ISSN 2723-6471

Financial Condition Prediction Using a Soft Voting Ensemble Model Based on Structured Financial Indicators and Unstructured News Data

Ibnu Rasyid Munthe[1,*], Sumijan[2], Muhammad Tajudin[3]

[1,2]*Univiersitas Putra Indonesia YPTK Padang, Jl. Raya Lubuk Begalung, Padang and 25145, Indonesia*

[3]*Universitas Bumigora, Jl. Ismail Marzuki, Mataram and 83127, Indonesia*

(Received: February 5, 2026; Revised: April 1, 2026; Accepted: June 12, 2026; Available online: July 12, 2026)

Abstract

The rapid growth of capital market participation has increased the need for reliable analytical tools to assess the financial conditions of listed companies. Conventional approaches commonly rely on structured financial indicators and may not fully capture contextual information reflected in financial news. This study aims to evaluate the predictive potential of structured financial data and unstructured textual data using a soft voting ensemble framework. The structured dataset consists of 1,263 financial records collected from the reports of manufacturing companies listed on the Indonesia Stock Exchange, whereas the unstructured dataset contains 6,329 online financial news records related to Indonesian stocks and listed companies. The structured data were processed through missing-value handling, normalization, and class balancing, while the textual data were processed through cleaning, case folding, tokenization, filtering, stemming, and Term Frequency–Inverse Document Frequency (TF-IDF) feature extraction. The proposed ensemble model combines the probability outputs of six base classifiers: Decision Tree, Support Vector Machine, Multinomial Naive Bayes, Logistic Regression, Random Forest, and K-Nearest Neighbors. Experimental results show that the soft voting ensemble achieved an accuracy of 92.66% on the structured dataset and 98.53% on the unstructured dataset. These findings indicate that both financial indicators and textual information can provide valuable predictive signals for identifying company financial conditions. The contribution of this study lies in the comparative evaluation of two distinct data sources using a consistent ensemble-learning framework. However, the two datasets were evaluated independently rather than integrated into a single multimodal pipeline. Therefore, future research should develop a data-fusion mechanism to assess whether combining financial indicators and news-based sentiment can further improve predictive performance and strengthen decision support for investors and financial analysts.

Keywords: Financial condition prediction, Soft voting, Structured data, Unstructured data, Financial news

1. Introduction

The rapid advancement of digital technology has significantly expanded public access to capital markets in Indonesia. Over the past several years, the number of capital market investors has increased substantially. Data from 2017 to May 2024 indicate that the number of investors grew from 1.12 million to approximately 13 million. The most notable increase occurred in 2021, when investor growth reached 93.3%, largely driven by the surge in digital investment activities during the COVID-19 pandemic. However, this growth did not continue at the same pace. After reaching its peak, the growth rate gradually declined to 37.68% in 2022, 18.01% in 2023, and 6.84% in May 2024 [1].

This slowdown suggests that the rapid expansion in the number of investors has not been accompanied by sufficient improvement in investor literacy and understanding of financial markets. One of the main challenges is the limited knowledge of novice investors regarding capital market mechanisms and investment risk management [2]. Many investors who entered the market during the pandemic were influenced by social trends, fear of missing out (FOMO), or expectations of short-term profits without adequate financial literacy or risk awareness [3]. Consequently, when market volatility occurs, many of these investors tend to become inactive or withdraw from investment activities, which contributes to the stagnation of long-term active investor growth [4].

***Corresponding author: Ibnu Rasyid Munthe (ibnurasyidmunthe@gmail.com)**
DOI: https://doi.org/10.47738/jads.v7i3.1417

In this context, the availability of analytical systems capable of providing reliable financial insights becomes increasingly important. One critical aspect in investment decision-making is the ability to identify the potential financial distress of companies, which may threaten business sustainability and investment value. Traditionally, financial distress analysis has relied on financial ratio analysis, including profitability, liquidity, and solvency indicators [5]. Although these indicators provide useful information about corporate financial health, traditional financial analysis alone often fails to capture the broader dynamics influencing company performance, particularly under conditions of global economic uncertainty.

Recent developments in machine learning (ML) have provided promising approaches for improving financial distress prediction. ML techniques are capable of identifying complex patterns in large-scale data and adapting to dynamic financial environments. In addition to utilizing structured financial data, ML models can also incorporate unstructured information such as online news articles and financial reports, thereby enabling more comprehensive and adaptive predictions [6]. The integration of structured and unstructured data sources allows predictive models to capture both financial performance indicators and external informational signals that may affect corporate stability.

Several previous studies have explored the application of machine learning models for financial condition prediction; however, several limitations remain. A previous study employed multiple models, including LDA, QDA, Logistic Regression, KNN, Decision Tree, Random Forest, Ridge, AdaBoost, SVM, Bagging, and GBC with GridSearchCV for parameter optimization, achieving an accuracy of 99.52% [7]. However, the study did not address the issue of class imbalance. Another study proposed a hybrid Decision Tree–Logit model that achieved an accuracy of 88.02%, although it did not apply data balancing techniques or parameter tuning [8]. A further study implemented Linear SVM, bagMLP, and Boost.CART models and reported an accuracy of 91.27%, but did not address class imbalance or hyperparameter optimization [9]. Other research attempted to address these limitations by combining Random Forest, SMOTE, and Optuna, achieving an accuracy of 86.7%, although the approach relied on a single primary model [10]. Meanwhile, another study integrated multiple models, including Logistic Regression, SVM, Random Forest, Decision Tree, XGBoost, and LightGBM with SMOTE, obtaining an accuracy of 87.3%. However, further improvements remain possible through more systematic parameter optimization and ensemble strategies [11].

Based on these observations, two key challenges remain prominent in financial distress prediction: the issue of class imbalance and the lack of optimal hyperparameter tuning. To address these challenges, this study applies the Synthetic Minority Over-sampling Technique (SMOTE) to balance class distributions within the dataset [12] and employs GridSearchCV to systematically identify the optimal hyperparameter configurations for each base model [13]. These techniques are expected to enhance model generalization and reduce classification bias toward majority classes.

Furthermore, this study proposes an ensemble learning approach based on soft voting that integrates several base classifiers, including Decision Tree, K-Nearest Neighbors, Logistic Regression, Random Forest, Support Vector Machine, and Naive Bayes. The soft voting mechanism aggregates the predicted probabilities from each classifier, enabling the final prediction to consider the confidence level of each model rather than relying solely on majority class votes. This strategy is expected to improve predictive robustness, reduce overfitting, and enhance overall classification performance when dealing with complex and imbalanced data [14].

Another distinctive contribution of this study lies in the evaluation of structured and unstructured data sources for financial condition prediction. Rather than implementing a full multimodal fusion framework, this study separately evaluates structured financial indicators and unstructured financial news to examine their respective predictive capabilities. The structured dataset represents company financial performance, while the unstructured dataset captures contextual information from financial news. This separation allows the study to compare how different data sources contribute to financial condition classification. By applying SMOTE for class balancing, optimizing model parameters using GridSearchCV, and utilizing a soft voting ensemble strategy, this study aims to develop a reliable prediction model for each data representation.

In this study, the terms “financial distress prediction” and “financial condition prediction” are used with a clear conceptual distinction. Financial distress prediction refers specifically to the classification of companies that show indications of financial difficulty or risk, while financial condition prediction refers to the broader classification of company financial status into “Good” and “Not Good” categories. Therefore, the main prediction target of this study

is financial condition classification, with the "Not Good" category interpreted as an indication of potential financial distress.

2. Literature Review

Financial distress prediction has become an important research area in financial analytics because it supports investors, analysts, and decision-makers in identifying companies that may experience financial instability. Traditional approaches commonly rely on structured financial indicators, such as profitability, liquidity, solvency, leverage, firm size, return on equity, sales growth, and debt-to-equity ratio. These indicators provide quantitative information about company performance and have been widely used to assess financial health and potential business failure [5], [6]. However, financial ratios alone may not fully capture external factors that influence company conditions, particularly market sentiment, corporate events, and macroeconomic uncertainty.

Machine learning has been increasingly applied to improve financial distress prediction because of its ability to identify complex and nonlinear patterns in large datasets. Several algorithms, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, Naive Bayes, AdaBoost, and Gradient Boosting, have been used in previous studies to classify companies into financially healthy or distressed categories. Study [7] proposed an ensemble learning model using a majority voting mechanism and reported high predictive performance in financial distress prediction. Study [8] developed a hybrid Decision Tree–Logistic Regression model for business failure prediction during economic downturns, showing that hybrid models can improve prediction capability compared with a single model. Study [9] also demonstrated that classifier ensemble approaches can improve financial distress prediction by combining multiple learning models.

Despite the promising performance of machine learning models, several challenges remain in financial distress prediction. One important challenge is class imbalance, where the number of financially healthy companies is often larger than financially distressed companies. This imbalance can cause machine learning models to become biased toward the majority class and reduce their ability to identify minority-class cases. To address this issue, resampling techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) have been widely used to balance class distributions and improve classification performance [12]. In addition, hyperparameter optimization is also important because model performance can be strongly influenced by parameter settings. Optimization techniques such as GridSearchCV can systematically search for the best parameter combinations and improve model generalization [13].

In addition to structured financial data, unstructured textual data such as financial news has become an increasingly valuable source of information for prediction tasks. News articles can capture external signals related to market perception, corporate reputation, public sentiment, and economic events that may not be directly reflected in financial statements. Textual data can be transformed into numerical features using text mining techniques such as Term Frequency–Inverse Document Frequency (TF-IDF). By incorporating textual information, prediction models can obtain a broader representation of company conditions and improve classification accuracy. Study [6] emphasized the importance of incorporating financial, textual, and social responsibility features into financial distress prediction systems.

Ensemble learning has also been widely recognized as an effective strategy for improving classification robustness. Instead of relying on a single classifier, ensemble methods combine several models to produce a more reliable final prediction. Soft voting is one ensemble technique that aggregates the predicted probabilities generated by multiple classifiers. Unlike hard voting, which only considers the predicted class label, soft voting considers the confidence score of each model. This allows the ensemble model to make more informed predictions and reduce the limitations of individual classifiers. Previous research has shown that voting-based ensemble models can improve predictive stability and classification accuracy in various domains [11], [14].

Based on previous studies, it can be concluded that financial distress prediction can benefit from three main strategies: integrating structured and unstructured data, addressing class imbalance, and applying ensemble learning. However, many previous studies still focus on either structured financial indicators or single-source data. Some studies also do not fully address class imbalance or systematic hyperparameter optimization. Therefore, this study proposes a soft voting ensemble framework that combines structured financial indicators and unstructured financial news data. By

 ISSN 2723-6471

applying preprocessing, TF-IDF feature extraction, SMOTE-based balancing, GridSearchCV optimization, and soft voting ensemble classification, this study aims to develop a more comprehensive and reliable model for financial condition prediction.

3. Methodology

This study proposes a machine learning framework to predict financial distress by integrating structured financial data and unstructured textual data. The methodology consists of several stages, including data preprocessing, base model training, ensemble classification using soft voting, and hyperparameter optimization with GridSearchCV. Structured data are processed through missing value handling and normalization, while unstructured text data undergo text preprocessing such as cleaning, tokenization, filtering, and stemming. The outputs from multiple base classifiers are then combined using a soft voting mechanism to obtain the final prediction. The overall research framework is illustrated in figure 1.

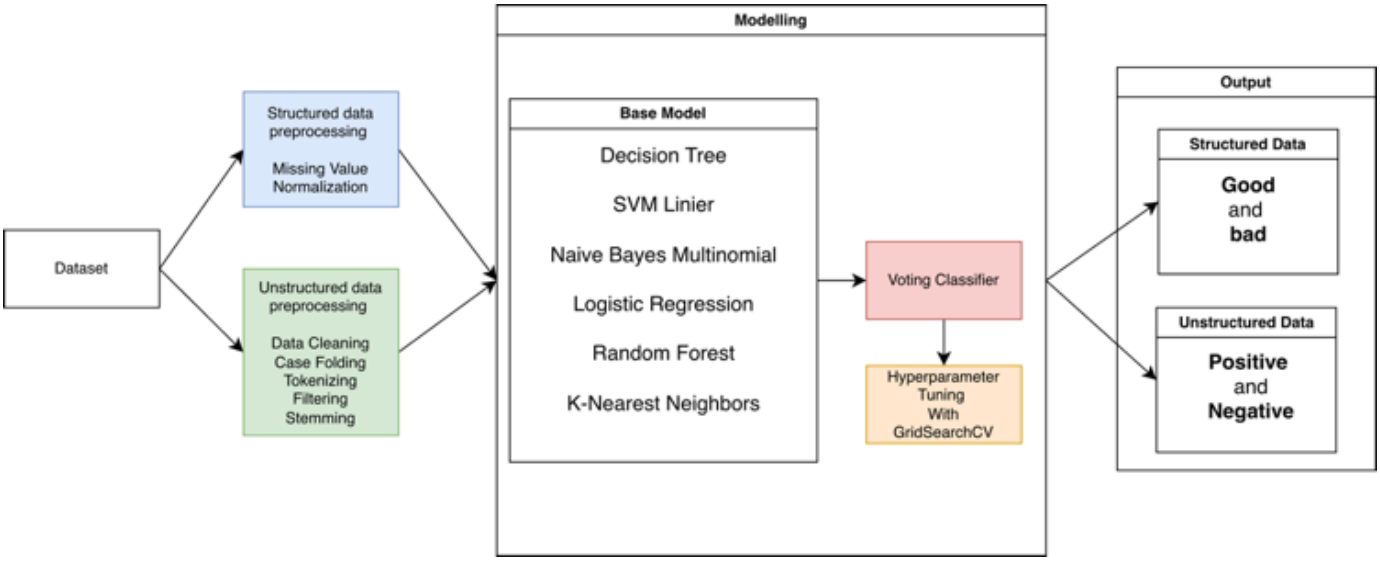

Figure 1. Methodology Flow

3.1. Dataset

This study uses two types of data, namely structured data and unstructured data, to support financial condition prediction based on both financial indicators and textual information. The integration of these two datasets enables the proposed model to capture numerical patterns derived from financial statements as well as contextual information obtained from online financial news.

The structured dataset consists of secondary financial data obtained from the official website of the Indonesia Stock Exchange (IDX). The data were collected from the financial reports of manufacturing companies listed on the IDX and include several financial and firm-related variables, namely profitability, liquidity, firm size, return on equity (ROE), sales growth, debt-to-equity ratio (DER), managerial ownership, institutional ownership, and firm value. In total, the structured dataset contains 1,263 records. Before the modelling stage, the structured data were processed through missing-value handling, normalization, and class balancing to improve data quality and support model performance.

The unstructured dataset consists of textual data collected from online financial news related to Indonesian stocks and listed companies. This dataset is used to capture external information that may reflect market sentiment, corporate conditions, and other contextual factors that are not directly represented in financial statements. In total, the unstructured dataset contains 6,329 news records. Before being used in the modelling stage, the textual data were processed through several natural language preprocessing steps, including data cleaning, case folding, tokenization, filtering, and stemming. The processed text was subsequently transformed into numerical feature vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) method.

The combination of structured and unstructured datasets provides a more comprehensive basis for prediction. The structured dataset represents the internal financial characteristics of the companies, whereas the unstructured dataset captures external information and market-related perspectives reflected in online news. By integrating these two types of data, the proposed approach can analyze both numerical financial patterns and textual market information.

Nevertheless, this study has a limitation regarding the scope of the structured financial dataset. The structured data used in this research were collected exclusively from manufacturing companies listed on the Indonesia Stock Exchange. Therefore, the findings should be interpreted specifically within the context of financial condition prediction in the manufacturing sector and should not be directly generalized to all industry sectors. Differences in financial structures, business models, capital intensity, and risk characteristics across sectors may influence model performance when the proposed approach is applied to other industries.

Although the proposed soft voting ensemble model demonstrates promising results, its external validity remains limited by the sector-specific nature of the structured dataset. Future research should evaluate the proposed approach using datasets from broader industry sectors, such as banking, services, mining, consumer goods, and technology companies. Such evaluations are necessary to assess the generalizability and robustness of the model across different financial contexts.

3.2. Labeling Process

The labeling process was conducted to define the target classes used in the financial condition classification task. In this study, companies were classified into two categories, namely “Good” and “Not Good.” The “Good” class represents companies with relatively healthy financial conditions, while the “Not Good” class represents companies that show indications of potential financial risk or financial distress.

For the structured dataset, class labels were assigned based on selected financial indicators obtained from company financial reports, including profitability, liquidity, solvency, leverage, return on equity, sales growth, debt-to-equity ratio, and firm value. Companies with stronger financial performance, such as positive profitability, stable liquidity, lower leverage risk, and positive growth indicators, were categorized as “Good.” In contrast, companies showing weaker financial performance, such as low or negative profitability, poor liquidity, high leverage, declining sales growth, or unfavorable firm value indicators, were categorized as “Not Good” or Bad.

For the unstructured dataset, the labels were assigned based on the financial meaning of news content related to each company. News containing positive signals, such as improved financial performance, business expansion, profit growth, positive corporate actions, or favorable market responses, was categorized as “Good” or positive. Meanwhile, news containing negative signals, such as losses, declining performance, debt problems, legal issues, operational disruption, bankruptcy risk, or unfavorable market responses, was categorized as “Not Good” or negative.

This labeling procedure was used to ensure that both structured financial indicators and unstructured financial news were mapped into comparable binary classification targets. Therefore, the prediction task in this study focuses on distinguishing companies with healthy financial conditions from those with potential financial risk. The explicit definition of the labeling criteria is expected to improve the reproducibility and conceptual clarity of the proposed financial condition prediction framework.

3.3. Data Preprocessing

Structured and unstructured datasets undergo different preprocessing procedures before being used for model training. For the structured dataset, the financial data are initially examined to identify missing values that may arise from incomplete financial reports or unavailable financial indicators. Appropriate imputation techniques are then applied to replace the missing values while preserving the statistical characteristics of the dataset and reducing the potential for bias during the model-learning process [15]. After missing-value handling, data normalization is performed to ensure that all numerical features are represented within a comparable range. This step is essential because financial indicators may have substantially different numerical scales, which can negatively affect the performance of scale-sensitive machine learning algorithms, particularly Support Vector Machine (SVM) and K-Nearest Neighbors (KNN). Through normalization, each numerical feature can contribute proportionally to the model-training process [16].

 ISSN 2723-6471

Meanwhile, the unstructured dataset, which consists of textual data collected from online financial news, is processed through several natural language preprocessing stages to transform raw text into meaningful features. The preprocessing pipeline includes data cleaning, case folding, tokenization, filtering, and stemming. Data cleaning is performed to remove irrelevant elements, such as punctuation marks, special symbols, and URLs [17]. Case folding subsequently converts all text into lowercase format to maintain consistency in text representation [18]. Tokenization is then applied to split sentences into individual words or tokens [19]. Afterward, filtering is conducted to remove stopwords that do not provide significant semantic information. Finally, stemming reduces words to their root forms to minimize linguistic variations and improve feature consistency [20]. These preprocessing stages enable the extraction of relevant textual features that can be used to analyze sentiment or trends associated with company performance.

3.4. Term Frequency–Inverse Document Frequency (TF-IDF)

The unstructured textual data were transformed into numerical representations using the Term Frequency–Inverse Document Frequency (TF-IDF) method. TF-IDF was selected because it can effectively measure the importance of words within documents while reducing the influence of overly frequent terms that carry limited discriminative information.

Before TF-IDF feature extraction, several preprocessing steps were applied, including data cleaning, case folding, tokenization, stopword removal, and stemming. Stopword handling was performed to remove common Indonesian words that do not contribute significant semantic meaning to the classification process, such as conjunctions, prepositions, and other frequently occurring functional words. This process helps reduce textual noise and improves the quality of extracted features.

The TF-IDF vectorization process used a unigram configuration to represent individual word importance within financial news articles. The vocabulary was automatically generated from the preprocessed corpus, resulting in a sparse numerical feature matrix representing the textual characteristics of the news dataset. The extracted TF-IDF features were then directly used as input for the classification models within the proposed soft voting ensemble framework. The use of TF-IDF enables the model to capture important financial terms, market-related expressions, and company-specific information contained in the news articles, thereby supporting the classification of company financial conditions based on textual information.

3.5. SMOTE

Class imbalance analysis was conducted before applying the Synthetic Minority Over-sampling Technique (SMOTE) to evaluate the distribution of the target classes within the dataset. The analysis showed that the number of instances in the majority class was considerably higher than the minority class, indicating the presence of class imbalance that could potentially bias the classification models toward the dominant category.

For the structured dataset, the original class distribution consisted of 823 instances in the “Good” class and 440 instances in the “Not Good” class. This imbalance could reduce the model’s ability to correctly identify minority-class instances associated with potential financial risk. Therefore, SMOTE was applied to synthetically increase the minority class samples and produce a more balanced class distribution for model training. After applying SMOTE, both classes contained an equal number of instances, resulting in a balanced dataset.

Similarly, the unstructured dataset also exhibited class imbalance, consisting of 4005 positive (“Good”) news records and 2324 negative (“Not Good”) news records. Since the positive class significantly outnumbered the negative class, SMOTE was applied after TF-IDF feature extraction to balance the textual dataset before the classification process. After resampling, both classes were adjusted to contain equal numbers of instances, allowing the models to learn more representative patterns from both categories. The application of SMOTE is expected to reduce classification bias toward majority classes and improve the capability of machine learning models in identifying minority-class patterns related to potential financial risk. By balancing the class distribution, the proposed framework aims to achieve more reliable and unbiased financial condition classification performance.

3.6. Train-Test Splitting Strategy

To evaluate the performance of the proposed model, the datasets were divided into training and testing subsets using an 80:20 split ratio. The training set was used to train the machine learning models and perform hyperparameter optimization, while the testing set was used to evaluate the predictive performance of the final models on unseen data.

In order to maintain the original class distribution between the "Good" and "Not Good" categories, stratified sampling was applied during the train-test splitting process. The use of stratified sampling ensures that the proportion of each class remains consistent in both the training and testing datasets, thereby reducing potential sampling bias and improving the fairness of the experimental evaluation.

For the structured dataset, stratified splitting was performed before the application of SMOTE to prevent information leakage between training and testing data. SMOTE was then applied only to the training dataset to balance the minority class during model training. Similarly, for the unstructured dataset, the textual data were first divided into training and testing subsets using stratified sampling before TF-IDF transformation and model training were conducted.

This experimental setup was designed to ensure a fair and consistent evaluation process while minimizing overfitting risk and preserving the reliability of the classification results.

3.7. Based Model Development

To build a robust prediction system, this study utilizes multiple machine learning algorithms as base classifiers. The selected models include Decision Tree (DT), Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), Logistic Regression (LR), Random Forest (RF), and K-Nearest Neighbors (KNN). Each algorithm has different learning characteristics, enabling the model ensemble to capture diverse data patterns.

Decision Tree is used due to its ability to model non-linear relationships and provide interpretable decision rules [21]. Support Vector Machine is effective for high-dimensional classification tasks and can handle complex decision boundaries [22]. Multinomial Naive Bayes is particularly suitable for text-based classification tasks involving probabilistic word distributions [23]. Logistic Regression provides a strong baseline linear classifier with stable performance in binary classification problems [24]. Random Forest enhances predictive accuracy by aggregating multiple decision trees [25], while K-Nearest Neighbors performs classification based on similarity between data points [26].

The selection of these algorithms is motivated by previous studies that demonstrated the effectiveness of hybrid and ensemble approaches in financial distress prediction. For example, [7] proposed a Hybrid Voting Model combining multiple classifiers such as LDA, QDA, Logistic Regression, KNN, Decision Tree, Random Forest, Ridge Classifier, AdaBoost, Support Vector Machine, Bagging, and Gradient Boosting Classifier, achieving an accuracy of 98%. Similarly, [8] introduced a hybrid Decision Tree–Logistic Regression (DT-LOGIT) model with an accuracy of 88% [9] proposed hybrid models combining Linear SVM, bagging-based MLP, and boosted CART, reporting an accuracy of 91.27% [10] implemented a Random Forest model with an accuracy of 86.7% while [11] combined Logistic Regression, SVM, Random Forest, Decision Tree, XGBoost, and LightGBM, achieving an accuracy of 87.3%.

Although these studies demonstrate the potential of hybrid models, many approaches rely on limited combinations of algorithms or single-model frameworks. Therefore, this research expands previous work by combining multiple base classifiers into a unified ensemble system using a voting mechanism.

3.8.Ensemble Voting Model

To improve predictive robustness, this study applies a soft voting ensemble strategy to aggregate the predictions from all base models. In soft voting, each classifier produces a probability score for each class label. The final prediction is determined by averaging the predicted probabilities across all models and selecting the class with the highest combined probability [27].

Compared to hard voting, which only considers the final class label from each classifier [28], soft voting incorporates the confidence level of each model's prediction. This mechanism allows the ensemble model to generate more stable and reliable predictions, especially when dealing with heterogeneous data sources such as financial ratios and textual news information.

The ensemble approach also reduces the risk of overfitting associated with individual models and improves generalization performance when applied to unseen data. The soft voting ensemble model integrates the probability outputs generated by six base classifiers: Decision Tree (DT), Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), Logistic Regression (LR), Random Forest (RF), and K-Nearest Neighbor (KNN). For each input instance x_i, the final probability of class c_k is calculated by averaging the class probabilities produced by the six

 ISSN 2723-6471

classifiers. The predicted class is subsequently determined by selecting the class with the highest aggregated probability, as expressed in the following equation:

$$\hat{y}_i = \underset{c_k \in C}{\text{argmax}} \frac{P_{DT}(c_k|x_i) + P_{SVM}(c_k|x_i) + P_{MNB}(c_k|x_i) + P_{LR}(c_k|x_i) + P_{RF}(c_k|x_i) + P_{KNN}(c_k|x_i)}{6} \quad (1)$$

$P_m(c_k|x_i)$ represents the probability assigned by classifier m to class c_k for input instance x_i, while $\hat{y}_i$ denotes the final predicted class. In the implementation of soft voting, the SVM model must be configured to produce probability estimates, for example through probability calibration.

3.9. Hyperparameter Optimization

To improve model performance and ensure robust parameter selection, hyperparameter optimization was conducted using the GridSearchCV method. GridSearchCV systematically evaluates multiple combinations of predefined hyperparameters for each classification model using a cross-validation procedure. In this study, the optimization process employed 5-fold cross-validation to ensure reliable model evaluation and reduce overfitting risk during training. The scoring metric used during optimization was classification accuracy, which was selected to identify the parameter combination producing the best predictive performance [29].

Several parameter grids were defined for each base classifier. For the K-Nearest Neighbors (KNN) model, the optimization process evaluated different values of the number of neighbors (n_neighbors = 3, 5, 7, 9) and distance metrics. For Logistic Regression (LR), different regularization strengths (C = 0.1, 1, 10) and solver configurations were tested. For Support Vector Machine (SVM), the optimization explored different kernel functions and regularization parameters (C = 0.1, 1, 10). For Random Forest (RF), the tuning process evaluated different numbers of trees (n_estimators = 100, 200), maximum tree depths, and minimum sample split values. For Decision Tree (DT), several maximum depth and split criterion configurations were examined. Meanwhile, for Multinomial Naive Bayes (MNB), smoothing parameter values (alpha = 0.1, 0.5, 1.0) were evaluated.

The best parameter combinations obtained from GridSearchCV were then used to train each classification model before being integrated into the proposed soft voting ensemble framework. The use of systematic hyperparameter optimization helps improve model generalization capability and ensures that each classifier operates under its most effective configuration for financial condition prediction [30].

4. Results and Discussion

This section presents the evaluation results of the proposed ensemble learning model applied to both structured and unstructured datasets. The evaluation aims to analyze the effectiveness of the soft voting approach in predicting financial conditions based on different types of data sources.

The structured dataset consists of financial indicators derived from company financial reports, which represent quantitative aspects of corporate performance such as profitability, liquidity, and capital structure. These variables provide important financial signals that can be used to detect potential financial distress. Meanwhile, the unstructured dataset consists of textual information obtained from online news related to Indonesian stocks and listed companies. This textual data captures external information such as market sentiment, corporate events, and economic conditions that may influence company performance.

By evaluating both datasets separately, this study aims to examine how the proposed model performs when processing numerical financial indicators as well as textual information. The evaluation results are analyzed using several classification performance metrics, including the confusion matrix, precision, recall, f1-score, and overall accuracy. These metrics provide a comprehensive understanding of the model's predictive capability and its ability to correctly classify financial conditions across different data representations. The detailed evaluation results for each dataset are presented in the following subsections

4.1. Evaluation Result on Structured Data

To evaluate the performance of the proposed ensemble model, experiments were conducted using the structured financial dataset. The model applies the soft voting ensemble technique, which combines the prediction probabilities

ISSN 2723-6471

generated by several base classifiers to produce the final classification result. The evaluation was performed using commonly used classification metrics, including the confusion matrix and classification report. These metrics provide a comprehensive assessment of the model's predictive capability in distinguishing between companies with good and poor financial conditions. The confusion matrix is used to visualize the distribution of correct and incorrect predictions produced by the model, while the classification report summarizes the precision, recall, f1-score, and accuracy values obtained from the prediction results. The evaluation results for the structured dataset are presented in figure 2 and table 1.

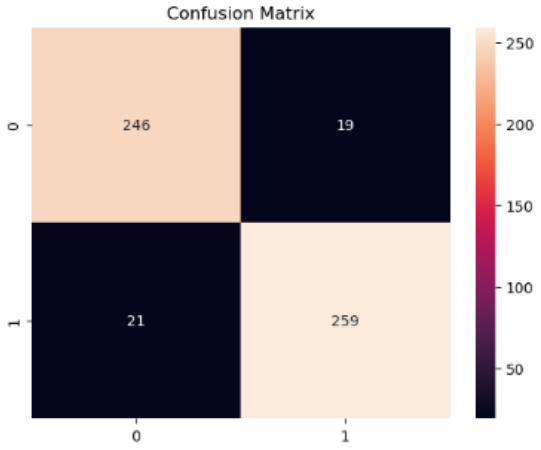

Figure 2. Confusion Matrix of Structured Data Using Soft Voting

The confusion matrix also provides important information regarding class sensitivity and error distribution. In the structured dataset, the model correctly classified most instances in both the "Good" and "Not Good" classes, indicating that the soft voting ensemble model has balanced sensitivity across the two financial condition categories. However, the remaining misclassifications should be interpreted carefully. A false positive prediction may cause a company with an actual "Not Good" financial condition to be classified as "Good," which can lead investors or analysts to underestimate potential financial risk. Conversely, a false negative prediction may classify a financially healthy company as "Not Good," potentially causing unnecessary caution or missed investment opportunities. Therefore, reducing both types of errors is important in financial condition prediction because each error type has different practical implications for investment decision-making and financial risk assessment.

Table 1. Classification Report of the Model

Class / Metric	Precision	Recall	F1-Score	Support
BAD	0.92	0.93	0.92	265
GOOD	0.93	0.93	0.93	280
Accuracy	—	—	0.93	545
Macro Average	0.93	0.93	0.93	545
Weighted Average	0.93	0.93	0.93	545

Table 1 presents the classification report of the soft voting ensemble model applied to the structured dataset. The classification report summarizes several important evaluation metrics, including precision, recall, f1-score, and support for each class. For the "Not Good" class, the model achieved a precision value of 0.92, recall of 0.93, and f1-score of 0.92, with a total of 265 observations. These results indicate that the model is able to correctly identify most of the instances belonging to the negative financial condition category. For the "Good" class, the model achieved precision of 0.93, recall of 0.93, and f1-score of 0.93, with 280 observations. This shows that the model performs consistently when predicting companies with healthy financial conditions.

Overall, the proposed model achieved an accuracy of approximately 0.93 across 545 testing instances. The macro average and weighted average scores for precision, recall, and f1-score are also around 0.93, indicating balanced classification performance across both classes. These results demonstrate that the soft voting ensemble approach is effective in improving classification stability and predictive accuracy for financial distress detection using structured financial data. In addition to evaluating the proposed ensemble voting model, this study also analyzes the performance

 ISSN 2723-6471

of several individual machine learning algorithms used as base classifiers. This evaluation aims to provide a clearer understanding of how each algorithm performs independently when applied to the structured financial dataset. The comparison between individual classifiers and the ensemble voting model helps illustrate the contribution of the ensemble approach in improving prediction performance. Figure 3 is the ROC result of the Soft Voting Ensemble Model Structured Data.

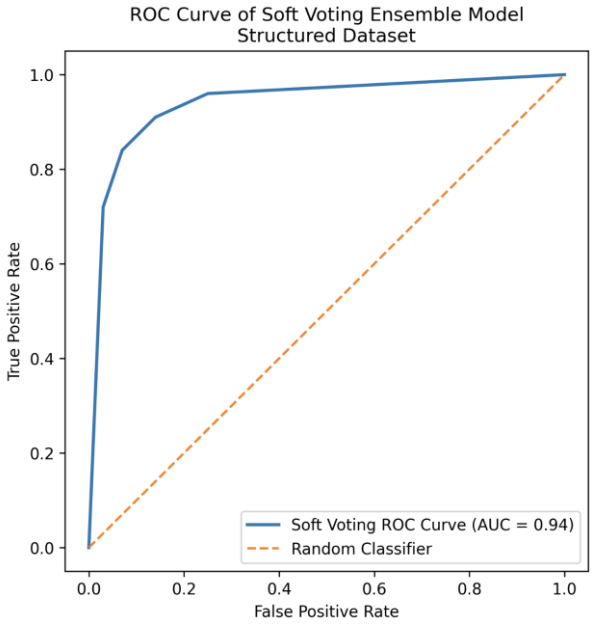

Figure 3. ROC Curve of Soft Voting Ensemble Model Structured Data

The ROC curve for the structured dataset shows the ability of the soft voting ensemble model to distinguish between "Good" and "Not Good" financial conditions based on financial indicators. The curve is positioned far above the diagonal random-classifier line, indicating that the model has good discriminative capability. This means that financial variables such as profitability, liquidity, leverage, ROE, sales growth, DER, and firm value provide useful signals for identifying companies with potential financial risk.

The performance of each classifier is measured using several commonly used classification metrics, namely accuracy, precision, recall, and F1-score. These metrics provide a comprehensive evaluation of the model's capability in correctly identifying the financial condition of companies. The evaluation results of the individual classification models are presented in table 2.

Table 2. Performance Base Classification Models

Model	Accuracy	Precision	Recall	F1 Score
KNN	76%	76%	75%	75%
LR	66%	66%	66%	66%
MNB	59%	60%	58%	56%
RF	87%	87%	87%	87%
SVM	78%	78%	78%	78%
DT	87%	87%	87%	87%

The results indicate that the performance differences among the algorithms are influenced by the characteristics of the structured financial data. Tree-based models such as Random Forest and Decision Tree achieved the highest performance because they are capable of capturing nonlinear relationships and complex interactions among financial indicators. Variables such as profitability, liquidity, leverage, return on equity, sales growth, debt-to-equity ratio, and

 ISSN 2723-6471

firm value may interact in complex ways, making nonlinear classification models more suitable for financial condition prediction.

The lower performance of Logistic Regression suggests that linear decision boundaries may not fully represent the complex relationships contained within financial indicators. Similarly, Multinomial Naive Bayes achieved the weakest performance because its assumption of feature independence is less appropriate for financial ratio data, where several financial indicators may be correlated with one another. Meanwhile, K-Nearest Neighbors produced moderate performance, possibly due to the sensitivity of distance-based methods to overlapping feature distributions and feature scaling.

These findings demonstrate that the effectiveness of machine learning algorithms strongly depends on the characteristics of the financial dataset. Models capable of representing nonlinear patterns tend to perform better for structured financial condition classification tasks. These results highlight the effectiveness of the soft voting ensemble strategy in enhancing financial distress prediction performance compared with individual machine learning classifiers.

4.2. Evaluation Result on Unstructured Data

In addition to evaluating the structured financial dataset, this study also analyzes the performance of the proposed soft voting ensemble model using unstructured textual data. The unstructured dataset consists of news articles related to Indonesian listed companies, which provide additional contextual information that may reflect market sentiment, company performance, and external economic conditions.

To evaluate the effectiveness of the proposed model on textual data, several classification evaluation metrics are used, including the confusion matrix and classification report. The confusion matrix provides a detailed representation of the model's prediction results, while the classification report summarizes important performance metrics such as precision, recall, f1-score, and overall accuracy. The evaluation results for the unstructured dataset are presented in figure 4 and table 3.

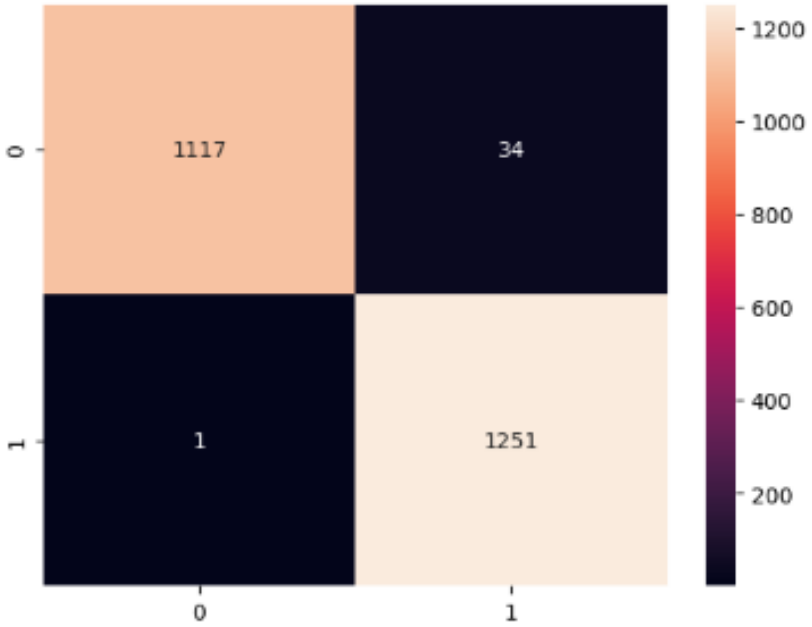

Figure 4. Confusion Matrix of Unstructured Data Using Soft Voting

The confusion matrix for the unstructured dataset indicates that the model achieved very high-class sensitivity, particularly in identifying both positive and negative textual signals from financial news. The small number of misclassified instances suggests that the TF-IDF features extracted from news articles contain strong discriminative patterns for distinguishing favorable and unfavorable company-related information. However, even a small number of false predictions may have practical implications. False positive predictions may cause negative financial news to be interpreted as positive, potentially misleading investors regarding company stability. Meanwhile, false negative

predictions may cause positive news to be classified as negative, which may create unnecessary concern about a company's financial condition. Therefore, the interpretation of the confusion matrix should not only focus on numerical accuracy but also consider how prediction errors may affect financial decision-making and risk perception.

Table 3. Classification Report of Unstructured Data Using Soft Voting

Class / Metric	Precision	Recall	F1-Score	Support
Negative	1.00	0.97	0.98	1,151
Positive	0.97	1.00	0.98	1,252
Accuracy	—	—	0.99	2,403
Macro Average	0.99	0.98	0.99	2,403
Weighted Average	0.99	0.99	0.99	2,403

Table 3 presents the classification report generated from the prediction results of the soft voting ensemble model applied to the unstructured dataset. The classification report provides detailed evaluation metrics including precision, recall, f1-score, and support for each class. For the negative class, the model achieved precision of 1.00, recall of 0.97, and f1-score of 0.98, with a total of 1151 observations. This result indicates that the model is highly accurate in identifying instances belonging to the negative category. For the positive class, the model achieved precision of 0.97, recall of 1.00, and f1-score of 0.98, with 1252 observations. The high recall value indicates that the model is able to correctly identify almost all positive instances within the dataset.

Overall, the model achieved an accuracy of approximately 0.99 across 2403 testing instances. The macro average and weighted average values for precision, recall, and f1-score are also close to 0.99, indicating highly balanced performance across both classes. These results demonstrate that the soft voting ensemble approach provides highly accurate and stable classification performance when applied to unstructured textual data, particularly in capturing patterns from news-based information related to company performance and financial conditions. To further evaluate the predictive capability of the proposed approach, experiments were conducted using the unstructured textual dataset. This dataset consists of news articles related to Indonesian listed companies, which provide contextual information such as market sentiment, corporate events, and other external factors that may influence company performance. The evaluation aims to measure how well individual machine learning algorithms perform when processing textual features derived from the TF-IDF representation. Next, figure 5 is the ROC result of the Soft Voting Ensemble Model on Unstructured Data.

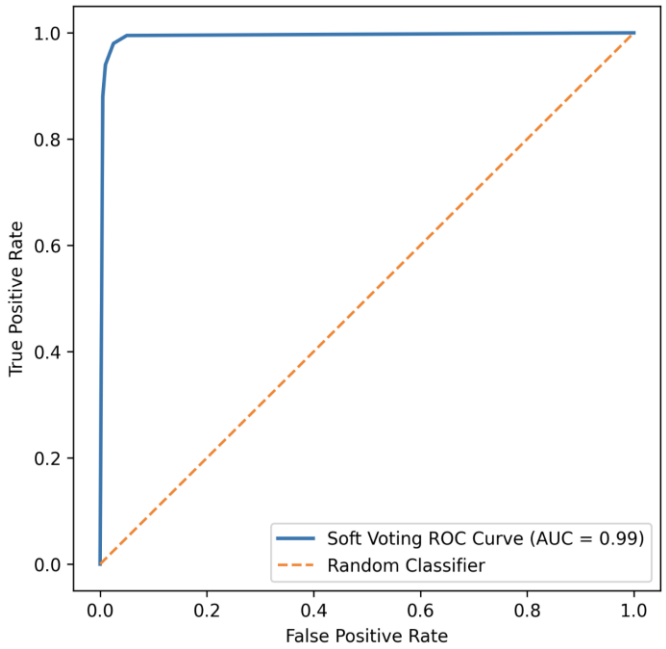

Figure 5. ROC Curve of Soft Voting Ensemble Model Unstructured Data

ISSN 2723-6471

The ROC curve for the unstructured dataset shows stronger discriminative performance than the structured dataset. The curve is closer to the top-left corner, indicating that the model can classify positive and negative financial news with very high accuracy across different thresholds. This suggests that textual features extracted using TF-IDF from news portals contain highly distinguishable patterns, such as positive business performance signals and negative financial risk signals.

Several commonly used machine learning algorithms were implemented as classification models, including K-Nearest Neighbors (KNN), Logistic Regression (LR), Multinomial Naive Bayes (MNB), Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT). The performance of each model was evaluated using four standard classification metrics: accuracy, precision, recall, and F1-score. The results of this evaluation are summarized in table 4.

Table 4. Performance of Individual Classification on Unstructured Data

Model	Accuracy	Precision	Recall	F1 Score
KNN	96%	96%	96%	96%
LR	98.08%	98%	98%	98%
MNB	98.08%	98%	98%	98%
RF	98.46%	99%	98%	98%
SVM	98.50%	99%	98%	98%
DT	97%	97%	97%	97%

The experimental results on the unstructured dataset show that most classification algorithms achieved significantly higher performance compared with the structured dataset. This indicates that textual features extracted from financial news contain highly discriminative information related to company financial conditions. Positive financial news commonly includes terms associated with profit growth, business expansion, improved performance, and favorable market responses, while negative news often contains terms related to financial losses, debt problems, operational disruption, declining performance, or legal issues.

Support Vector Machine achieved the highest performance because it is highly effective in handling high-dimensional sparse textual representations generated by TF-IDF. Logistic Regression also demonstrated strong performance due to the relatively separable textual patterns in TF-IDF feature space. In addition, Multinomial Naive Bayes performed effectively on textual data because probabilistic word-frequency representations are well suited for news classification tasks.

In contrast, K-Nearest Neighbors achieved comparatively lower performance because distance-based methods are generally less efficient in handling sparse and high-dimensional text vectors. The strong overall performance of the unstructured dataset suggests that financial news contains explicit contextual signals that are highly informative for financial condition classification.

Although the soft voting ensemble model achieved very high accuracy on the unstructured dataset, the result should be interpreted carefully. The accuracy approaching 99% indicates that the textual features extracted from financial news provide strong discriminative patterns for classification. However, such high performance may also be influenced by several factors, including similar news writing patterns, repeated company-related terms, duplicated or near-duplicated news content, and highly explicit positive or negative financial expressions in the dataset.

To reduce the possibility of data leakage and overfitting, several precautions were applied during data preparation and model evaluation. Duplicate news records were removed before the preprocessing stage, and the training and testing data were separated before model evaluation to ensure that the same news article did not appear in both sets. Text preprocessing, TF-IDF transformation, and classification were conducted within the experimental workflow to minimize unintended information leakage from the testing data into the training process.

Nevertheless, the unusually high performance on the unstructured dataset remains an important issue that requires careful interpretation. The result may reflect the strong separability of positive and negative financial news in the collected dataset rather than the model's generalization capability across all possible financial news contexts. Therefore, future studies should conduct additional validation using unseen news data from different time periods, different news portals, and more diverse company sectors. Further evaluation using cross-validation, temporal data splitting, and duplicate-similarity checking is also recommended to provide stronger evidence of model robustness.

Based on this consideration, the result of the unstructured dataset should not be overclaimed as definitive evidence of generalizable performance. Instead, it should be understood as an indication that news-based textual features have strong predictive potential for financial condition classification under the dataset and experimental settings used in this study.

4.3. Comparison of Structured and Unstructured Data Testing

The experimental results demonstrate that the proposed soft voting ensemble model performs effectively in predicting financial conditions using both structured financial data and unstructured textual data. The findings indicate that combining multiple machine learning models through a soft voting mechanism improves classification stability and predictive accuracy compared to individual base classifiers.

For the structured dataset, the soft voting model achieved an overall accuracy of approximately 93%, which is higher than the performance of individual models presented in table 1. The confusion matrix results show that the number of correctly classified instances is significantly higher than the number of misclassifications. This indicates that the model can effectively capture patterns within financial indicators such as profitability, liquidity, and capital structure. The improvement over individual models such as KNN, Logistic Regression, and Naive Bayes demonstrates the advantage of the ensemble approach in integrating the strengths of multiple algorithms while reducing their individual weaknesses.

The evaluation results for the unstructured dataset show even higher performance. The soft voting ensemble model achieved an accuracy of approximately 99%, indicating that textual information derived from financial news provides strong predictive signals for classification. The confusion matrix results reveal that the model correctly classified the majority of instances with only a small number of misclassifications. This suggests that textual features extracted using TF-IDF are highly informative in capturing sentiment, corporate events, and market perceptions that may influence company performance.

When comparing the two datasets, the unstructured dataset produces higher classification performance than the structured dataset. One possible explanation is that news-based textual data contain more contextual and real-time information related to company conditions, market reactions, and external economic factors. In contrast, financial ratios in structured data represent historical financial conditions, which may not fully capture external influences affecting company performance.

Additionally, the evaluation of individual classifiers on the unstructured dataset shows that most algorithms already achieve high accuracy above 96%, particularly SVM and Random Forest. However, the soft voting ensemble model further improves the overall prediction performance by aggregating predictions from multiple classifiers. This confirms that ensemble learning is effective in improving model robustness and reducing the risk of model bias.

Overall, the experimental results highlight that the proposed soft voting ensemble approach is capable of producing reliable and accurate predictions for financial distress classification, particularly when integrating different types of data sources. The combination of structured financial indicators and unstructured textual information provides a more comprehensive representation of company conditions, which ultimately improves the predictive capability of the model.

The superior performance of the soft voting ensemble model demonstrates that combining multiple classifiers can improve prediction robustness and reduce the weaknesses of individual algorithms. Each classifier contributes different learning characteristics to the ensemble framework. Tree-based models capture nonlinear feature interactions, linear models capture global decision boundaries, probabilistic models capture word-distribution patterns, and distance-based models capture local similarity among data instances.

By aggregating prediction probabilities from multiple classifiers, the soft voting mechanism produces more stable and reliable predictions compared with relying on a single classification model. These findings indicate that ensemble learning is particularly beneficial when dealing with heterogeneous data representations such as structured financial indicators and unstructured financial news. Therefore, the results should not only be interpreted as numerical performance differences but also as evidence that model effectiveness depends on the interaction between algorithm characteristics and data representation.

5. Conclusion

This study proposes a soft voting ensemble model to predict financial conditions by evaluating structured financial indicators and unstructured textual information derived from financial news. The proposed approach combines multiple machine learning algorithms, including K-Nearest Neighbors, Logistic Regression, Multinomial Naive Bayes, Random Forest, Support Vector Machine, and Decision Tree, to improve prediction robustness and classification accuracy.

The experimental results show that the ensemble voting model performs effectively in both structured and unstructured datasets. For structured financial data, the model achieved an accuracy of approximately 93%, demonstrating its capability to capture financial patterns from company financial reports. For unstructured textual data, the model achieved an accuracy of approximately 99%, indicating that news-based textual information provides strong contextual signals for financial condition prediction.

However, this study does not integrate both data sources into a single multimodal predictive pipeline. Instead, the structured and unstructured datasets are evaluated separately to compare their respective predictive performance. Therefore, the results should be interpreted as evidence of the individual predictive potential of each data source, not as proof that data fusion significantly improves model performance. Future research should develop a true multimodal fusion framework that combines financial indicators and textual news features within one unified prediction pipeline.

Despite the promising results, several practical limitations should be considered. The implementation of this framework in real-world financial decision-making requires continuous data updating, consistent preprocessing, and sufficient computational resources, especially when processing large volumes of online news data. In addition, online news may contain reporting bias, duplicated information, sensational headlines, or speculative content that does not always accurately reflect actual company financial conditions. Therefore, the proposed model should be used as a decision-support tool rather than as the sole basis for investment or financial decisions.

This study is also limited by the scope of the structured dataset, which only includes manufacturing companies listed on the Indonesia Stock Exchange. Therefore, the findings should not be generalized directly to all industry sectors. Future studies should evaluate the proposed framework using broader sectoral datasets, additional financial indicators, and more diverse news sources to improve external validity, robustness, and practical applicability.

6. Declarations

6.1. Author Contributions

Conceptualization: I.R.M., S., and M.T.; Methodology: I.R.M.; Software: I.R.M.; Validation: I.R.M., S., and M.T.; Formal Analysis: I.R.M.; Investigation: I.R.M.; Resources: I.R.M.; Data Curation: I.R.M.; Writing Original Draft Preparation: I.R.M.; Writing Review and Editing: I.R.M., S., and M.T.; Visualization: I.R.M.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. Alfarizi, T. Widiastuti, and Ngatindriatun, "Exploration of Technological Challenges and Public Economic Trends Phenomenon in the Sustainable Performance of Indonesian Digital MSMEs on Industrial Era 4.0," *Journal of Industrial Integration and Management*, vol. 9, no. 1, pp. 65–96, Mar. 2024, doi: 10.1142/S2424862223500045.

[2] B. Dima, Ş. M. Dima, and R. Ioan, "The short-run impact of investor expectations' past volatility on current predictions: The case of VIX," *Journal of International Financial Markets, Institutions and Money*, vol. 98, no. 1, pp. 1–37, Jan. 2025, doi: 10.1016/j.intfin.2024.102084.

[3] N. W. S. Martaningrat and Y. Kurniawan, "The Impact of Financial Influencers, Social Influencers, and FOMO Economy on the Decision-Making of Investment on Millennial Generation and Gen Z of Indonesia," *Journal of Ecohumanism*, vol. 3, no. 3, pp. 1319–1335, Apr. 2024, doi: 10.62754/joe.v3i3.3604.

[4] P. Malhotra, "The rise of passive investing: a systematic literature review applying PRISMA framework," *Journal of Capital Markets Studies*, vol. 8, no. 1, pp. 95–125, Jul. 2024, doi: 10.1108/JCMS-12-2023-0046.

[5] M. S. Murugan and S. K. T, "Large-scale data-driven financial risk management & analysis using machine learning strategies," *Measurement: Sensors*, vol. 27, no. 1, pp. 1–9, Jun. 2023, doi: 10.1016/j.measen.2023.100756.

[6] Y. Song, M. Jiang, S. Li, and S. Zhao, "Class-imbalanced financial distress prediction with machine learning: Incorporating financial, management, textual, and social responsibility features into index system," *Journal of Forecasting*, vol. 43, no. 3, pp. 593–614, Apr. 2024, doi: 10.1002/for.3050.

[7] M. Muniappan and N. D. P. Subramanian, "A Majority Voting Mechanism-Based Ensemble Learning Approach for Financial Distress Prediction in Indian Automobile Industry," *Journal of Risk and Financial Management*, vol. 18, no. 4, pp. 1–31, Apr. 2025, doi: 10.3390/jrfm18040197.

[8] S. Y. Kim and A. Upneja, "Majority voting ensemble with decision trees for business failure prediction during economic downturns," *Journal of Innovation and Knowledge*, vol. 6, no. 2, pp. 112–123, Apr. 2021, doi: 10.1016/j.jik.2021.01.001.

[9] D. Liang, C. F. Tsai, A. J. Dai, and W. Eberle, "A novel classifier ensemble approach for financial distress prediction," *Knowledge and Information Systems*, vol. 54, no. 2, pp. 437–462, Feb. 2018, doi: 10.1007/s10115-017-1061-1.

[10] Y. Ling and P. P. Wang, "Ensemble Machine Learning Models in Financial Distress Prediction: Evidence from China," *Journal of Mathematical Finance*, vol. 14, no. 2, pp. 226–242, 2024, doi: 10.4236/jmf.2024.142013.

[11] M. Nguyen, K. Cao-Van, L. Gia Minh, T. X. Bui, and S. Hong, "Hybrid Machine Learning Models Using Soft Voting Classifier for Financial Distress Prediction," *SSRN*, vol. 1, no. 1, pp. 1–19, 2024, doi: 10.2139/ssrn.4941751.

[12] G. Husain *et al.*, "SMOTE vs. SMOTEENN: A Study on the Performance of Resampling Algorithms for Addressing Class Imbalance in Regression Models," *Algorithms*, vol. 18, no. 1, pp. 1–16, Jan. 2025, doi: 10.3390/a18010037.

[13] K. Alemerien, S. Alsarayreh, and E. Altarawneh, "Diagnosing Cardiovascular Diseases using Optimized Machine Learning Algorithms with GridSearchCV," *Journal of Applied Data Sciences*, vol. 5, no. 4, pp. 1539–1552, Dec. 2024, doi: 10.47738/jads.v5i4.280.

[14] A. A. Pravitasari *et al.*, "STraVEns: Sentence Transformer Voting Ensemble for Intent Classification-Based Chatbot Model," *IEEE Access*, vol. 12, no. December, pp. 197187–197200, 2024, doi: 10.1109/ACCESS.2024.3519223.

[15] A. Ghasemieh, A. Lloyed, P. Bahrami, P. Vajar, and R. Kashef, "A novel machine learning model with Stacking Ensemble Learner for predicting emergency readmission of heart-disease patients," *Decision Analytics Journal*, vol. 7, no. June, pp. 1–13, Jun. 2023, doi: 10.1016/j.dajour.2023.100242.

[16] B. P. Doppala, D. Bhattacharyya, M. Janarthanan, and N. Baik, "A Reliable Machine Intelligence Model for Accurate Identification of Cardiovascular Diseases Using Ensemble Techniques," *Journal of Healthcare Engineering*, vol. 2022, no. March, pp. 1–13, Mar. 2022, doi: 10.1155/2022/2585235.

[17] Y. C. Hua, P. Denny, J. Wicker, and K. Taskova, "A systematic review of aspect-based sentiment analysis: domains, methods, and trends," *Artificial Intelligence Review*, vol. 57, no. 11, pp. 1–51, Nov. 2024, doi: 10.1007/s10462-024-10906-z.

[18] R. J. Dhanal and V. R. Ghorpade, "Aspect-based sentiment analysis using topic modelling and machine learning," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 6, pp. 6689–6698, Dec. 2024, doi: 10.11591/ijece.v14i6.pp6689-6698.

[19] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.

[20] M. K. Anam, T. P. Lestari, H. Yenni, T. Nasution, and M. B. Firdaus, "Enhancement of Machine Learning Algorithm in Fine-grained Sentiment Analysis Using the Ensemble," *ECTI Transactions on Computer and Information Technology*, vol. 19, no. 2, pp. 159–167, Mar. 2025, doi: 10.37936/ecti-cit.2025192.257815.

[21] I. D. Mienye and N. Jere, "A Survey of Decision Trees: Concepts, Algorithms, and Applications," *IEEE Access*, vol. 12, no. June, pp. 86716–86727, Jun. 2024, doi: 10.1109/ACCESS.2024.3416838.

[22] N. W. Susanto and H. Suparwito, "SVM-PSO Algorithm for Tweet Sentiment Analysis #BesokSenin," *Indonesian Journal of Information Systems (IJIS)*, vol. 6, no. 1, pp. 36–47, 2023, doi: 10.24002/ijis.v6i1.7551.

[23] N. Qomariah, "Sentiment Analysis on Coffee Consumer Perceptions on Social Media Twitter Using Multinomial Naïve Bayes," *Journal of Intelligent Computing & Health Informatics*, vol. 2, no. 1, pp. 6–11, Apr. 2021, doi: 10.26714/jichi.v2i1.7241.

[24] A. Desfiandi and B. Soewito, "Student Graduation Time Prediction Using Logistic Regression, Decision Tree, Support Vector, And Adaboost Ensemble Learning," *International Journal of Information System and Computer Science (IJISCS)*, vol. 7, no. 3, pp. 195–199, 2023, doi: 10.56327/ijiscs.v7i2.1579.

[25] O. C. B. Omankwu, E. C. Osodoeke, and V. I. Ubah, "Disease Prediction Using Random Forest Machine Learning Algorithm," *NIPES – Journal of Science and Technology Research*, vol. 6, no. 1, pp. 174–181, Feb. 2024, doi: 10.5281/zenodo.11005308.

[26] X. T. Dang and T. T. Le, "KNN-SMOTE: An Innovative Resampling Technique Enhancing the Efficacy of Imbalanced Biomedical Classification," in *Machine Learning and Other Soft Computing Techniques: Biomedical and Related Applications*, N. Hoang Phuong, N. T. Huyen Chau, and V. Kreinovich, Eds. Cham: Springer Nature Switzerland, 2024, pp. 111–121, doi: 10.1007/978-3-031-63929-6_11.

[27] K. Cao-Van, T. C. Minh, L. G. Minh, T. T. B. Quyen, and H. M. Tan, "Soft-Voting Ensemble Model: An Efficient Learning Approach for Predictive Prostate Cancer Risk," *Vietnam Journal of Computer Science*, vol. 11, no. 4, pp. 531–552, Nov. 2024, doi: 10.1142/S2196888824500155.

[28] H. G. Jabbar, "Advanced Threat Detection Using Soft and Hard Voting Techniques in Ensemble Learning," *Journal of Robotics and Control (JRC)*, vol. 5, no. 4, pp. 1104–1116, 2024, doi: 10.18196/jrc.v5i4.22005.

[29] M. I. Elim and E. Utami, "Performance Comparison of Child Stunting Prediction Support Vector Machine vs Random Forest with Grid Search Optimization," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 5, pp. 5305–5319, Oct. 2025, doi: 10.52436/1.jutif.2025.6.5.5285.

[30] A. Efendi, I. Fitri, and G. W. Nurcahyo, "Improving Student Graduation Timeliness Prediction Using SMOTE and Ensemble Learning with Stacking and GridSearchCV Optimization," *Data and Metadata*, vol. 4, no. 1, pp. 1–10, Jan. 2025, doi: 10.56294/dm2025917.