Asia Pacific Fraud Journal
E-ISSN: 2502-695X, ISSN: 2502-8731
Volume 10, Issue 1 (January-June) 2025
*Asia Pacific Fraud Journal now has been accredited "**SINTA 3**" by Ministry of Research and Technology of The Republic of Indonesia (No SK. 225/E/KPT/2022).*
Available online at: http://apfjournal.or.id/index.php/apf

Reading Between the Lines: Incorporating Text Mining and Machine Learning in Financial Fraud Detection

✉Agung Septia Wibowo & Iis Istianah
Universitas Gadjah Mada, Indonesia

ARTICLE INFORMATION
Article History:
Received November 13, 2024
Revised May 20, 2025
Accepted June 10, 2025

DOI:
10.21532/apfjournal.v10i1.382

ABSTRACT
Notwithstanding rigorous oversight in the Indonesian capital market, the manipulation of financial reports continues to occur. This study examines the potential for employing machine learning (ML) models, which utilize linguistic features and financial ratios, in effectively detecting deception or manipulation. Drawing upon publicly listed Indonesian companies as the samples, this research validates the predictive capabilities of the Beneish M-Score, confirms the occurrence of negative language in fraudulent reports, and demonstrates the superiority of the Gradient Boosting ML model in identifying anomalies within financial and textual data. The study distinctively adapts to Indonesian-language annual reports, thereby addressing a gap in the linguistic-based fraud detection literature. These findings not only advance our comprehension of how linguistic features and financial ratios provide practical tools for fraud detection, thereby preparing the academic and professional community in this domain.

Keywords: *Financial Statement Fraud, Text Mining, Machine Learning (ML), Natural Language Processing (NLP), Financial Ratios.*

How to Cite:
Wibowo A. S., & Istianah I. (2025). Reading Between the Lines: Incorporating Text Mining and Machine Learning in Financial Fraud Detection. *Asia Pacific Fraud Journal, 10*(1), 73-93. http://doi.org/10.21532/apfjournal.v10i1.382.

✉Corresponding author :
Email: agungseptiawibowo@gmail.com

**The 2nd Best Paper of National Call for Paper ACFE Indonesia Chapter 2024*

Association of Certified Fraud Examiners (ACFE) Indonesia Chapter
Page. 73-93

1. INTRODUCTION

Financial statement fraud (FSF) has far-reaching consequences. It impacts stakeholders both financially and emotionally, affecting entire communities. When financial reports are manipulated, the resulting information becomes inaccurate, leading to misguided decisions. The damage can be severe, ranging from lost of public trust to corporate bankruptcy. Research by Dechow et al. (2010) shows that manipulating financial statements often reduces share value and increases market volatility, ultimately harming shareholders and financial stability. Aghghaleh et al. (2016) found that financial statement fraud results in significant financial losses, with organizations losing an average of 6% of their annual revenue. High-profile cases-including Enron, WorldCom, and Satyam-caused losses exceeding US$20 billion, underscoring the urgent need for preventive measures. Hogan et al (2008) highlighted those 30 financial scandals led to a loss of more than $900 billion in market capitalization. Proactive steps, such as independent audits, strict internal controls, and whistleblower hotlines, are crucial. According to "Occupational Fraud 2024: A Report to the Nations," Financial Statement Fraud has a median loss of $766,000 per case, making it the least common but most expensive category of fraud (ACFE, 2024).

FSF is often concealed from the public and auditors, making prompt detection challenging (Aghghaleh et al., 2016). The complexity of manipulating financial data requires advanced detection tools (Aghghaleh et al., 2016). While three 'red flags' do not always indicate fraud (Hogan et al., 2008), perpetrators' adaptability underscores the urgency for better detection methods (Zhou & Kapoor, 2011). Analyzing financials with the Beneish M-Score can reveal manipulation. This model uses ratios to identify anomalies, differentiating between manipulated and accurate reports (Beneish, 1999). Natural Language Processing (NLP) advancements also offer opportunities to detect false reports by analyzing language and sentiment patterns (Saquete et al., 2020; Faccia et al., 2024).

Prior research has employed financial ratios such as the Beneish M-Score and natural language processing (NLP) for fraud detection (Beneish, 1999; Loughran & McDonald, 2011). However, these studies had not yet integrated linguistic features alongside financial indicators. Furthermore, Ribeiro et al. (2016) illustrated that machine learning models possess the capability to unveil intricate patterns that remain undetected through conventional analyses, thereby fulfilling the requirements for FSF detections. The integration of these elements could potentially enhance the accuracy and efficiency of detecting financial statement manipulation. This integration could help stakeholders proactively identify and address financial fraud before its impact spreads. In addition, most of the literature uses English-language data, such as LM Dictionary and NRC Lexicons that pose challenges in directly applying linguistic analysis of different languages. Therefore, contextual adaptation is needed in the linguistic analysis of Indonesian-language financial statements.

This study propose a comprehensive fraud detection framework for Indonesian capital market companies. The research finds that the Gradient Boosting ML model offers superior prediction accuracy and consistency. Moreover, combining linguistic features with financial ratios improves model performance, showing this approach's effectiveness in fraud identification. Furthermore, sentiment analysis reveals that negative words and emotions like fear and sadness are common in reports from fraudulent companies, indicating negative sentiment as a key fraud indicator. Additionally, fraudulent companies often avoid strong language in reports to evade scrutiny.

This research demonstrates significant advancement over previous studies by effectively integrating linguistic analysis with financial forensics to

detect manipulation in corporate communications. This study validates the Beneish M-Score's predictive capabilities while simultaneously employing machine learning to identify subtle anomalies across both financial and textual data. Furthermore, the study approach in conducting careful contextual adaptation to Indonesian-language annual reports, signify innovation that addresses a critical gap in fraud detection literature which has predominantly focused on English-language materials. By developing language-specific NLP techniques that account for cultural and linguistic nuances in Indonesian corporate communications, the study offers practical applications for regulators, investors, and auditors operating in emerging markets where such sophisticated detection tools were previously unavailable or underdeveloped.

2. LITERATURE REVIEW AND HYPOTHESIS

FSF is the deliberate manipulation of financial statements to present a false image of a company's financial status (Hogan et al., 2008; Kanapickiene & Grundienė, 2015). This fraudulent act often carried out through misrepresenting or omitting information in financial reports is a calculated deception to gain unauthorized or illegal financial benefits (Gotelaere & Paoli, 2022; Soltani et al., 2023). It is a material omission or misrepresentation, a serious violation of generally accepted accounting principles, due to a deliberate failure to report financial information (Craja et al., 2020; Hajek & Henriques, 2017). FSF can be seen in the form of Fictitious Income and Sales, which is essentially Income that has not been fully earned (Aghghaleh et al., 2008). Another form is False Revenue Recording, where products that were not ordered are sent and charged (Craja et al., 2020). Asset Manipulation is another type, involving the Manipulation of asset values to enhance financial reports (Dong et al., 2016). Changes in accounting records, which involves altering records to hide irregularities, and overly high company valuations, which inflate company worth are also common types of financial statement fraud (Gotelaere & Paoli, 2022; Hajek & Henriques, 2017).

One highly effective and adaptable technique in the battle against FSF is financial text analysis. This method, which utilizes text mining techniques, is adept at detecting fraud in financial reports. By analyzing textual content such as Management Discussion and Analysis (MD&A) sections and managerial comments in typical annual reports (see Craja et al., 2020; Dong et al., 2016; Hajek & Henriques, 2017), it extracts valuable information and patterns from finance-related texts, such as financial reports, emails, and other documents. Financial text analysis, with its significant benefits, is a powerful tool in detecting financial statement fraud. Research demonstrates that the combination of text features and financial ratios can significantly enhance fraud detection accuracy (Craja et al., 2020). Moreover, text analysis-based methods outperform traditional methods based on financial ratios (Dong et al., 2016). By enabling early fraud detection, text analysis provides crucial support for auditors in their decision-making (Hajek & Henriques, 2017). The use of features from multiple categories in text analysis further enhances overall financial fraud (Throckmorton et al., 2015).

Financial text analysis is based on Systemic Functional Linguistics (SFL) Theory by Halliday (2014). This approach emphasizes that language is a system used to create meaning in a social context and provides a comprehensive understanding of language functions. SFL's focus on the paradigmatic axis in understanding how language functions is crucial to its comprehensive nature. In SFL, every act of communication involves a choice, and SFL maps the choices available in various language variations using a representation tool as a 'network of systems.' Metafunction in SFL refers to organizing a functional framework around a choice system.

Three metafunctions are simultaneously involved in all languages, and SFL plays a key role in providing a comprehensive framework for understanding language, ensuring clarity, interpersonal, and textual (Halliday & Matthiessen, 2014).

3. METHODS

We use financial data and textual data from Indonesian sample firms. Financial ratios and numbers are derived from firms' financial statements; meanwhile, for the textual data, we obtained data from 404 corporate annual reports from the Indonesian Exchange (IDX) and company websites. From the initial 404 firm-year data, we further checked for missing data and potential outliers from the sample, resulting in the final sample of 351 firm-year data. The reports in the sample were meticulously grouped into two, namely the fraudulent group identified by sanctions and under special supervision related to financial reports from IDX in the reporting period. In contrast, the next group was non-fraudulent reports, with no cases related to financial reports in the reporting year. The details of the samples based on the firm sector are presented in Table 1.

We use the Loughran-McDonald Financial Dictionary (LM Dictionary) to assess sentiment and extract information from corporate annual reports. This dictionary was designed to handle financial texts that often have unique characteristics compared to general texts (Loughran & McDonald, 2011). Furthermore, we also use emotion lexicons from the National Research Council Canada (NRC), which provides a list of keywords to analyze emotions in text using emotion lexicons (Mohammad & Turney, 2013). As the annual reports are all in Bahasa Indonesia, The LM Dictionary and NRC lexicons need to be translated first. Additionally, each category in the LM Dictionary and NRC Lexicons was mapped into the ideational metafunction of the SFL framework, whereas the Sentiment information type was derived from LM Dictionary word categories, while the Emotions information type was derived from NRC Lexicons groups. We assigned an interpersonal information type adopting the approach of Dong et al. (2016), where FirstPerson and ThirdPerson categories were included in this group. We add society word features in the context as we find that the words "masyarakat" (society) and "Indonesia" were among the words that were highly mentioned in the company's annual reports. Therefore, we suggest that the analysis of this word category provides additional value to our research and opportunity for further analysis. Furthermore, we also follow textual metafunctions approach by Dong et al. (2016) where the total number of words in the reports provides a proxy for textual information type. Table 2 shows the details of the linguistic features derived based on the SFL framework.

We also gathered financial data from firms' financial reports to be analyzed with Beneish's M-Score, a mathematical model to identify whether a company has manipulated its financial statement. The Beneish ratio is a series of indicators used to detect possible manipulation of financial statements by companies. These

Table 1. **Final Report Samples by Sectors**

Sector	Non-Fraudulent Report	Fraudulent Report	All Report
Consumer Cyclicals	111	26	137
Infrastructures	64	10	74
Consumer Non-Cyclicals	46	8	54
Energy	30	13	43
Other Sectors	27	16	43
Total	278	73	351

Source: Processed Data

ratios include Days Sales in Receivables Index (DSRI), Gross Margin Index (GMI), Asset Quality Index (AQI), Sales Growth Index (SGI), Depreciation Index (DEPI), Sales, General, and Administrative Expenses Index (SGAI), Leverage Index (LVGI), and Total Accruals to Total Assets (TATA). Details of measures to calculate the ratios are provided in Table 3. DSRI measures changes in the receivables to

Table 2. **Features Based on SFL Metafunction**

Metafunction	Information Type	Features	Feature Explanation
Ideational	Sentiments	Negative	Total number of negative sentiment words divided by total number of words.
		Positive	Total number of positive sentiment words divided by total number of words.
		Uncertainty	Total number of uncertainty sentiment words divided by total number of words.
		Strong_Modal	Total number of strong modal sentiment words divided by total number of words.
		Weak_Modal	Total number of weak modal sentiment words divided by total number of words.
		Litigious	Total number of litigious sentiment words divided by total number of words.
		Constraining	Total number of constraining sentiment words divided by total number of words.
	Emotions	Anticipation	Total number of anticipation emotion words divided by total number of words.
		Joy	Total number of joy emotion words divided by total number of words.
		Trust	Total number of trust emotion words divided by total number of words.
		Surprise	Total number of surprise emotion words divided by total number of words.
		Anger	Total number of anger emotion words divided by total number of words.
		Disgust	Total number of disgust emotion words divided by total number of words.
		Fear	Total number of fear emotion words divided by total number of words.
		Sadness	Total number of sadness emotion words divided by total number of words.
Interpersonal	Personal Pronoun	FirstPerson	Total number of first person singular pronouns divided by total words in documents.
		ThirdPerson	Total number of all other person pronouns divided by total words in documents.
		Society	Total number of references to society divided by total words in documents.
Textual	Writing Style	LnTotal	Natural Log (the number of words in documents)

Source: Processed Data

sales ratio, where a higher value may indicate manipulation to increase sales. GMI measures changes in gross profit margin, and a higher value could indicate a decrease in profit margin, which may be a sign of manipulation to maintain profits. AQI measures changes in asset quality, indicating an increase or decrease in low-quality assets. SGI measures sales growth, where high growth can encourage companies to manipulate to meet market expectations. The Depreciation Index (DEPI) measures changes in the depreciation rate and could indicate a possible decrease in the depreciation rate to delay expense recognition. The SGAI measures changes in selling, general, and administrative costs relative to sales, where increases in these costs could indicate manipulation to cover a decline in profitability. LVGI measures changes in financial leverage and increases in leverage can indicate financial stress that drives reporting manipulation. TATA measures total accruals relative to total assets, where higher values indicate an increase in accruals, which could be a sign of earnings manipulation. A combination of these ratios is used in the Beneish M-Score model to indicate whether a company's financial statements have been manipulated, with a BM-Score value greater than -2.22 indicating a greater likelihood that the financial statements have been manipulated.

Tabel 3. **Beneish M-Score Ratios**

Ratio	Measure	Formula
DSRI	changes in the accounts receivable to sales ratio	(This Year's Receivables / This Year's Sales) / (Last Year's Receivables / Last Year's Sales)
GMI	changes in gross profit margin	[(Last Year's Sales - Last Year's Cost of Goods Sold) / Last Year's Sales] / [(This Year's Sales - This Year's Cost of Goods Sold) / This Year's Sales]
AQI	changes in asset quality	[1 - (Current Assets + PP&E + Investments and Other Assets This Year) / Total Assets This Year] / [1 - (Current Assets + PP&E + Investments and Other Assets Last Year) / Total Assets Last Year]
SGI	sales growth	This Year's Sales / Last Year's Sales
DEPI	changes in depreciation rates	(Last Year's Depreciation Expense / (Last Year's Depreciation Expense + Last Year's PP&E)) / (This Year's Depreciation Expense / (This Year's Depreciation Expense + This Year's PP&E))
SGAI	changes in selling, general, and administrative costs relative to sales	(SG&A This Year / Sales This Year) / (SG&A Last Year / Sales Last Year)
LVGI	changes in financial leverage	(Total Debt This Year / Total Assets This Year) / (Total Debt Last Year / Total Assets Last Year)
TATA	total accruals relative to total assets	(Net Profit - Operating Cash Flow) / Total Assets
BM-Score	Overall manipulation score	= -4.84 + 0.92*DSRI + 0.528*GMI + 0.404*AQI + 0.892*SGI + 0.115*DEPI – 0.172*SGAI + 4.679*TATA – 0.327*LVGI

Source: Processed Data

We utilize ML algorithms to develop and evaluate the classification models. ML uses algorithms to find patterns in the data and, intriguingly, provides predictions or judgments without the need for explicit programming (Hajek & Henriques, 2017; Perols, 2010). ML works by identifying patterns in data without guidance from human analysts or experts. This technique helps detect and prevent fraud by allowing automatic pattern recognition in large amounts of data (Ashtiani & Raahemi, 2022). In this study, we utilize seven ML models. The first three models, Logistic Regression, k-nearest Neighbors (kNN), and Naive Bayes, are classic classification models known for their reliability. Logistic Regression uses a logit function to map inputs to binary outputs, while kNN, a non-parametric algorithm, uses the entire dataset as a model. The classification is based on most labels from the nearest neighbors of the data to be classified (Cover & Hart, 1967). On the other hand, the Naive Bayes algorithm family is exceptionally reliable for text categorization, relying on applying Bayes' Theorem and the assumption of feature independence and compatibility (McCallum & Nigam, 1998). Furthermore, we also employ ensemble models, a technique that enhances predictive performance by combining multiple models that consistently outperforms a single model. The ensemble models we use are Adaboost, Gradient Boosting, and Random Forest. Adaboost (Adaptive Boosting) is a boosting algorithm that aims to enhance the performance of weak models (Freund & Schapire, 1997). It assigns greater weight to errors from previous iterations, focusing on challenging data. Gradient Boosting is a boosting technique that constructs a model incrementally by optimizing the loss function using gradient descent (Friedman, 2001). Each new model is designed to rectify the errors of the previous model, making it an iterative process. Random Forest aggregates forecasts from several decision trees trained on various subsets of data (Breiman, 2001). Finally, the last ML model is Neural Network, which draws inspiration from biological neural networks. It is made up of layers of networked neurons that, with training, can be trained to represent complex data (LeCun et al., 2015). Neural Networks, as the basis model of deep learning, is superior in term of their flexibility in modeling complex non-linear relationships in data enables them to handle high-dimensional data, capture complex and abstract patterns, and generate new data that resembles the training data and exhibits creative variations (Goodfellow et al., 2016).

4. RESULTS AND DISCUSSION

Descriptive Analysis

Descriptive statistics provides an overview of the basic characteristics of the dataset, which helps identify patterns, trends, and data distribution. Furthermore, it can reveal data errors or important information, providing a crucial basis for further statistical analysis. It also offers an initial understanding of the relationships between variables, enlightening us and enhancing our knowledge of the data.

Table 4 presents descriptive statistics for various textual features analyzed in the 351 samples. The Negative feature averages 556 words, or the highest mean among the sentiment category types. Meanwhile, the standard deviation is 367, indicating significant variation in the number of negative words between samples. The Trust feature has the overall highest mean among emotion features with an average of 4,135 words, with a standard deviation of 2,954, indicating that words reflecting trust vary widely in the analyzed texts. The Anticipation and Joy features also show relatively high frequencies, with an average of 2,573 and 1,271 words, respectively. The personal Pronoun that shows the most significant mean was Society with 907, followed by First Person with 791; meanwhile, Third Person words were mentioned relatively low with only a mean of 108. Lastly, Total Words summarizes the total words

Table 4. **Descriptive Statistics of Textual Features**

Information Type	Linguistics Features	Mean	Median	Std Error	Std Dev	N
Sentiments	Negative	556	439	20	367	351
	Positive	302	259	18	337	351
	Uncertainty	187	172	10	186	351
	Strong_Modal	122	110	6	120	351
	Weak_Modal	18	18	1	11	351
	Litigious	252	208	14	254	351
	Constraining	84	69	4	84	351
Emotions	Anticipation	2,573	2,116	97	1,823	351
	Joy	1,271	1,028	45	842	351
	Trust	4,135	3,372	158	2,954	351
	Surprise	453	339	20	377	351
	Anger	489	311	23	423	351
	Disgust	302	151	17	311	351
	Fear	845	583	36	672	351
	Sadness	746	528	29	538	351
Personal Pronoun	FirstPerson	791	724	22	405	351
	ThirdPerson	108	101	3	57	351
	Society	907	809	25	476	351
Writing Style	TotalWords	144,164	125,337	4,777	89,501	351

Source: Processed Data

analyzed in each sample, with a mean of 144,164 words and a standard deviation of 89,501, indicating large variability in text length.

Table 5 presents the descriptive statistics for various Beneish ratios, which are of utmost importance in detecting possible FSF based on 351 samples. The DSRI, with a mean of 1.080 and a standard deviation of 0.844, reveals a considerable variation in the receivables-to-sales ratio. The GMI, with a mean of 0.949 and a standard deviation of 2.129, shows some outliers with significant changes in gross profit margin. The AQI, with a mean of 0.923 and a standard deviation of 1.688, indicates significant variation in asset quality. SGI demonstrates relatively stable sales growth, with an average of 1.202 and a standard deviation of 0.531. DEPI, with the highest mean of 1.333 and a standard deviation of 1.885, indicates significant variations in depreciation rates. The SGAI and LVGI, with averages close to 1 (1.025 and 1.003, respectively) and low standard deviations, suggest relative stability in SG&A expenses and financial leverage. TATA (Total Accruals to Total Assets) averages -0.047, indicating a negative accrual tendency, with a standard deviation of 0.253. The Beneish M-Score, with a mean of -2.472 and a standard deviation of 1.996, reveals significant variations in indications of possible financial statement manipulation. This data provides a comprehensive overview of the distribution and variability of each Beneish ratio in the sample, which is invaluable in the detection of potential FSF.

Table 5. **Descriptive Statistics of Beneish Features**

Beneish ratios	Mean	Median	Std Error	Std Dev	N
DSRI	1.080	0.936	0.045	0.844	351
GMI	0.949	0.985	0.114	2.129	351
AQI	0.923	1.050	0.090	1.688	351
SGI	1.202	1.089	0.028	0.531	351
DEPI	1.333	1.038	0.101	1.885	351
SGAI	1.025	0.976	0.023	0.422	351
LVGI	1.003	0.978	0.014	0.263	351
TATA	-0.047	-0.023	0.013	0.253	351
BMScore	-2.472	-2.484	0.107	1.996	351

Source: Processed Data

Table 6. **Factor Analysis**

Lingustic Type	Factors	Component	
		1: Ideational	2: Interpersonal
Sentiment	Negative	**0.812**	0.346
	Positive	**-0.806**	-0.167
	Uncertainty	**-0.861**	0.178
	Strong_Modal	**-0.845**	-0.012
	Weak_Modal	**-0.632**	0.293
	Litigious	**-0.850**	0.081
	Constraining	**-0.825**	0.062
Emotion	Anticipation	**0.900**	-0.019
	Joy	**0.892**	0.061
	Trust	**0.911**	0.012
	Surprise	**0.848**	-0.020
	Anger	**0.963**	0.162
	Disgust	**0.948**	0.061
	Fear	**0.943**	0.133
	Sadness	**0.875**	0.307
Pronouns	FirstPerson	-0.022	**0.918**
	ThirdPerson	-0.155	**0.748**
	Society	0.348	**0.705**

Source: Processed Data

Factor Analysis

When we use lexicons with numerous categories in our text analysis, the resulting variables can be quite complex. Factor analysis is a powerful tool that simplifies these complex variables and creates a more manageable model. It effectively reduces data dimensions, compressing a large number of variables into fewer factors. Importantly, it still manages to explain

most of the variance in the original data, thanks to its ability to reveal the latent factors underlying these variables (Hair et al., 2010).

Principal Component Analysis (PCA) is used to reduce data dimensionality based on factor data extracted from annual report texts. By identifying the underlying factors, factor analysis with PCA can reduce the number of variables that must be analyzed, simplifying the data structure without losing important information. The results of PCA, presented in Table 6, reveal two components: 'Ideational' for component 1 and 'Interpersonal' for component 2. All of the LM Dictionary and NRC lexicon word categories are significantly mapped into the Ideational component, while Pronoun words are mapped into the Interpersonal component. This consistent mapping with the SFL Framework provides a robust foundation for further analysis, opening up exciting possibilities for future research in the field of linguistics and text analysis.

Classification Model Performances

The classification models built for the analysis were based on the two classes for the target variable: *non-fraud* class and *fraud* class. In a comprehensive testing process, we examined the use of BM-Score, Textual features, and the combination of M-Score and textual features to analyze the effect of each approach in predicting the class in the target variable. This thorough testing process instills confidence in the robustness of our research methodology.

Standard evaluation metrics like classification accuracy (CA), precision (Prec), recall, F1 score, the area under the receiver-operating curve (AUC), and Matthews correlation coefficient (MCC) are used to analyze the performance of the ML models built. CA provides a general idea of how often the model makes correct predictions (Aronoff, 1982), while Prec is a metric that measures the proportion of optimistic predictions that are genuinely positive. Precision is essential when the cost of false positives is high, such as in disease detection (Hicks et al., 2022). Recall (also called Sensitivity or True Positive Rate) measures the proportion of positive cases the model successfully identifies. Recall is necessary when the cost of false negatives is high, as in cancer detection (Hicks et al., 2022).

F1 Score harmoniously combines Precision and Recall, offering a balanced assessment of model performance, mainly when either metric alone is insufficient (Jaballi et al., 2024). Furthermore, AUC measures a model's ability to differentiate between classes and provides a general idea of the model's performance at various thresholds, with higher values indicating better discrimination capabilities (Lobo et al., 2008). Lastly, MCC is a metric that measures the correlation between predictions and actual values. It provides a more balanced assessment, even on imbalanced data, which can be reassuring in the face of skewed datasets (Chicco & Jurman, 2023). All of the metrics range

Table 7. **Classification Performance with BM-Score as Feature**

Model	AUC	CA	F1	Prec	Recall	MCC
AdaBoost	0.802	0.866	0.866	0.865	0.866	0.607
Gradient Boosting	0.802	0.858	0.849	0.849	0.858	0.551
kNN	0.720	0.769	0.750	0.741	0.769	0.237
Logistic Regression	0.591	0.795	0.729	0.782	0.795	0.221
Naive Bayes	0.573	0.781	0.684	0.609	0.781	-
Neural Network	0.581	0.798	0.738	0.782	0.798	0.244
Random Forest	0.801	0.835	0.828	0.825	0.835	0.484

Source: Processed Data

from 0 to 1 with 0 indicating a random prediction, and 1 indicating a perfect prediction, except for MCC that ranges from -1 to 1, where -1 indicates a prediction opposite to the actual value.

Table 7 provides results on the classification performance based on BM-Score as a single feature to predict the probability of fraudulent reports. The AdaBoost algorithm performs the best, with all metrics topping the chart. An Area Under the Curve (AUC) of 0.802 indicates the robust ability to differentiate between different classes. With Accuracy (CA) and F1 Score reaching 0.866 and Precision and Recall almost identical at 0.865 and 0.866, AdaBoost balances positive and negative predictions. Matthew's Correlation Coefficient (MCC) of 0.607, while not as high as other metrics, still indicates good consistency in the overall performance of this model. Gradient Boosting, on the other hand, also showed strong performance with the same AUC of 0.802, indicating comparable discrimination capabilities to AdaBoost. The CA metrics of 0.858 and F1 Score of 0.849 indicate slightly lower performance than AdaBoost in terms of accuracy and agreement between Precision (0.849) and Recall (0.858). However, this model does not have MCC values provided, making it difficult to assess the overall consistency of performance in the context of the relationship between positive and negative predictions.

Table 8 displays the evaluation results of the classification model from the linguistic features. The Random Forest model shows the best performance with an AUC value of 0.864, while the CA (0.886), F1 Score (0.880), Precision (0.882), and Recall (0.886) values also show that this model is not only accurate but also consistent in its predictions. The MCC of 0.644 strengthens this result by showing a strongest correlation between predicted and actual values among all models. The Gradient Boosting and AdaBoost algorithms also show competitive performance, with AUCs of 0.844 and 0.826, respectively. However, Gradient Boosting is slightly behind in F1 Score (0.860) and Precision

Table 8. **Classification Performance with Texts Features**

Model	AUC	CA	F1	Prec	Recall	MCC
AdaBoost	0.826	0.866	0.869	0.873	0.866	0.626
Gradient Boosting	0.844	0.866	0.860	0.859	0.866	0.582
kNN	0.760	0.778	0.760	0.753	0.778	0.270
Logistic Regression	0.626	0.772	0.680	0.608	0.772	-0.049
Naive Bayes	0.634	0.764	0.724	0.713	0.764	0.143
Neural Network	0.710	0.786	0.703	0.779	0.786	0.138
Random Forest	0.864	0.886	0.880	0.882	0.886	0.644

Source: Processed Data

Table 9. **Classification Performance with Combined Features**

Model	AUC	CA	F1	Prec	Recall	MCC
AdaBoost	0.842	0.892	0.892	0.892	0.892	0.684
Gradient Boosting	0.846	0.915	0.910	0.914	0.915	0.737
kNN	0.777	0.801	0.788	0.783	0.801	0.359
Logistic Regression	0.645	0.781	0.716	0.728	0.781	0.138
Naive Bayes	0.638	0.775	0.742	0.736	0.775	0.207
Neural Network	0.728	0.815	0.769	0.812	0.815	0.345
Random Forest	0.870	0.892	0.887	0.888	0.892	0.665

Source: Processed Data

(0.859) compared to AdaBoost, which has an F1 Score of 0.869 and Precision of 0.873. AdaBoost's MCC of 0.626, indicating that its overall performance comfortably put it in as the second best model.

Based on the evaluation results in Table 9, the overall performance of the combined features exceeded the performance of the model with BM-Score and linguistics features alone. The Gradient Boosting model shows a superior performance among all the tested algorithms with the metrics of AUC (0.846), CA (0.915), F1 Score (0.910), Precision (0.914), and Recall (0.915), tops the chart and show that this model shows a strong ability to distinguish between different classes and consistent in making predictions. Matthew's Correlation Coefficient (MCC) of 0.737 also confirms that this model has a satisfactory correlation between predictions and actual values, indicating high stability and reliability in its performance. The AdaBoost and Random Forest models also show competitive performance, although slightly below Gradient Boosting. Overall, the results with the combined features found to be better in all of the models, indicating the value of combination of linguistic features and financial ratio features.

Additionally, we conduct a T-test of differences between the results of the classification models to investigate the significant effect of linguistics features and financial ratios features in the combined models. Table 10 shows that most algorithms significantly differed when linguistics features and BM Score were combined. This shows that the combination of text and financial ratios features contributes to improving classification model performance. They both provide significantly different effects that contribute to the combined model generated.

Feature Importance

We conduct feature importance analysis to identify which feature has significant influence in the model. As shown in Figure 1, the feature importance of the Gradient Boosting model, as the best performance model, was analyzed based on the AUC metrics. The Interpersonal SFL construct (PERS_Factor), has the most significant influence on the model, as indicated by the longest bar. Its removal results in the most significant decrease in AUC, highlighting its importance in the model predictions. Equally essential is the LnTotal feature, which, if removed, causes a significant decrease in AUC. BMScore, with its moderate influence on the model, causes a decrease in AUC when removed, but less than the previous two features. Lastly, Ideational construct (IDEA Factor), has a minor influence on the model, as indicated

Table 10. **T-Test of Different two-tail Between Feature Types**

Model	BMScore-Linguistics **P(T<=t)**	p-val	BMScore-Combined **P(T<=t)**	p-val	Linguistics-Combined **P(T<=t)**	p-val
AdaBoost	0.084	*	0.007	***	0.006	***
Gradient Boosting	0.027	**	0.015	**	0.033	**
kNN	0.021	**	0.013	**	0.024	**
Logistic Regression	0.132		0.324		0.089	*
Naive Bayes	0.103		0.078	*	0.057	*
Neural Network	0.843		0.049	**	0.083	*
Random Forest	0.009	***	0.010	***	0.017	**

*p-val of *, ** and *** represent statistical significance at 10%,rate of 5% and 1% level.*
Source: Processed Data

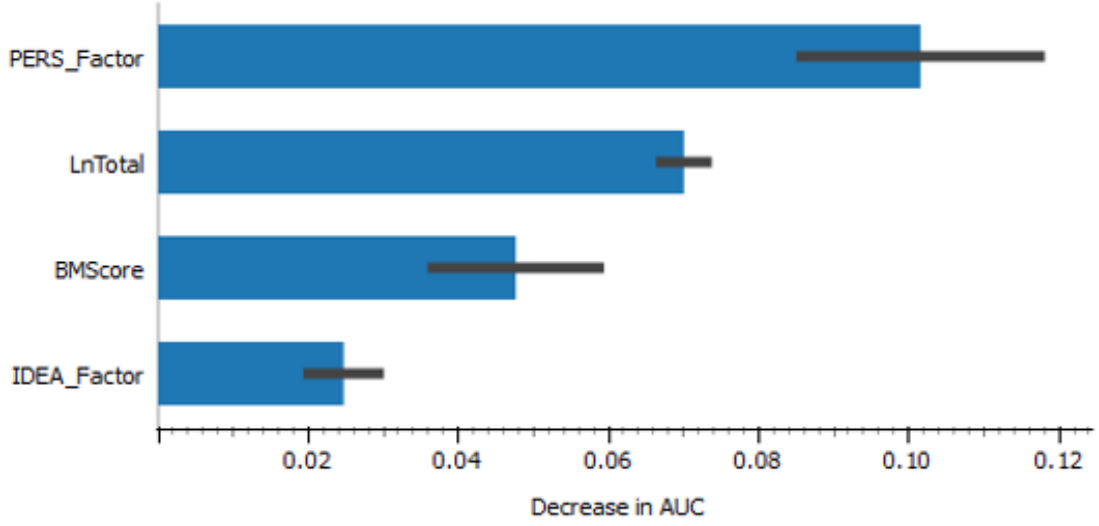
by the shortest bar. The removal of this feature causes the slightest decrease in AUC.

Furthermore, we analyze the influence of each feature with SHAP (Shapley Additive explanations), which is a method used to explain the classification model's output by measuring each feature's contribution to the model predictions. SHAP can provide global and local interpretations of how features influence model predictions. This method can also reveal interactions between features and is agnostic to model type so that it can be applied to various classification models. We conducted the SHAP analysis on the Gradient Boosting Model as it scored the highest performance in the classification task.

Based on Figure 2, the numbers show the specific contribution of each feature to the model predictions in a single sample. They describe how much each feature influences the prediction and in which direction (positive or negative) it influences. Meanwhile, the numbers outside the bar show the cumulative total contribution of each feature in the entire dataset, providing a comprehensive understanding of the model's behavior. The segments in the blue bar (positive association) are BMScore with a SHAP value of 0.69 and IDEA_Factor with a SHAP value 1.39. The SHAP values represent the magnitude of the influence of each feature on the model predictions. Meanwhile, the red bar (negative association), namely LnTotal, has a SHAP value of 0.37, and PERS_Factor shows a SHAP value of 0.29. The value for the IDEA_Factor on the outside is 2.08, which makes the most significant positive contribution to the prediction. BMScore makes a positive contribution with a value of 0.35, while LnTotal makes a negative contribution of 10.16, and the PERS_Factor feature provides a negative contribution of -0.28.

Based on the result presented, the findings indicate that the higher the BMScore, the higher the possibility of reports containing fraud. An increase in BMScore indicates an anomaly in the financial statements, which can strongly indicate that a company may engage in fraudulent practices (Beneish, 1999). On the other hand, the higher the total number of words in the annual report, the lower the possibility of fraud in the report. More extended annual reports reflect transparency and openness from company management when providing information to stakeholders. Research by Li (2008) shows that companies more open to disclosing information tend to have healthier financial practices and are less likely to engage in accounting

Figure 1. **Feature Importance of Gradient Boosting Model**



Source: Processed Data

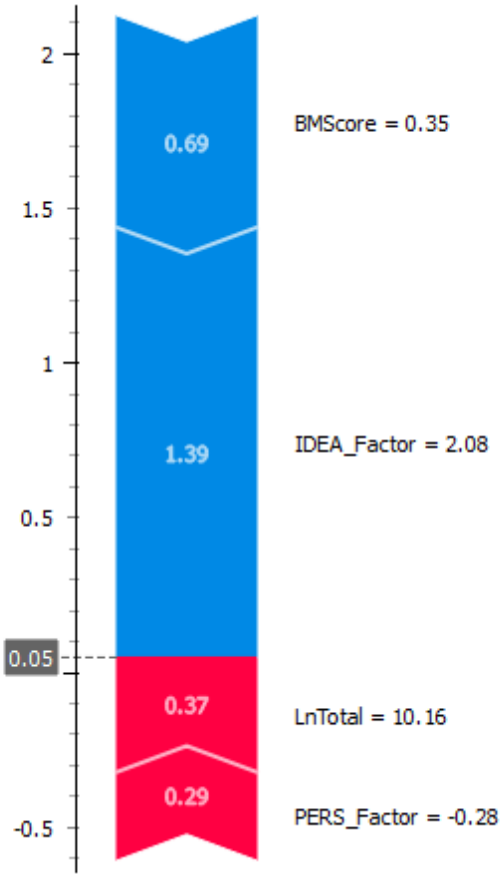
manipulation. Therefore, linguistic analysis that considers the length and depth of annual reports can be a valuable additional tool in detecting potential fraud, providing a more comprehensive understanding of a company's financial health.

Additional Analysis

Our primary analysis focuses on how financial features, as represented by Beneish M-Score, and linguistic features contribute to the classification performance in identifying fraudulent reports. However, to provide a more detailed understanding, we conduct an additional analysis. This analysis is designed to examine how each feature category contributes to predicting fraudulent reports, thereby enhancing our overall comprehension. The test results in Figure 3 show the importance of the features, which are measured based on the decrease in AUC of the Gradient Boosting model when the feature is removed. The bar in the figure shows the magnitude of each feature in the model, indicating the relative influence on the model prediction.

First in the Beneish ratio group (left) shows that DEPI, AQI, and TATA considerably influence the model, with DEPI making the most significant contribution to the decrease in AUC. DEPI, which reflects depreciation expense, is an essential indicator in detecting fraudulent reports because depreciation manipulation can significantly affect a company's reported profits. AQI, which measures asset quality, also showed a significant impact, indicating that companies with low-quality or questionable assets were more likely to engage in fraud. TATA, which reflects total accruals, is another essential feature, as high accruals can indicate accounting manipulation.

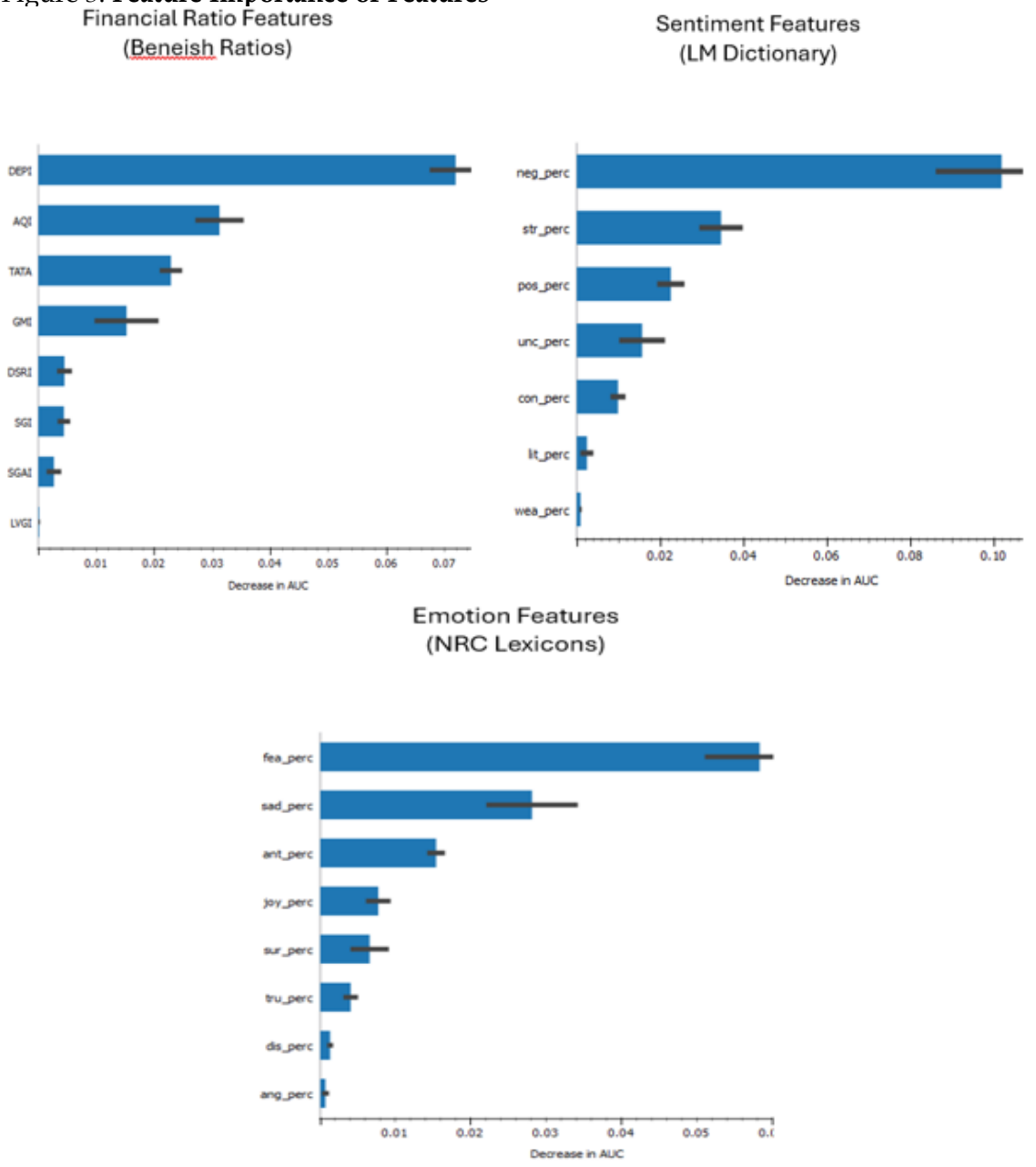
Figure 2. **SHAP Analysis of Gradient Boosting Model**



Source: Processed Data

Secondly, in the sentiment features based on the LM Dictionary (middle), the percentage of negative emotions (neg_perc) and strong modals (str_perc) show the most significant influence, followed by the percentage of positive emotions (pos_perc). These features indicate that the sentiments expressed in annual reports can provide important insights into a company's intentions. High negative emotions may reflect dissatisfaction or worry that is attempted to be hidden. In contrast, significant positive emotions and strong modals may indicate attempts to reassure or obscure reality. Lastly, features based on the NRC lexicon (right), shows that fear (fea_perc) and sadness (sad_perc) have the most significant impact on the model. These emotions were found to be most influential in the model's predictions of fraud, indicating that companies involved in fraud were more likely to convey fear and sadness in their reports. This could be due to the internal and

Figure 3. **Feature Importance of Features**



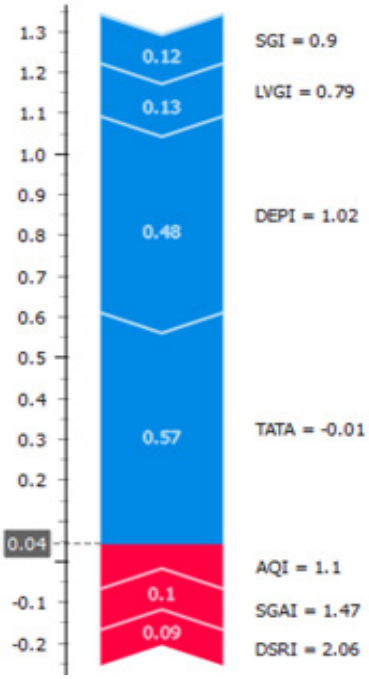
Source: Processed Data

external pressures they face or attempts to cover up the grim reality. Thus, this linguistic analysis of emotions provides a valuable tool in detecting deception, as certain emotional expressions can indicate attempts at manipulating or concealing accurate information.

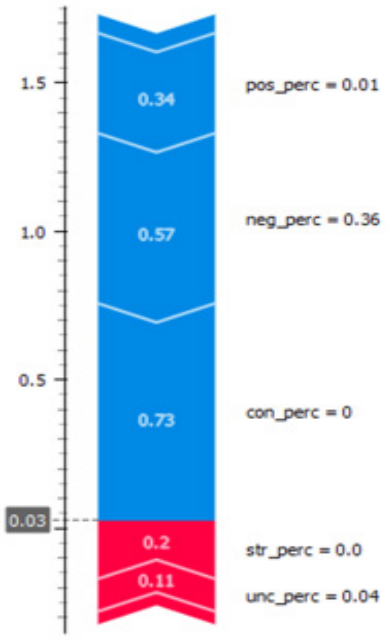
The SHAP test results in Figure 4 show the contribution of various features to model predictions in three feature groups. In the Beneish ratio features (left), DEPI, with a SHAP value of 0.48 and a cumulative contribution of 1.02, has the most significant positive impact on model

Figure 4. **SHAP Analysis of Features**

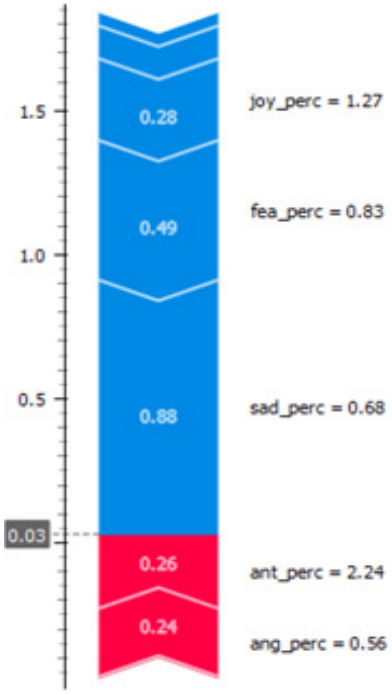
Financial Ratio Features (Beneish Ratios)



Sentiment Features (LM Dictionary)



Emotion Features (NRC Lexicons)



Source: Processed Data

predictions, followed by SGI (0.12) and LVGI (0.13). Meanwhile, DSRI provided the most significant negative contribution with a SHAP value of -0.04 and a cumulative value of -2.06. Fraudulent firms have higher DEPI, SGI, and LVGI ratios. A high DEPI suggests they depreciate assets less to inflate profits. A high SGI could indicate rapid, possibly manipulative, sales growth. A high LVGI implies they use more debt to mask financial weaknesses. Conversely, fraudulent firms have lower AQI, SGAI, and DSRI ratios. A low AQI reflects poorer asset quality, potentially from manipulative recording. A lower SGAI suggests reduced costs to boost perceived profits. A low DSRI indicates sales are recognized faster, or receivables are delayed, to enhance apparent cash flow.

In the sentiment LM Dictionary features (middle), neg_perc (Negative) with a SHAP value of 0.57 and a cumulative contribution of 0.36 and con_perc (Constraining) with a SHAP value of 0.73 had the most significant positive impact. Meanwhile, str_perc (Strong Modal) and unc_perc (Uncertainty) provide negative contributions with values of -0.2 and -0.11 respectively. Fraudulent reports often use more negative and constraining words but fewer strong words. This may indicate fraudsters' cunning nature, as they aim to obscure poor financial results by using negative language generally while avoiding strong words that could raise suspicion. Furthermore, the SHAP analysis of the NRC lexicons (right), sad_ perc (Sadness) with a SHAP value of 0.88 and a cumulative contribution of 0.68, and fea_perc (Fear) with a SHAP value of 0.49 has a significant impact on prediction. Furthermore, ant_perc (anticipation) provides the most significant negative contribution with a SHAP value of -0.26 and a cumulative -2.24. Overall, fraudulent reports often use sad and fearful words, creating a negative tone that influences perceptions. Sad words may indicate internal problems or declining performance, while fearful words reflect concern about the company's future due to operational issues or external threats. Meanwhile, anticipation-related words are less frequent, suggesting a lack of long-term planning or attempts to avoid unfulfilled promises. This linguistic analysis offers insight into the communication strategies of fraudulent companies.

This study shows that combining linguistic features and financial ratios greatly enhances fraud detection accuracy in financial statements. A multidimensional approach using both quantitative (Beneish ratio) and qualitative (language patterns) data is crucial. The Gradient Boosting model with combined features outperformed others, achieving over 91% accuracy, confirming that this integrative approach is essential for effective machine learning-based fraud detection. The findings are valuable for auditors, regulators, and investors. Unlike prior studies focusing separately on NLP or ML, this research demonstrates that linguistic features have stronger predictive power when paired with financial indicators. Identifying negative emotions, modality expression, and document length as key indicators is vital for fraud detection. This study also advances fraud detection in the Indonesian market by adapting linguistic resources. Comparing SHAP and feature importance enhances model interpretability and validity. The results validate the hybrid approach and expand fraud detection analysis, highlighting linguistic dimensions as critical indicators of suspicious behavior, informing adaptive detection systems in market supervision and governance.

5. CONCLUSION

This study evaluates classification ML models for detecting manipulation in corporate reports using combined features of financial ratios and linguistic features. This study found that the combination of features complement each other and can improve the overall performance of the classification models. Gradient Boosting models perform the best among the ML

models, achieving 91% correctness in CA, F1, Precision and Recall metrics. The result shows that adding linguistics features based on the SFL framework significantly improves the performance of the classification models. These findings underscore the value of text mining and ML in financial accounting, making the audience realize the importance of the study's conclusions.

The results also provide evidence of Beneish M-Score's ability to predict fraudulent reports as well as linguistic features and power in identifying false reports. The findings show that increased BMScore is associated with a higher likelihood of fraud in financial reports. Furthermore, firms involved in fraud tend to have higher DEPI, SGI, and LVGI ratios than non-fraud companies. Manipulated depreciation indicated by DEPI, low or questionable asset quality measured by AQI, and high accruals indicated by TATA are all critical indicators in detecting fraudulent financial statements. On the other hand, linguistic analysis conducted shows that a more extended number of words in annual reports tends to reflect the transparency of company management and tend to have healthier financial practices. In addition, sentiment and emotion words analysis shows that fraudulent companies tend to use negative and constraining language, and express emotions such as fear and sadness in their reports to hide the unfavorable performance. These communication patterns could indicate that companies are trying to cover up their dishonesty by creating a more vague and indirect narrative as well as avoid greater scrutiny from regulators or other external parties.

This study has limitations that could be addressed in future research. First, the analysis was carried out on annual reports in Indonesian, which required the translation process of the LM Dictionary and NRC lexicons. This translation process causes a decrease in the number of unique words identified, due to differences in structure, grammar, and time references in Indonesian compared to English. This can reduce the depth of analysis and obscure language context. In addition, the use of pronouns in Indonesian may have a different impact on the context of communication than in English. This research has added the word "society" to the category, but opportunities might not have been revealed. It is crucial that future research prioritizes the development of an Indonesian dictionary that is more relevant in the context of financial reporting. Secondly, this research only uses the Beneish M-Score as a financial ratio to detect fraud. Using other ratios, such as Dechow's F-Score or other fraud detection measures, might be beneficial in providing more robust results.

REFERENCES

Aghghaleh, S. F., Mohamed, Z. M., & Rahmat, M. M. (2016). Detecting Financial Statement Frauds in Malaysia: Comparing the Abilities of Beneish and Dechow Models. *Asian Journal of Accounting and Governance, 7,* 57–65. https://doi.org/10.17576/ajag-2016-07-05.

Aronoff, S. (1982). Classification Accuracy: A User Approach. *Photocrammetric Engineering and Remote Sensing, 48*(8), 1299–1307.

Ashtiani, M., & Raahemi, B. (2022). Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review. *IEEE Access, 10*(6), 72504–72525. https://doi.org/10.1109/ACCESS.2021.3096799.

Association of Certified Fraud Examiners (ACFE). (2024). *Occupational Fraud 2024: A Report to the Nations*. Association of Certified Fraud Examiners (ACFE).

Beneish, M. (1999). The Detection of Earnings Manipulation. *Financial Analysts Journal - FINANC ANAL J, 55*(5), 24–36. https://doi.org/10.2469/faj.v55.n5.2296.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324.

Chan, S., & Chong, M. (2016). Sentiment Analysis in Financial Texts. *Decision Support Systems, 94*(2017), 53-64. https://doi.org/10.1016/j.dss.2016.10.006.

Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining, 16*(4). 1-23. https://doi.org/10.1186/s13040-023-00322-4.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964.

Craja, P., Kim, A., & Lessmann, S. (2020). Deep learning for detecting financial statement fraud. *Decision Support Systems, 139*(1), 113421. https://doi.org/https://doi.org/10.1016/j.dss.2020.113421.

Dechow, P., Ge, W., Larson, C., & Sloan, R. (2010). Predicting Material Accounting Misstatements. *Contemporary Accounting Research, 28*(1), 17-82. https://doi.org/10.1111/j.1911-3846.2010.01041.x.

Dong, W., Liao, S., & Liang, L. (2016). Financial Statement Fraud Detection Using Text Mining: A Systemic Functional Linguistics Theory Perspective. *Proceeding of the 20th Pacific Asia Conference on Information Systems (PACIS 2016)*.

Faccia, A., McDonald, J., & George, B. (2024). NLP Sentiment Analysis and Accounting Transparency: A New Era of Financial Record Keeping. *Computers, 13*(5), 1-18. https://doi.org/10.3390/computers13010005.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences, 55*(1), 119–139. https://doi.org/https://doi.org/10.1006/jcss.1997.1504.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451.

Fuller, C., Biros, D., & Delen, D. (2011). An investigation of data and text mining methods for real world deception detection. *Expert Syst. Appl., 38*(7), 8392–8398. https://doi.org/10.1016/j.eswa.2011.01.032.

Gotelaere, S., & Paoli, L. (2022). Prevention and Control of Financial Fraud: a Scoping Review. *European Journal on Criminal Policy and Research, 31*(1), 1–21. https://doi.org/10.1007/s10610-022-09532-8.

Hajek, P., & Henriques, R. (2017). Mining Corporate Annual Reports For Intelligent Detection of Financial Statement Fraud – A comparative study of machine learning methods. *Knowledge-Based Systems, 128*(C), 139–152. https://doi.org/https://doi.org/10.1016/j.knosys.2017.05.001.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *An Introduction to Functional Grammar (Third Edition)*. Hodder Arnold.

Healy, P. M., & Wahlen, J. M. (1999). A Review of the Earnings Management Literature and Its Implications for Standard Setting. *Accounting Horizons, 13*(4), 365–383. https://doi.org/10.2308/acch.1999.13.4.365.

Hicks, S., Strumke, I., Thambawita, V., Hammou, M., Riegler, M., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports, 12*(1), 1-9. https://doi.org/10.1038/s41598-022-09954-8.

Hogan, C., Rezaee, Z., Riley, J., & Velury, U. (2008). Financial Statement Fraud: Insights from the Academic Literature. *Auditing, 27*(2), 231–252. https://doi.org/10.2308/aud.2008.27.2.231.

Jaballi, S., Zrigui, S., NICOLAS, H., & Zrigui, M. (2024). Analyzing Multilingual Conversations During COVID-19: An Imbalanced Class-Ensemble Learning Approach with Reweighted AdaBoost-SVM for Code-Switched Text Classification. https://doi.org/10.21203/rs.3.rs-3978507/v1

Kanapickiene, R., & Grundienė, Ž. (2015). The Model of Fraud Detection in Financial Statements by Means of Financial Ratios. *Procedia - Social and Behavioral Sciences, 213*(2015), 321–327. https://doi.org/10.1016/j.sbspro.2015.11.545.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539.

Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography, 17*(2), 145–151. https://doi.org/https://doi.org/10.1111/j.1466-8238.2007.00358.x.

Lokanan, M., & Sharma, S. (2024). The use of machine learning algorithms to predict financial statement fraud. *The British Accounting Review, 56*(6), 101441. https://doi.org/https://doi.org/10.1016/j.bar.2024.101441.

Loughran, T., & Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance, 66*(1), 35–65. https://doi.org/https://doi.org/10.1111/j.1540-6261.2010.01625.x.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing A Word–Emotion Association Lexicon. *Computational Intelligence, 29*(3), 436–465. https://doi.org/https://doi.org/10.1111/j.1467-8640.2012.00460.x.

Perols, J. (2010). Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. *Auditing A Journal of Practice & Theory, 30*(2), 19-50. https://doi.org/10.2308/ajpt-50009.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust You?" Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016, 1135–1144. https://doi.org/10.1145/2939672.2939778

Saquete, E., Tomás, D., Moreda, P., Martínez-Barco, P., & Palomar, M. (2020). Fighting post-truth using natural language processing: A review and open challenges. *Expert Systems with Applications, 141*(7), 112943. https://doi.org/https://doi.org/10.1016/j.eswa.2019.112943.

Soltani, M., Kythreotis, A., & Roshanpoor, A. (2023). Two decades of financial statement fraud detection literature review; combination of bibliometric analysis and topic modeling approach. *Journal of Financial Crime, 30*(5), 1367-1388. https://doi.org/10.1108/JFC-09-2022-0227.

Throckmorton, C., Mayew, W., Venkatachalam, M., & Collins, L. (2015). Financial Fraud Detection Using Vocal, Linguistic and Financial Cues. *Decision Support Systems, 74*(2015), 78-87. https://doi.org/10.1016/j.dss.2015.04.006.

Zhou, W., & Kapoor, G. (2011). Detecting Evolutionary Financial Statement Fraud. *Decision Support Systems, 50*(3), 570–575. https://doi.org/10.1016/j.dss.2010.08.007.