4].

A key issue in the FSA regime remains class imbalance, whereby one sentiment class (like the neutral sentiment) would be in abundance with all others (for instance, positive or negative sentiment) being underrepresented, thus biased model development is engendered. Some techniques for dataset balancing, such as SMOTE, may go some way toward mitigating this problem but can also provoke overfitting or losing interesting data. This research attempts to review many recent studies, evaluate techniques of dataset balancing, and gauge different ML models like LR, RF, AdaBoost, Gradient Boost, XGBoost, SVM, KNN, Decision Tree, BERT, RoBERTa and DeBERTa on both balanced and imbalanced datasets. In this regard, the datasets will go through preprocessing, including missing value imputation, text vectorization, and hyperparameter tuning before classification.

Performance will be analyzed by means of accuracy, precision, and F1-score to assess the impinging influence of balancing upon model efficacy. Results will also be contrasted with prior studies to show improvements within the ML-based frameworks. Optimizing these strategies of data balancing aims to make FSA petrifyingly effective for financial decision making and diminishing wastage of resources.

II. Literature Review

Financial sentiment analysis (FSA) has become an important technique for understanding market dynamics, investor sentiment, and economic trends. By analyzing textual data in the form of news articles, social media posts, and financial reports, it can produce insights on market behavior for informed decision-making and the evidence-based forecast. In addition, the growth of new-age machine learning and deep learning models opens many avenues toward the financial sentiment analysis. In addition to this, it has been explored in several studies to develop methods that tackle class imbalance in sentiment analysis. This paper will discuss some related work in the areas of machine learning, deep learning, preprocessing techniques, and insights into data balancing and handling imbalanced datasets.

Machine learning and its techniques have always served as a foundation for financial sentiment analysis as they processed both structured and unstructured data effectively. Karanikola et al. (2023) compared traditional ML models with advanced neural networks like BERT and FinBERT, highlighting their effectiveness in understanding financial language [5]. Dogra et al. demonstrated that Random Forest with SMOTE achieved the best performance in categorizing bank-oriented financial news, with balanced bagging also performing well [6]. Carta et al. employed explainable AI for feature selection in financial forecasting but lacked causal analysis and relied on Random Forest, limiting flexibility [7]. Jie Sun et al. found that imbalance-oriented SVM improved financial distress prediction, though limited sample size affected generalization [8]. Chiong et al. showed that SVM optimized with Particle Swarm Optimization outperformed deep learning in accuracy and computation time but struggled with limited historical stock data [9].

Deep learning models of mostly neural networks and transformers transform the sphere of sentiment analysis by being much more capable of understanding contextual and semantic subtleties in textual data. Sohangir et al. applied deep learning algorithms to the StockTwits dataset, finding that CNN outperformed LSTM and doc2vec in stock price prediction [10]. Kefah Alissa and Omar Alzoubi demonstrated that fine-tuned BERT and RoBERTa models achieved superior results when ensembled using majority voting [11]. Aditya Gupta and Vijay Kumar Tayal explored RoBERTa for financial sentiment analysis, emphasizing the impact of balanced versus imbalanced datasets on model accuracy [12]. Day et al. highlighted deep learning's ability to identify sentiment trends in financial news but did not analyze media source influence [13].

This review examines strategies for addressing class imbalance in financial sentiment analysis, highlighting techniques, model performance, and the impact of balancing on improving precision, recall, and robustness, while also identifying the research gap of limited evaluation of balancing methods on real-world financial datasets.

III. Methodology

The methodology section shows a structure for analyzing financial statements. At the initial step Data collection subsection describes the data sources, selection, and some preprocessing steps. Next Model Preprocessing subsection indicates how data is prepared for training, like cleaning. Then model evaluation matrix subsection describes how the models' performance was measured. Fig. 1 graphically represents the whole process.

A. Data Collection

For this research, a Financial Sentiment Analysis dataset was used from [5], which is a combination of two datasets, FiQ and Financial PhraseBank. There are two columns: sentences, sentiments, and 5322 rows. Both columns contain textual data such that sentences refer to financial news or statements, and sentiments identify the sentiment label corresponding to each sentence.

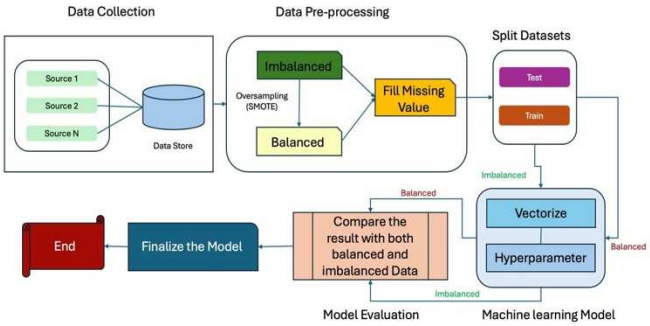

Fig. 1. Overview of process

B. Preprocessing

1) Data Balancing: The data is imbalanced, and this can result in the model favoring majority classes and low accuracy for minority classes. For this purpose, SMOTE was used to create synthetic samples, balancing all classes to 2508 instances each.

2) Train Data and Test Data: The model is trained on the training set and tested on another set to ensure performance. Both balanced and imbalanced sets were divided 80-20 for training and testing, respectively, to ensure enough data for learning and testing.

3) ML Models: Preprocessing the data makes the model more accurate. Missing values in the "Sentence" column were filled, and text was converted by a TF-IDF Vectorizer with bi-grams for greater context. Positive, Negative, Neutral labels were numericized. Model-specific parameters like C for Logistic Regression and n_estimators for Random Forest were tuned using GridSearchCV for improving precision and recall.

C. Machine Learning Model Evaluation

The performance of the classification model can be evaluated using several key metrics: accuracy, precision, recall, and F1-score, all together giving insight into the effectiveness of the model, each from a different viewpoint on the abilities of the classifier. Accuracy basically gives the overall correctness of the model's prediction that is the number of correctly classified instances divided by the total number of instances for which the model has made a prediction [20].

IV. Result and Discussion

This section demonstrates the performance across different models for both balanced and imbalanced datasets. The discussion section elaborates on a comparative analysis of the model's performance with results from previous studies.

1) Hyperparameters for each model (Balanced and imbal- anced:): Hyperparameters are crucial for a model's ability to generalize and impact its performance [15]. Table I lists optimized hyperparameters for balanced and imbalanced datasets. For Logistic Regression and SVM, regularization strength (C = 1) ensures a balance between model flexibility and overfitting the data [16]. Tree-based models (Random Forest, XGBoost, Decision Tree) use max_depth to limit tree growth. For balanced datasets, max_depth = 'None' allows the tree to grow fully, capturing complex patterns. In contrast, smaller depths (max_depth = 3 for XGBoost and max_depth = 5 for Decision Tree) simplify the model and mitigate overfitting in imbalanced datasets. The n_estimators parameter, which controls tree count, was set to 100 for Random Forest and AdaBoost for efficiency [17]. Gradient Boosting used n_estimators = 200 for balanced datasets but n_estimators = 100 for imbalanced data to avoid overfitting. The learning_rate was 0.5 for fast training on balanced datasets and 0.1 for imbalanced data to improve stability. SVM utilized an RBF kernel for feature extraction of complex patterns from balanced datasets and a linear kernel for quick computation in imbalanced datasets. KNN used 3 neighbors on balanced datasets for local pattern extraction and 7 on imbalanced ones for added robustness. Transformer models (BERT, RoBERTa, DeBERTa) were trained for 5 iterations in both cases to prevent underfitting and overfitting and to achieve efficiency. The optimized settings improved model performance in the different scenarios.

TABLE I. Optimal Hyperparameters For ML Models

Model	Parameters	
	Balanced	Imbalanced
Logistic Regression	{C: 1}	{C: 1}
Random Forest	{max_depth: none, n_estimators: 100}	{max_depth: none, n_estimators:100}
AdaBoost	{n_estimators: 100}	{n_estimators: 100}
Gradient Boosting	{n_estimators: 200, learning_rate: 0.5}	{n_estimators: 200, learning_rate: 0.1}
XGBoost	{max_depth: 7, n_estimators: 100}	{max_depth: 3, n_estimators: 100}
SVM	{C: 1, kernel: rbf}	{C: 1, kernel: linear}
KNN	{n_neighbors: 3}	{n_neighbors: 7 }
Decision Tree	{max_depth: none}	{max_depth: 5 }
Bert	{Total Epoch: 5}	{Total Epoch: 5}
Roberta	{Total Epoch: 5}	{Total Epoch: 5}
Deberta	{Total Epoch: 5}	{Total Epoch: 5}

2) Classification Performance on Imbalanced Dataset: Table II shows Negative, Neutral, and Positive classes model performance. Due to data imbalance, scores are highly volatile. Negative class is the hardest to predict. Random Forest performed worst (F1: 0.13). Decision Tree performed high Precision (0.59) but low Recall (0.09), with F1-score of 0.15. Logistic Regression, Gradient Boost, and KNN performed F1 value between 0.22 and 0.36. Roberta (0.54) and Deberta (0.36) performed better. The easiest was Neutral class. Bert, Roberta, and Deberta all had F1 scores greater than 0.79. Gradient Boost (Recall: 0.92) and Decision Tree (0.91) performed well, even if Precision was from 0.66 (Random Forest) to 0.72 (Logistic Regression). For the Positive class, Bert and Deberta worked best (F1: 0.78). Decision Tree (0.47) and KNN (0.57) didn't work well. Logistic Regression, SVM, and Roberta scored between 0.73 and 0.82. Random Forest and Gradient Boost had low Recall, which lowered their F1 scores. Generally speaking, models performed best with the Neutral class but poorly with Negative and Positive. Roberta and Deberta were most stable.

TABLE II. Performance Comparison Of ML Models On Imbalanced Dataset

Model	Class	Precision	Recall	F1-Score	Support
Logistic Regression	Negative	0.44	0.15	0.23	175
	Neutral	0.72	0.89	0.79	622
	Positive	0.76	0.69	0.73	372
Random Forest	Negative	0.19	0.10	0.13	175
	Neutral	0.66	0.85	0.74	622
	Positive	0.80	0.59	0.68	372
AdaBoost	Negative	0.49	0.34	0.40	175
	Neutral	0.69	0.84	0.76	622
	Positive	0.71	0.56	0.63	372
Gradient Boost	Negative	0.44	0.18	0.25	175
	Neutral	0.68	0.92	0.78	622
	Positive	0.81	0.55	0.66	372
XGBoost	Negative	0.38	0.16	0.22	175
	Neutral	0.69	0.90	0.78	622
	Positive	0.80	0.60	0.69	372
SVM	Negative	0.48	0.19	0.27	175
	Neutral	0.73	0.91	0.81	622
	Positive	0.78	0.69	0.73	372
KNN	Negative	0.41	0.18	0.25	175
	Neutral	0.67	0.91	0.77	622
	Positive	0.70	0.48	0.57	372
Decision Tree	Negative	0.59	0.09	0.15	175
	Neutral	0.61	0.91	0.73	622
	Positive	0.68	0.36	0.47	372
Bert	Negative	0.45	0.30	0.36	172
	Neutral	0.75	0.89	0.81	626
	Positive	0.85	0.71	0.78	371
Roberta	Negative	0.49	0.60	0.54	172
	Neutral	0.87	0.73	0.79	626
	Positive	0.76	0.89	0.82	371
Deberta	Negative	0.45	0.30	0.36	172
	Neutral	0.75	0.89	0.81	626
	Positive	0.85	0.71	0.78	371

Table III compares a few machine learning models on an imbalanced classification dataset by showing their performance regarding accuracy and the F1-score. Among the traditional models, the best performance is normally given by SVM, which tends to yield a test accuracy of 73.22%, an F1-score Macro of 61.0, and an F1-score Weighted of 71.0, hence effectively handling class imbalance. Logistic Regression, XGBoost, and AdaBoost had just average performance. At the same time, the worst was KNN and Decision Tree, likely because these models suffered more from overfitting problems and poor handling of imbalanced data. Transformer-based models, such as BERT, RoBERTa, and DeBERTa, exhibit the strongest performance, where RoBERTa provides the top results in test accuracy (76.13%) and F1-scores (72.0 Macro, 76.0 Weighted), demonstrating how well these models cope

with complex and imbalanced distributions. Given this, transformer models are more competitive compared to traditional algorithms and are more fitted for this dataset.

TABLE III. COMPARISON OF MODELS ON IMBALANCED DATA

Model	Accuracy (%)		F1-Score	
	Validation	Test	Macro	weighted
Logistic Regression	69.65 ± 2.02	71.69	58	69
Random Forest	63.69 ± 1.14	65.44	52	63
AdaBoost	67.00 ± 1.08	67.75	60	66
Gradient Boost	67.43 ± 1.08	69.75	56	66
XGBoost	67.56 ± 1.64	69.29	56	67
SVM	70.17 ± 1.85	73.22	61	71
KNN	65.23 ± 1.35	66.12	53	63
Decision Tree	60.41 ± 1.24	61.42	45	56
BERT	-	74.68	65	74
RoBERTa	-	76.13	72	76
DeBERTa	-	74.68	65	74

TABLE IV. PERFORMANCE COMPARISON OF ML MODELS ON BALANCED DATASET

Model	Class	Precision	Recall	F1-Score	Support
Logistic Regression	Negative	0.81	0.83	0.82	472
	Neutral	0.76	0.80	0.78	515
	Positive	0.86	0.81	0.83	518
Random Forest	Negative	0.83	0.79	0.81	472
	Neutral	0.68	0.82	0.74	515
	Positive	0.89	0.75	0.81	518
AdaBoost	Negative	0.77	0.72	0.75	472
	Neutral	0.64	0.80	0.71	515
	Positive	0.77	0.64	0.70	518
Gradient Boost	Negative	0.80	0.78	0.79	472
	Neutral	0.71	0.77	0.74	515
	Positive	0.87	0.81	0.83	518
XGBoost	Negative	0.80	0.78	0.79	472
	Neutral	0.72	0.79	0.75	515
	Positive	0.88	0.82	0.85	518
SVM	Negative	0.83	0.85	0.84	472
	Neutral	0.75	0.81	0.78	515
	Positive	0.93	0.83	0.88	518
KNN	Negative	0.79	0.94	0.86	472
	Neutral	0.75	0.76	0.80	515
	Positive	0.89	0.72	0.80	518
Decision Tree	Negative	0.71	0.74	0.72	472
	Neutral	0.63	0.83	0.64	515
	Positive	0.75	0.65	0.73	518
Bert	Negative	0.83	0.70	0.83	501
	Neutral	0.74	0.83	0.78	502
	Positive	0.92	0.82	0.87	502
Roberta	Negative	0.82	0.86	0.84	501
	Neutral	0.80	0.77	0.78	502
	Positive	0.87	0.87	0.87	502
Deberta	Negative	0.84	0.83	0.84	501
	Neutral	0.76	0.78	0.77	502
	Positive	0.87	0.86	0.87	502

1) Classification Performance on balanced Dataset: Table IV summarizes model performance for an imbalanced dataset by Negative, Neutral, and Positive classes. F1-scores represent model performance. KNN, SVM, and transformers (BERT, RoBERTa, DeBERTa) work well with the Negative class (F1: 0.83–0.87), whereas Decision Tree lags with F1 of 0.72. For the Neutral class, performance is not good. RoBERTa and BERT are best (F1: 0.78), while Decision Tree is bad (F1: 0.64). Most models have higher recall than precision. Positive class has the maximum scores by RoBERTa, DeBERTa, and SVM (F1: 0.88). Most balanced are RoBERTa and DeBERTa with values 0.87 for both classes. In general, transformer models work the best in all classes and suit best for the case of balanced data. Classic models like Decision Tree and AdaBoost work bad, especially for Neutral and Positive classes. Again, RoBERTa and DeBERTa prevail over the others.

Table V shows results for different machine learning models on a balanced dataset in terms of accuracy and F1-scores. Among the traditional models, the best is the SVM model, which achieved 82.86% accuracy on the test set and an F1-score of 83.0 for both Macro and Weighted respectively, thus establishing its strong performance on balanced data. Logistic Regression, XGBoost, and KNN perform equally well with about 80% test accuracy, having F1-scores at around 80-81%. Decision Tree presents the minimum test accuracy of 69.50% with F1-scores of 70.0. The highlights, among them, are the Transformer-based models: RoBERTa has an accuracy of 83.12% and F1-scores of 83.0 for both Macro and Weighted, which can be considered as superior capability on balanced data. The next goes DeBERTa with the same F1 scores and an accuracy of 82.46%, whereas BERT has an accuracy of 76.05% and F1-scores of 65.0 for Macro and 74.0 for Weighted. Therefore, RoBERTa and DeBERTa have scored overall best results obviously for this balanced dataset.

TABLE V. COMPARISON OF ML MODELS ON BALANCED DATA

Model	Accuracy (%)		F1-Score	
	Validation	Test	Macro	weighted
Logistic Regression	80.75 ± 0.0087	81	81	81
Random Forest	78.31 ± 0.0107	78.34	79	79
AdaBoost	72.43 ± 0.0052	71.89	72	72
Gradient Boost	79.08 ± 0.0152	78.74	79	79
XGBoost	79.41 ± 0.0113	79.47	80	80
SVM	82.52 ± 0.0035	82.86	83	83
KNN	79.26 ± 0.0100	80.53	81	80
Decision Tree	69.64 ± 0.0117	69.5	70	70
BERT	-	76.05	65	74
RoBERTa	-	83.12	83	83
DeBERTa	-	82.46	82	82

Table VI provides a comparative analysis of model performance across balanced and imbalanced datasets. On this data set transform-based model, Roberta achieved height accuracy (83.12%) against the non-transfer model SVM (82.86). There are a few differences between the two models. On the other hand, for the imbalanced dataset, Roberta also performed the superior result, which is 76.13%. The overall optimal model is identified as the Roberta model, with an accuracy of 83.12% on the balanced dataset. Figure 2 shows the performance of the model Roberta (Balanced dataset) for three classes through the ROC curve of multi-classes, where each plots the True Positive Rate vs. False Positive Rate. Each

class has high separability: the AUC for Class 0 is 0.93, and that for Class 1 and Class 2 is 0.96. Curves above the random classifier line justify strong classification capability, specifically for Class 1 and Class 2, reflecting how well the model performs in distinguishing between classes.

TABLE VI. COMPARISON OF MODELS ON BALANCED AND IMBALANCED DATASET

Dataset	ML Model	Accuracy (%)
Balanced	SVM (Non-transformer)	82.86
	RoBerta (Transformer)	83.12
Imbalanced	SVM (Non-transformer)	73.22
	RoBerta (Transformer)	76.13

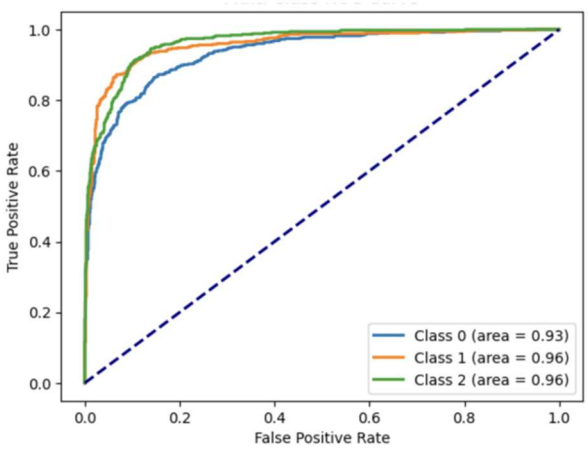

Fig. 2. ROC for Roberta on Balanced Dataset

Table VII shows the performance from previous studies on similar tasks and datasets. Karanikola et al. [5] used Roberta and achieved 77.58% accuracy. Kefah Alissa et al. [11] finding slightly higher accuracy (81.95%) with Majority Voting approach. Aditya Gupta et al. [12] used RoBERTa Large and got 80.71% accuracy. These previous results show the competitive performance between transformer models and ensemble methods for achieving high accuracy.

TABLE VII. COMPARISON OF OTHER RESEARCHERS WORK ON A SIMILAR DATASET

Study	Model	Accuracy
Karanikola et al. [5]	RoBERTa	77.58
Kefah Alissa et al. [11]	Majority Voting	81.95
Aditya Gupta et al [12]	RoBERTa large	80.71

This analysis demonstrates that the balanced dataset has produced favorable and satisfactory outcomes. RoBERTa achieves an accuracy of 83.12% on the balanced dataset, which is much higher than in earlier studies (Table 7). However, the effectiveness of our approach was validated. Nonetheless, the effectiveness of our methodology was validated. RoBERTa achieves an accuracy of 83.12% on the balanced dataset, significantly exceeding the outcomes of other studies (Table VII). A recent work by M. Mujahid et al. utilizing SMOTE on an imbalanced dataset revealed an accuracy of only 78% . RoBERTa achieves an accuracy of 83.12% on the balanced dataset [18]. SVM also performed similarly on the same balanced dataset, with an accuracy of 82.86%, which is both better than in the previous study and less than 0.26% away as compared to RoBERTa. Having an SVM as a non-transformer model saves on time and computational resources as well . Thus, other researchers can use the SVM model when time constraints restrict financial sentiment analysis, whereas the RoBERTa model is recommended for better accuracy.

V. CONCLUSION

The research investigates the effect of balanced and imbalanced datasets on machine learning models for financial sentiment classification. Traditional ML classifiers with three transformer models, that is, Logistic Regression, Random Forest, AdaBoost, Gradient Boosting, XGBoost, SVM, KNN, and Decision Tree, are compared against BERT, RoBERTa, and DeBERTa. The dataset contained financial statements labeled either positive, negative, or neutral, and was, as such, imbalanced and required SMOTE application for even class representation. The individual sentiment discrimination capability of the model was enhanced through resampling methods, resulting in improved classification performance. These findings confirmed that training on balanced datasets yields a higher precision score, recall score, and F1 score for each class of sentiment. Among traditional ML models, TF-IDF with SMOTE consistently enhanced classification accuracy. The transformer models, on the other hand, RoBERTa and BERT, outperformed the traditional classifiers. The RoBERTa model attained maximum accuracy of 83.12% after dataset balancing, establishing the efficacy of balancing techniques for improving sentiment classification performance. Future work should also explore hybrid ensembles further, including both traditional ML and transformer models, taking advantage of the strengths each possesses. This study underlines the efficiency of transformers in financial sentiment analysis and points out that strategic preprocessing and hyperparameter tuning enhance the performance of traditional models.

ACKNOWLEDGMENT

This research was supported by Universiti Malaysia Pahang Al-Sultan Abdullah International Matching Grant , Grant ID: RDU242726 and Tabung Persidangan Dalam Negara (TPDN).

REFERENCES

[1] 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, Oct. 2024, doi: 10.1145/3649451.

[2] K. Naithani and Y. P. Raiwani, "Realization of natural language processing and machine learning approaches for text-based sentiment analysis," *Expert Syst.*, vol. 40, no. 5, p. e13114, 2023, doi: 10.1111/exsy.13114.

[3] S. Sohangir, D. Wang, A. Pomeranets, and T. M. Khoshgoftaar, "Big Data: Deep Learning for financial sentiment analysis," *J. Big Data*, vol. 5, no. 1, p. 3, Jan. 2018, doi: 10.1186/s40537-017-0111-6.

[4] X. Man, T. Luo, and J. Lin, "Financial sentiment analysis (fsa): A survey," in *2019 IEEE international conference on industrial cyber physical systems (ICPS)*, IEEE, 2019, pp. 617–622. Accessed: Sept. 25, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8780312/

[5] A. Karanikola, G. Davrazos, C. M. Liapis, and S. Kotsiantis, "Financial sentiment analysis: Classic methods vs. deep learning

models," *Intell. Decis. Technol.*, vol. 17, no. 4, pp. 893–915, Nov. 2023, doi: 10.3233/IDT-230478.

[6] V. Dogra, S. Verma, K. Verma, N. Z. Jhanjhi, U. Ghosh, and D. N. Le, "A comparative analysis of machine learning models for banking news extraction by multiclass classification with imbalanced datasets of financial news: Challenges and solutions.," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 7, no. 3, pp. 35–52, 2022.

[7] S. Carta, A. S. Podda, D. Reforgiato Recupero, and M. M. Stanciu, "Explainable AI for Financial Forecasting," in *Machine Learning, Optimization, and Data Science*, vol. 13164, G. Nicosia, V. Ojha, E. La Malfa, G. La Malfa, G. Jansen, P. M. Pardalos, G. Giuffrida, and R. Umeton, Eds., in Lecture Notes in Computer Science, vol. 13164. , Cham: Springer International Publishing, 2022, pp. 51–69. doi: 10.1007/978-3-030-95470-3_5.

[8] J. Sun, Z. Shang, and H. Li, "Imbalance-oriented SVM methods for financial distress prediction: a comparative study among the new SB-SVM-ensemble method and traditional methods," *J. Oper. Res. Soc.*, vol. 65, no. 12, pp. 1905–1919, Dec. 2014, doi: 10.1057/jors.2013.117.

[9] R. Chiong, Z. Fan, Z. Hu, M. T. P. Adam, B. Lutz, and D. Neumann, "A sentiment analysis-based machine learning approach for financial market prediction via news disclosures," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, Kyoto Japan: ACM, July 2018, pp. 278–279. doi: 10.1145/3205651.3205682.

[10] S. Sohangir, D. Wang, A. Pomeranets, and T. M. Khoshgoftaar, "Big Data: Deep Learning for financial sentiment analysis," *J. Big Data*, vol. 5, no. 1, p. 3, Dec. 2018, doi: 10.1186/s40537-017-0111-6.

[11] K. Alissa and O. Alzoubi, "Financial sentiment analysis based on transformers and majority voting," in *2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA)*, IEEE, 2022, pp. 1–4. Accessed: Sept. 25, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10017941/

[12] A. Gupta and V. K. Tayal, "Analysis of Twitter sentiment to predict financial trends," in *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*, IEEE, 2023, pp. 1027–1031. Accessed: Sept. 26, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10085195/

[13] M.-Y. Day and C.-C. Lee, "Deep learning for financial sentiment analysis on finance news providers," in *2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, IEEE, 2016, pp. 1127–1134. Accessed: Sept. 25, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7752381/

[14] C. Goutte and E. Gaussier, "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation," in *Advances in Information Retrieval*, vol. 3408, D. E. Losada and J. M. Fernández-Luna, Eds., in Lecture Notes in Computer Science, vol. 3408., Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 345–359. doi: 10.1007/978-3-540-31865-1_25.

[15] C. Arnold, L. Biedebach, A. Küpfer, and M. Neunhoeffer, "The role of hyperparameters in machine learning models and how to tune them," *Polit. Sci. Res. Methods*, vol. 12, no. 4, pp. 841–848, 2024.

[16] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020.

[17] A. Mansoori, M. Zeinalnezhad, and L. Nazarimanesh, "Optimization of Tree-Based Machine Learning Models to Predict the Length of Hospital Stay Using Genetic Algorithm," *J. Healthc. Eng.*, vol. 2023, no. 1, p. 9673395, Jan. 2023, doi: 10.1155/2023/9673395.

[18] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *J. Big Data*, vol. 11, no. 1, p. 87, June 2024, doi: 10.1186/s40537-024-00943-4.