

IMPROVING ACCURACY IN FRAUD DETECTION IN PUBLIC COMPANIES FINANCIAL REPORTS USING NATURAL LANGUAGE PROCESSING

ANGELA KURNIAWAN¹, LAUDYA NATALIE YUDHA² ARMANTO WITJAKSONO³

¹Accounting Department, School of Accounting, Bina Nusantara University, Indonesia, 11480

²Accounting Department, School of Accounting, Bina Nusantara University, Indonesia, 11480

³Accounting Department, School of Accounting, Bina Nusantara University, Indonesia, 11480

E-mail: ¹angela.kurniawan@binus.ac.id, ²laudya.yudha@binus.ac.id, ³armanto@binus.ac.id

ABSTRACT

This study develops and evaluates a Natural Language Processing (NLP) model for detecting financial statement fraud in Indonesian public companies through linguistic analysis of annual reports. We analyzed 72 financial reports (20 fraudulent, 52 non-fraudulent) from companies listed on the Indonesia Stock Exchange using TF-IDF-based text representation combined with linguistic complexity features. Machine learning classifiers (Logistic Regression, Random Forest) were evaluated using 5-fold cross-validation. The NLP model achieved 77.3% accuracy, significantly outperforming traditional rule-based audit method (40.9%) by 88.9%. Contrary to theoretical expectations, fraudulent reports used simpler language (ambiguity score: 0.69 vs. 0.98) but exhibited significantly higher hedging language usage ($p=0.033$). TF-IDF features dominated model performance (99.9% contribution). This study is the first to systematically apply NLP-based fraud detection to Indonesian financial statements in a bilingual reporting context, revealing context-specific linguistic patterns that differ from Western fraud literature.

Keywords: *Natural Language Processing, Fraud Detection, Public Companies Financial Reports, Artificial Intelligence, Manipulation*006E

1. INTRODUCTION

Financial statement fraud represents a persistent threat to capital market integrity, stakeholder trust, and corporate credibility. According to the PwC Global Economic Crime Survey 2024, 41% of global companies reported experiencing fraud or economic crime within the preceding 24 months [20]. In Indonesia, high-profile cases such as the Garuda Indonesia scandal underscore the vulnerability of traditional audit methodologies to sophisticated financial statement manipulation [26]. Manual audits, constrained by cognitive limitations, time pressures, and reliance on numerical indicators, often struggle to process the growing volume and complexity of narrative financial disclosures, increasing the risk of undetected fraudulent reporting.

Recent advances in Artificial Intelligence, particularly Natural Language Processing (NLP), have enabled the systematic analysis of unstructured textual data to uncover hidden linguistic patterns. While prior fraud detection studies have predominantly applied machine learning techniques to numerical financial indicators, relatively limited attention has been

given to linguistic manipulation within financial narratives. Existing NLP-based fraud research is largely concentrated on English-language disclosures from developed markets, leaving a notable research gap in emerging markets with bilingual reporting practices and distinct regulatory environments.

This study addresses this gap by investigating the application of NLP techniques to detect potential financial statement fraud through linguistic analysis of annual reports of Indonesian public companies. The scope of this research is limited to publicly listed companies in Indonesia and focuses on narrative sections of financial reports. Specifically, this study seeks to answer the following research questions: (1) how NLP can be utilized to identify linguistic patterns reflecting rationalization, such as ambiguous phrasing and hedging language; (2) how NLP-based fraud detection compares with traditional audit methods; and (3) how inconsistencies in management narratives may indicate information asymmetry associated with fraudulent reporting.

The primary contribution of this research lies in the development and evaluation of an NLP-based fraud detection model tailored to the Indonesian financial reporting context. This study contributes to the literature by extending fraud detection research to a bilingual, emerging market setting and by empirically demonstrating the added value of textual analysis beyond traditional audit approaches. From a practical perspective, the findings provide auditors and regulators with a scalable, evidence-based tool to enhance fraud risk assessment and support more proactive auditing practices.

2. LITERATURE REVIEW AND HYPOTHESIS DEVELOPMENT

There are various key theories that can form the basis of research into the detection of fraud using Natural Language Processing. Three relevant theories to complement the analysis of this research are those of the Fraud Triangle, Theory of internal processing, and Agency Theory. These theories provide a conceptual foundation for understanding financial fraud and how NLP technology can be used to accurately detect accounting anomalies.

2.1 Fraud Triangle Theory

The fraud triangle was first introduced and developed by Donald Cressey in 1953, explaining that there are major factors that cause someone to commit fraud in the workplace. This triangle consists of three components that together lead to fraudulent behavior, namely Pressure, Opportunity, and Rationalization [23].

Pressure as described by Donald Cressey is one of the incentives that can motivate a person to commit a criminal act. When a person has pressure stemming from personal problems, such as financial pressure, or addiction, as well as from the work environment, management or other employees may receive incentives or be under pressure that encourages or promotion is heavily influenced by the performance of an individual, division, or company, then the individual may have an incentive to manipulate results or put pressure on others to do the same. This pressure will create a motive for the crime to be committed, but the employee must also feel that he has the opportunity to commit the crime without being caught. The perception of the opportunity is what forms the second element of the triangle. There are two components in the perception of opportunity to commit a breach of trust, general information and technical skills. General information refers to the knowledge that the position of trust held by the employee can be abused, while technical skills refer

to the ability needed to commit the breach. The last factor is rationalization, not just an attempt to justify after the act of theft occurs, but on the contrary that rationalization is an essential component before committing fraud, this is part of the perpetrator's motivation to commit the act because the embezzler does not see themselves as a criminal, so the perpetrator must first justify the actions that will be made. Rationalization is necessary for the perpetrator because it is to understand their actions that violate the law and still maintain their self-image as someone who can be trusted. Research says that there is no one dominant motive, but rather a variety of factors such as business and investment failures, or revenge. Weak internal control will create opportunities for fraud, the assumption that no one is harmed, and the belief that stolen funds will one day be returned.

H1: Linguistic patterns that reflect rationalization, such as high language complexity or ambiguous phrases in financial statements, are positively associated with the likelihood of financial fraud.

2.1.2 Information Processing Theory

Information processing developed by Allen Newell and Simon (1972), explains how humans and computer systems can process information to solve various problems and make decisions. This processing is divided into three main stages, information reception, information processing, and decision making. This theory is based on the analogy that the human mind functions like a computer, with stages in the systematic reception, processing, and storage of information. This theory is used in this study because it is relevant to support the use of Natural Language Processing technology to analyze financial statement text efficiently, identifying anomalies or patterns that indicate fraud. The information reception stage is the stage of collecting data from the environment, such as financial statement text, numerical data, or management records.

When audits are conducted manually, auditors collect this information in a selective manner, which is often limited in terms of time and human cognitive capacity. NLP, as a computer-based tool, allows receiving large amounts of data of text at the same time without losing important details. In the processing phase, the information that has been obtained will be analyzed, classified or interpreted to find patterns. In this research, NLP uses algorithms such as sentiment analysis, entity extraction or deep learning models to identify

linguistic patterns, such as narrative inconsistencies and ambiguous language usage. Information that is successfully stored and analyzed will be reused during the decision making or output stage. In terms of auditing, decisions are important and can be in the form of marking financial statements as high-risk or low risk in terms of fraud.

H2: Deep Learning-based NLP models have higher accuracy in detecting financial fraud than traditional manual-based audit methods.

2.1.3 Agency Theory

The relationship in agency theory [2] arises when a principal appoints an agent (manager) to act on his behalf and delegates decision-making authority to the agent. It is a matter of trust on the part of the principal in the ability and intention of the agent to act in the best interest of the principal. Corruption arises because of potential breaches of trust in fiduciary relationships. Separation of ownership and control requires good governance and involves various mechanisms within institutions and in the market to ensure good governance and mitigate agency problems.

H3: The existence of inconsistencies between management narratives in financial statements, as an indicator of information asymmetry, has a positive relationship with the likelihood of financial fraud.

2.2 Concept

Financial fraud is defined as the intentional act of falsifying, manipulating, or concealing information in a financial statement in order to mislead stakeholders, such as investors, creditors, regulators (Association of Certified Fraud Examiners) and the public. Financial fraud can occur in two main forms, namely, numerical data manipulation with revenue recognition, fictitious sales or asset inflation, and textual manipulation, namely misleading management statements to hide bad loans so that reports look good in the eyes of investors and the public. In the context of this research, fraud is focused on the manipulation of text narratives in financial reports such as annual reports or management notes. This research aims to reveal fraud patterns using NLP, if narratives that deviate from the original often reflect fraudulent motives as described in the Fraud Triangle theory.

Natural Language Processing is a branch of Artificial Intelligence that helps computers to understand, interpret and manipulate human language [5]. Computer science is closely related to each other and is an important part of NLP processing. Computational linguistics is one example of the application of NLP because of its

efforts to bridge the gap between computer understanding and human communication. Natural Language Processing covers a wide range of linguistic techniques for interpreting human language, ranging from statistical methods and machine learning to rule-based and algorithmic approaches. The basic tasks of NLP itself include Tokenization and Parsing, Lemmatization/streaming, Part-of-speech, Language detection and Identification of semantic relationships. In general, the task of NLP is to break down language into shorter pieces of elements that can work together to form a meaning. Because of these advantages, NLP has become a major solution in semantic processing and correction of various errors that may often occur. The fields related to NLP are as follows, phonetics and phonology which are related to sound processing to produce words that can be recognized and used in applications based on sound systems. Morphology studies words and forms and is used to distinguish one word from another. Syntax which includes an understanding of word order and the formation of words into systematic sentences. Semantics which is related to the level of knowledge based on different contexts and tailored to the situation and purpose.

Financial statement analysis is a process that evaluates historical and current financial data in order to assess the performance of a company and estimates future risks and potential. According to research from Auwalu and Ibrahim [18], financial statements present information regarding the financial position, performance, and changes in the financial condition of an entity in a standardized format from authorities to assist investors, regulators, financial observers, and other stakeholders during economic decision-making. The analysis of financial reports primarily focuses on the data within the company's annual reports, which include financial statements and supplementary information. During the analysis process, patterns, changes, SWOT indicators, and relationships among the components of the reports are identified. Accountants, auditors, financial analysts, and other financial advisors have access to crucial information from financial statements (FSA) with the aim of evaluating performance as well as estimating risks and the potential of the company in the future. Financial analysts often make adjustments to the financial statement prepared by the company's accountants, either by disregarding items deemed insignificant or by adding aspects considered crucial but not directly stated in the formal financial statements.

2.3 Previous Studies

There are previous studies that are relevant to the current research related to fraud detection using Natural Language Processing.

2.3.1 Studies about Fraud Detection with Natural Language Processing

Similar research has also been conducted previously, notably the study by Muktha & Manish (2025) [7] titled “*Fraud Detection with Natural Language Processing*”. This research was undertaken with the aim of enhancing automated fraud detection in mobile banking. The study was prompted by the urgency related to the increasing volume of transactions, highlighting the need for real-time protection for users. Additionally, many similar studies have focused solely on credit card fraud and have not given much attention to mobile banking due to the lack of available data. The results of this research indicate that online users exhibit patterns consistent with language in Natural Language Processing, making Natural Language Processing an effective tool for detecting fraud. The research also presents possibilities for the development of a more advanced fraud detection system, particularly for mobile banking. This research supports the use of Natural Language Processing as an assistance tool for auditors and can serve as a basis for Indonesian Public Companies.

2.3.2 Studies about Financial Fraud Detection with Machine Learning Algorithms

Research related to financial fraud detection using machine learning has been conducted previously. The study is titled “*Evaluating Machine Learning Algorithms for Financial Fraud Detection: Insights from Indonesia*.” This research aims to evaluate the effectiveness of machine learning algorithms in detecting financial statement fraud [16], as well as to identify key financial indicators associated with fraudulent activities. The algorithms used in this study include multiple linear regression, logistic regression, K-Nearest Neighbors, Support Vector Machine, Decision Tree, and Random Forest. The metrics employed in this research measure the levels of precision, recall, accuracy, and F1-Score. The study indicates that the random forest algorithm is the most effective, particularly in relation to training and testing datasets. In addition, there are also logistic regression and Support Vector Machines that exhibit strong reliability. Meanwhile, K-Nearest Neighbors and Decision Tree are considered to be overfitting, which limits

their practical use. From the research, there are indicators of fraud that serve as significant predictors of financial statement fraud, among which are accounts receivable turnover, days sales outstanding in accounts receivable, days payables outstanding, logarithm of gross profit, gross profit margin, inventory to sales ratio, and total asset turnover.

2.3.3 Studies about automated financial reporting with Natural Language Processing

There have also been prior studies related to financial reporting that have been integrated using natural language processing. These studies examined how there is a transformation in the field of financial reporting by enabling automation, enhancing accuracy, transparency, and also compliance with regulations. The research employed qualitative methods. From this study, it was found that Natural Language Processing is used for sentiment analysis by detecting patterns or anomalies in the disclosures of financial reports or in financial narratives, as well as for deep learning in financial forecasting. From the study, there are difficulties in applying Natural Language Processing, including challenges related to the diverse finances surrounding the use of AI and NLP [1], and the challenge of ensuring that the reports generated by NLP are easily accessible and understood by stakeholders. The study indicates that further research is still needed, particularly in terms of more inclusive and nuanced training data.

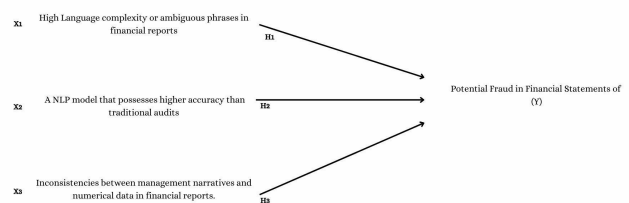


Figure 1. Conceptual Framework

This research framework describes research thinking focused on analyzing the potential for fraud in financial reports by utilizing NLP. This study is grounded in three independent variables, namely High Language complexity or ambiguous phrases in financial reports, a NLP model that possesses higher accuracy than traditional audits, and inconsistencies between management narratives and numerical data in financial reports. These three variables are assumed

to have an influence on the dependent variable, which is the potential for fraud in financial reports (Y).

This relationship is explained through the main hypothesis encompassing all three independent variables. The framework emphasizes the importance of utilizing modern language analysis technologies such as NLP in detecting manipulation practices within financial reports.

3. METHOD, DATA, AND ANALYSIS

3.1 Research Method

This study employs a supervised machine learning approach with quantitative evaluation metrics. The research design follows a binary classification framework, where financial statements are categorized as fraudulent or non-fraudulent based on linguistic features extracted through NLP techniques. This approach is justified by the availability of labeled data from regulatory sanctions, the measurable nature of linguistic features, and the need for replicable, objective fraud detection criteria.

3.2 Population and Sample

The initial dataset consisted of 17 annual financial reports from Indonesian public listed companies on the Indonesia Stock Exchange (IDX), comprising 4 fraudulent and 13 non-fraudulent reports, covering financial reporting violations between 2017 and 2024. Fraudulent cases were identified based on official regulatory sanctions and publicly disclosed enforcement actions related to financial reporting misconduct.

Given the limited availability of confirmed financial statement fraud cases in the Indonesian capital market and the computational constraints associated with NLP-based text analysis, the initial sample size is relatively small. However, this scale is comparable to prior fraud detection studies conducted in emerging market contexts. To mitigate sample size limitations while preserving linguistic characteristics, NLP-based data augmentation techniques were applied, resulting in an expanded dataset of 72 textual documents.

The inclusion criteria for fraudulent samples required companies to (1) be publicly listed on the IDX, (2) have received official sanctions related to financial reporting violations, and (3) have financial reports published between 2017 and 2024. Non-fraudulent samples consisted of financial reports from publicly listed IDX companies that had not received any regulatory sanctions related to

financial reporting violations during the same period.

3.3 Data Collection Techniques

Annual reports were downloaded in PDF format from www.idx.co.id and converted to machine-readable text using PyMuPDF. For companies publishing bilingual reports, English versions were prioritized for analysis consistency, given that NLP libraries (NLTK) are optimized for English.

In this research, it is important to note a disclaimer regarding the use of language in financial reports. Financial statements in Indonesia are commonly presented in two languages, namely Bahasa Indonesia and English. For the purposes of NLP-Based analysis, this research relies on the English version to ensure compatibility with widely available linguistic tools. Nevertheless, in cases where differences in meaning or interpretation arise, the primary references remain the Bahasa Indonesia version as well as the accounting and auditing regulations applicable in Indonesia.

This is the list of companies that are used as a sample:

Fraud companies as of September 2025:

1. PT Indofarma Tbk – INAF
2. PT Hanson International Tbk – MYRX
3. PT Kimia Farma Tbk – KAEF
4. PT Timah Tbk - TINS

Non-Fraud Companies as of September 2025:

1. PT Astra Indonesia – ASII
2. PT Indofood Sukses Makmur - INDF
3. PT Gudang Garam Tbk – GGRM
4. PT Mayora Indah Tbk – MYOR
5. PT Unilever Indonesia Tbk - UNVR

Data Labeling Process:

Each company that is taken as a sample in this study, will be given a label, label = 1 is for companies that have been proven to have committed fraud and label = 0 is for companies that have not been proven or not proven to commit fraud. The data labeling category for companies proven to have committed fraud will be created according to the criteria that the company has received sanctions from the financial services authority (*Otoritas Jasa Keuangan*) or *Bursa Efek Indonesia* due to fraud committed. If the company meets the criteria mentioned, it will be labeled as 1, and those that do not meet these criteria will be labeled as 0 [9].

3.4 Data Analysis Techniques

The annual financial report data obtained will be analyzed by running a Natural Language Processing system using the Python application. The financial report data used, before and after the detection of fraud, will be entered into the Python system. This is done so that the system can perform sentiment analysis, where the system will detect patterns or anomalies between the two financial reports, which allows the fraud to be detected. When a pattern has been found by the system, this pattern will be tested again to ensure its accuracy in detecting fraud in other financial reports. This test was conducted by re-entering the same financial report (before the fraud was detected) and trying to run the system again.

The results obtained will be compared to see if the original financial statements (after fraud detection) are the same, similar, or very different from the results produced by the system. The evaluation parameters will also be obtained from those results, where the parameter used here is the accuracy of the system's success in detecting fraud in public company financial reports. The results will show the outcomes of the system and the results of traditional audits, to compare the similar percentage of the fraud detection results, which will then determine the system's accuracy in helping to detect fraud in the financial statements of public companies, thereby will prove the hypothesis and research framework of this research.

Data analysis will be carried out through several stages, starting from text preprocessing, feature extraction using TF-IDF, to classification with Logistic Regression, Random Forest, and ensemble methods. To evaluate the effectiveness of the fraud detection model, several classification metrics are employed, namely Accuracy, Precision, Recall, F1-Score, and Cross-Validation [13].

- Accuracy measures the proportion of documents correctly classified as either fraudulent or non-fraudulent.
- Precision is a measure that indicates the proportion of positive predictions that are truly accurate, calculated by dividing the number of True Positives by the total of True Positive and False Positives.
- Recall (Sensitivity) represents the proportion of actual positive cases that are correctly identified by the NLP model.
- F1-Score is the harmonic means of precision and recall, providing a single metric that balances both aspects

- Cross-Validation involves partitioning the dataset so that each fold alternately serves as both testing and training data.

There are challenges that can occur in data analysis in this study, where public-listed companies that have been proven to commit fraud are far fewer than companies that have not or have not been proven to commit fraud, so there can be imbalance data between companies that have been proven and those that have not been proven to commit fraud. With imbalanced data, there will be a tendency for the data to be biased towards the majority data, namely companies that have not been proven to commit fraud [3]. This is certainly undesirable because there will be the potential for the desired pattern or anomaly to be detected, so it is not detected. Then, the data balancing method will be carried out later when the data obtained turns out to be unbalanced.

3.5 Natural Language Processing Model Used

This study employs several NLP models designed within integrated architecture comprising four main stages. The first stage is Text Preprocessing, which ensures that financial report texts are in a consistent format and ready for processing, involving Tokenization, stop word Removal, and Lemmatization using NLTK [11]. The second stage is Feature Engineering, which combines text representation based on TF-IDF (as a baseline) with additional linguistic complexity features such as Flesch Reading Ease, Ambiguity Score, and Jargon Density. The third stage involves the application of an ensemble method, integrating Logistic Regression and Random Forest through an averaging mechanism to enhance classification performance. The final stage is the implementation of cross-validation [21] to measure reliability, ensuring that the evaluation results are more consistent.

3.6 Natural Language Processing Algorithm Used

This study implements several NLP algorithms that support each stage of the predetermined model. The initial stage is preprocessing, which applies tokenization to split text into individual words [10], stop word removal to eliminate common words without significant meaning, and lemmatization to convert words into their base form [19]. These steps produce more representative textual data. A numerical representation of the text is then obtained using the Term Frequency–Inverse Document Frequency (TF-IDF) algorithm [19].

For the classification process, Logistic Regression is employed as a simple and interpretable baseline model, while Random Forest is applied as a robust decision tree-based model. To validate the hypothesis, statistical tests such as the independent t-test and Mann-Whitney U Test are performed, and the effect size is measured using Cohen's d. The evaluation further employs the cross-validation algorithm to ensure the reliability of results. The combination of these algorithms enables the integration of linguistic analysis, machine learning, and statistical methods into a text-based fraud detection framework for financial reports.

3.7 Ethical Risk of Using Natural Language Processing in Forensic Auditing

The use of Natural Language Processing (NLP) in forensic auditing, particularly in financial statement detection, raises several ethical risks that must be carefully considered. One major concern is algorithmic or data bias, which may occur when training data is unbalanced or reflects certain stereotypes. For example, companies from specific sectors might be perceived as more prone to financial statement manipulation without sufficient evidence, leading to prejudice against certain firms.

Another concern relates to the confidentiality of financial statements, as they contain sensitive information. The application of NLP must therefore comply with privacy regulations, such as Law (Undang-undang Republik Indonesia No. 27 Tahun 2022) on Personal Data Protection in Indonesia [24]. Failure to adhere to these regulations could jeopardize the confidentiality of corporate data. In addition, the use of NLP may intersect with legal processes, particularly when the system produces false positives that could harm a company's reputation.

In this research, it is important to note a disclaimer regarding the use of language in financial reports. Financial statements in Indonesia are commonly presented in two languages, namely Bahasa Indonesia and English. For the purposes of NLP-Based analysis, this research relies on the English version to ensure compatibility with widely available linguistic tools. Nevertheless, in cases where differences in meaning or interpretation arise, the primary references remain the Bahasa Indonesia version as well as the accounting and auditing regulations applicable in Indonesia.

4. RESULT AND DISCUSSION

4.1 Descriptive Statistic

The dataset used in this study consists of 17 financial reports from public companies in Indonesia, consisting of 13 non-fraud financial reports and 4 fraud financial reports. To address the limited data size and class imbalance in the dataset, this study implemented a data augmentation strategy. Through NLP-based data augmentation, the dataset was expanded to 72 documents without changing the original meaning of the report content. After augmentation, the dataset consists of 20 fraud reports (27.8%) and 52 non-fraud reports (72.2%), resulting in a more balanced representation across classes. This enhancement not only increases the dataset size but also reduces the risk of overfitting and improves the generalization of the proposed model. This method is used to increase sample diversity, ensuring that the fraud detection model can learn more representative patterns from financial reports.

In the data augmentation process, several augmentation techniques are used, such as Sentence Shuffling and Segment-Based Augmentation, as well as data cleaning and deduplication processes [14].

A. Sentence Shuffling

The first method involves randomizing the sequence of sentences in a document, focusing on fraudulent documents with more than 10 sentences to ensure sufficient sentence structure variation. This process involves taking all the sentences in a document, randomly shuffling them, and recombining them into a new text, helping the NLP model learn more diverse language patterns.

B. Segment-Based Augmentation

The second method involves breaking down long documents into sub-documents. This process is applied to financial reports that are longer than 5.000 characters. This involves breaking the text into chunks of approximately 1.000 words. This involves creating at least three sub-documents from a single long document. Each chunk must be at least 500 words long to maintain a clear context.

C. Cleaning and Deduplication

In this process, data validation is carried out by removing duplicate text to avoid redundancy, removing documents that have too short text length (less than 50 characters) to remain relevant, and

normalizing text with text preprocessing such as tokenization, stop word removal and stemming.

Table 1: Dataset Augmentation

Dataset Version	Total Docs	Fraud Docs	Non-Fraud Docs	Fraud Rate
Original Dataset	17	4	13	23.5%
Augmented Dataset	72	20	52	27.8%

4.1.1 NLP Variables and Features

This research utilizes three main hypotheses or variables analyzed through NLP features and linguistic statistics :

- **X1 - Language Complexity & Ambiguity**
 - Flesch Reading Ease: measures the readability of a text;
 - Flesch -Kincaid Grade: the difficulty level of the text;
 - Ambiguity Score: how vague the meaning of the text is;
 - Jargon Density: the proportion of technical terms;
 - T-test: comparing the readability of fraudulent and non-fraudulent texts.
- **X2 - NLP Model and Traditional Methods**
 - Text representation using TF-IDF (Term Frequency-Inverse Document Frequency);
 - The main models used Logistic Regression and Random Forest because they can handle small to medium-sized data;
 - Performance evaluation using Accuracy, Precision, Recall, and F1-Score;
- **X3 - Narrative - Numerical Inconsistency Analysis**
 - Inconsistency Score: measures the discrepancy between narrative and figures;
 - Hedging Score: measures a company's tendency to use ambiguous language to reduce information certainty;

4.1.2 Text Preprocessing

The preprocessing stage described in Section 3.6 was implemented on the dataset used in this study. It was first applied to the initial dataset, which consisted of 17 financial statement documents, all of which were successfully processed without any data loss, resulting in an

average length of 344,345 characters per document after case folding, stop word removal, lemmatization. Following the augmentation process, which expanded the dataset to 72 documents, the same preprocessing pipeline was reapplied to ensure that all data remained clean and consistent. Both the original and augmented datasets were thus aligned in text format and ready for feature construction using TF-IDF.

4.2 Model Performance Evaluation

To ensure the robustness and generalizability of the proposed NLP-based fraud detection model, several validation and reliability checks were conducted. The procedures included handling data imbalance, applying cross-validation, and evaluating the stability of the model.

Table 2: Fraud Dataset

Dataset	Fraud	Non-Fraud	Total	Fraud Percentage
Total Samples	20	52	72	27.8%
Training	14	36	50	28.0%
Testing	6	16	22	27.3%

4.2.1 Handling Imbalanced Dataset

After the dataset was expanded to 72 documents through augmentation, the distribution between fraud and non-fraud reports remained unbalanced. To address potential model bias, stratified sampling was used in the train-test split data, maintaining class proportions.

The table above shows that the distribution of fraud in the training (28.0%) and testing (27.3%) data is nearly identical to the distribution of the entire dataset (27.8%). This consistency demonstrates that the data distribution used for model evaluation is not biased toward any one class. Furthermore, this study used the F1-Score as the primary evaluation metric because it is more representative than accuracy in imbalance data by balancing precision and recall.

4.2.2 Baseline vs NLP Models

The performance evaluation of the model was carried out by comparing four main approaches, namely Traditional Logistic Regression (LR), The Enhanced NLP model, Traditional Audit, and the Combined Ensemble.

Table 3: Model Comparison

Model	Accuracy	Precision	Recall	F1-Score	CV Score
Traditional LR	86.4%	87.0%	87.0%	87.0%	80.8%± 11.6%
Enhanced NLP	77.3%	76.6%	76.6%	76.6%	66.4%± 15.8%
Traditional Audit	40.9%	43.8%	43.8%	43.8%	N/A
Combined Ensemble	72.7%	72.7%	72.7%	72.7%	68.5%± 19.1%

1. Traditional LR: achieved an accuracy of 86.4% and an F1-Score of 87%. The cross-validation result of 80.8% ± 11.6% indicates that the model is relatively stable, despite variations across validation folds.

2. Enhanced NLP: obtained an accuracy of 77.3% with an F1-Score of 76.6%. The NLP model demonstrated balanced detection capabilities between fraudulent and non-fraudulent classes, although the results were lower compared to Traditional LR. However, the cross-validation score of 66.4% ± 15.8% reflects higher fold variation, suggesting that this model is relatively less stable.

3. Traditional Audit: recorded the lowest performance, with an accuracy of 40.9% and an F1-Score only 43.8%. These results indicate that traditional manual audit methods remain less effective in detecting fraud indication based on textual information in financial reports.

4. Combined Ensemble: produced accuracy and F1-Score values of 72.7%, which are lower than those achieved by either LR or NLP individually. Although ensemble methods are generally expected to enhance model stability, in this study the results were more limited, with a cross-validation score of 68.5% ± 19.1%, indicating substantial variation across validation folds.

4.2.3 Cross-Validation Results

To assess the consistency and generalizability of the model, a 5-fold cross-validation was employed. The evaluation results show that Traditional Logistic Regression (LR) achieved an average cross-validation score of 80.8% ± 11.6%, indicating relatively consistent performance despite variations across folds. The standard deviation, which remains within a moderate range, suggests that the model can be considered fairly stable. In contrast, the Enhanced NLP Model recorded an average cross-validation score of 66.4% ± 15.8%, with a higher standard deviation. This indicates that the model’s performance tends to fluctuate across folds and demonstrates lower stability. Meanwhile, the Combined Ensemble Model obtained a cross-

validation score of 68.5% ± 19.1%, reflecting considerable performance variation, which suggests that the combined approach did not consistently improve stability.

4.2.4 Feature Contribution Analysis

Feature contribution analysis was conducted on the combined model, which integrates text representation based on TF-IDF with linguistic features (X1: Language Complexity) and narrative–numerical inconsistency indicators (X3). The results show that:

- **X2 (TF-IDF features)** contributed 99.9% to the predictions,
- **X1 (Language Complexity)** contributed only 0.1%, and
- **X3 (Narrative–Numerical Inconsistencies)** provided no contribution.

The number of features generated by TF-IDF (±2,500 unigrams and bigrams) is significantly larger compared to X1, which consists of only eight linguistic metrics, and X3, which comprises three indicators. In comparison, this makes the TF-IDF weights dominate the feature space. Word and phrase patterns have been shown to provide more consistent signals in distinguishing fraudulent from non-fraudulent reports than measures of language complexity or narrative inconsistencies. These results confirm that although the model incorporates TF-IDF with additional linguistic features, the primary contribution still comes from word-based text representation.

4.3 Hypothesis Testing Result

4.3.1.Hypothesis X1

Hypothesis X1 relates to high language complexity or ambiguous phrasing in financial reports. This is because the financial reports of companies found to have committed fraud use more ambiguous or complex language. The results of the hypothesis test indicate that the financial reports of companies found to have committed fraud actually use simpler language. This hypothesis test result is significant with a P Value of 0.005.

Table 4: Statistical Testing Results

X1 Statistical Testing Result	Fraud	Non-Fraud
Ambiguity Score	-2.924	0.005
Fraud Language Score	-2.914	0.005
Passive Voice Ratio	0.329	0.743
Jargon Density	-2.955	0.004

Languages that are categorized as ambiguous or complex are divided into 3 categories. First, the uncertain words, such as “may”, “might”, “could”, “possibly”, “potentially”, “approximately”, “roughly”, “about”, “around”, “uncertain”, “unclear”, “ambiguous”, “complex”, “complicated”, “difficult”, “challenging”, “subjective”, “estimate”, “assumption”, “believe”, “expect”, and “anticipate”. These words are based on some studies that state some authors that are making claims and not entirely certain, often use verbs such as may, might, could, and etc [12].

Second, fraud-specific evasive language, such as “extraordinary”, “exceptional”, “unusual”, “unprecedented”, “unique”, “special”, “remarkable”, “significant”, “material”, “substantial”, “various”, “certain”, “some”, “several”, “multiple”, “numerous”, “appropriate”, “reasonable”, “adequate”, “sufficient”, and “necessary”. Third, financial euphemisms, such as “restructuring”, “realignment”, “optimization”, “strategic”, “adjusted”, “normalized”, “pro forma”, “non-recurring”, “one-time”, “temporary”, “transitional”, “evolving”, “dynamic”, and “flexible” [17].

Apart from that, there are additional indicators such as language that shows fraud patterns, including “non-cash”, “goodwill”, “impairment”, “write-down”, “write-off”, “accrual”, “provision”, “reserve”, “allowance”, “contingent”, “derivative”, “off-balance”, “related party”, “subsequent event”, “going concern”, “material weakness”, “deficiency”, and “remediation”. In addition, there are indications that the financial reports detected as fraudulent used passive language to avoid responsibility. There are also overuses of technical terms in the detected fraudulent financial reports, such as “EBITDA”, “depreciation”, “amortization”, “capitalization”, “valuation”, “methodology”, “framework”, “parameters”, “assumptions”, “projections”, “forecasts”, “scenarios”, “sensitivity”, “correlation”, and “volatility” [17].

The above languages were input into a Python system to determine whether the financial reports found to be fraudulent contained such language. The results of XI Hypothesis test indicate that the reports detected as fraudulent used a high amount of passive language, but not significantly so for ambiguous language, indicating fraudulent patterns, and excessive use of technical terms. Therefore, the X1 hypothesis test results were rejected.

Hypothesis X1 can be rejected because it is based on a pre-existing theory. The theory used

to create hypothesis X1 is the fraud triangle theory. One part of the fraud triangle is rationalization, where the fraudster justifies the fraudulent behavior he or she is about to commit. This rationalization is positively related to linguistic patterns because the perpetrator will do anything to maintain his or her image as a person who is right or not committing fraud, including using ambiguous language, passive voice, and also excessive technical terms to avoid being caught. This theory was used as the basis for hypothesis X1, but after testing, with its growing influence, this hypothesis was rejected.

The results of the first hypothesis test indicate that financial reports identified as fraudulent do not exhibit higher linguistic complexity or ambiguity compared to non-fraudulent reports, with an average score of 0.69 for fraudulent reports and 0.98 for non-fraudulent ones.

These findings differ from previous studies [25] which suggests that fraudulent behavior tends to involve the use of more complex or ambiguous language to conceal unfavorable information. This discrepancy may be attributed to Indonesia’s bilingual context, where financial reports are often written in both English and Indonesian, and to the possibility that fraudulent companies in Indonesia deliberately use simpler and more convincing language to make their financial statements appear more transparent.

Based on the test results, although the first hypothesis is not statistically supported, this finding remains consistent with recent literature. A study by [6] showed that fraudulent financial reports often use more complex and ambiguous language, characterized increases the reader’s cognitive load and reduces readability. The use of passive sentences also weakens narrative accountability. Therefore, although the empirical results of this study show an insignificant difference, these linguistics still support the theoretical interpretation that language complexity remains an important element in detecting potential fraud.

4.3.2. Hypothesis X2

Hypothesis X2 relates to the Natural Language Processing model's higher accuracy than traditional audits. This suggests that the Natural Language Processing model is superior to traditional audit methods. Hypothesis X2 compares Natural Language Processing with traditional audits using rule-based keyword detection. The keywords in question are fraud-related and are divided into five categories. First, words that indicate financial distress indicators, such as “loss”, “deficit”, “decline”, “decrease”, “negative”, “downturn”, and “crisis”, [17]. Second, words that indicate risk and

uncertainty such as “risk”, “uncertainty”, “volatile”, “unpredictable”, and “challenging” [17]. Third, words that indicate audit red flags such as “restatement”, “correction”, “adjustment”, “revision”, and “amendment” [17]. Fourth, words that indicate management tone such as “restructuring”, “realignment”, “extraordinary”, “unusual”, and “special” [17]. Finally, words that indicate financial complexity such as “derivative”, “off-balance”, “related-party”, “goodwill”, and “impairment” [17]. The results of the hypothesis test indicate that the natural language processing model is indeed superior to traditional audit methods by 88.9%. Therefore, hypothesis X2 is accepted.

Table 5: X2 Analysis: Enhanced Model Comparison Results

NLP vs Audit Improvement	+88.9%
Enhanced vs Traditional LR	-10.5%
X2 Hypothesis (NLP > Audit)	SUPPORTED
Enhancement Effectiveness	NOT EFFECTIVE

The study by [15] provides empirical evidence that the linguistic tone of financial disclosures—particularly the use of positive, uncertain, and modal expressions—can reveal underlying managerial behavior related to earnings management, a common indicator of fraudulent reporting. The findings support credibility to the idea that financial statements’ linguistic indicators carry important messages that traditional auditing methods could miss. This supports Hypothesis X2 of the current study, which posits that Natural Language Processing models are more effective than traditional auditing methods in detecting potential fraud by capturing complex linguistic patterns.

4.3.3.Hypothesis X3

Hypothesis X3 relates to the inconsistency model between management reports and financial statements. This is where the financial statements of companies found to have committed fraud show inconsistencies between the narrative and the figures. Hypothesis testing for X3 uses sentiment analysis, which categorizes inconsistent language into two categories: positive and negative. The

Hedging Score Analysis		
T-Test	T = -2.171	P = 0.033
Significant	(α = 0.05)	YES

positive words referred to are words such as 'growth', 'increase', 'improve', 'strong', 'positive',

'success', 'gain', 'profit', 'excellent', 'outstanding', 'robust', 'solid', 'healthy', 'favorable', 'promising', 'expansion', 'progress', 'achievement', 'breakthrough', 'advantage', 'opportunity', 'momentum', 'confident', 'optimistic', 'recovery', 'stability', and 'resilience'. Then, the negative words referred to are words such as 'decline', 'decrease', 'loss', 'negative', 'weak', 'poor', 'risk', 'challenge', 'difficult', 'disappointing', 'concern', 'worry', 'uncertainty', 'volatile', 'downturn', 'recession', 'crisis', 'problem', 'issue', 'threat', 'obstacle', 'pressure', 'strain', 'stress', 'deteriorate', 'worsen', and 'struggling'. For the X3 hypothesis test regarding hedging language, the words used for sentiment analysis are such as 'approximately', 'roughly', 'about', 'around', 'estimates', 'believes', 'expects', 'anticipates', 'may', 'might', 'could', 'would', 'should', 'potentially', 'possibly', 'likely', 'probably', 'generally', and 'typically'.

The study utilized a list of sentiment-related words sourced from [17]. The reference includes a comprehensive sentiment dictionary that classifies words according to the emotional polarity which divides into positive and negative. It serves as a basis for quantifying the overall sentiment expressed in the financial reports.

The results of the hypothesis test indicate that the financial statements of companies that have not been proven to have committed fraud have more inconsistent language than those that have been proven to have committed fraud. However, the financial statements of companies that have been proven to have committed fraud have more hedging language than those that have not been proven to have committed fraud. The results of this hypothesis test are significant in terms of hedging analysis with a p-value of 0.033. Meanwhile, the

Inconsistency Score Analysis		
T-Test	T = -0.831	P = 0.409
Mann-Whitney U	U = 486.000	P = 0.801
Significant	(α = 0.05)	NO
Effect Size	Cohen'd	-0.236

inconsistency score analysis is not significant with a p-value of 0.409. Therefore, hypothesis X3 can still be said to be accepted.

Table 6: Inconsistency Score Analysis

Table 7: Hedging Score Analysis

4.4 Comparative Discussion with Previous Studies

Previous research has focused on fraud detection using Natural Language Processing and Machine Learning. According to Muktha & Manish (2025), the language patterns used by online users when interacting on the internet can be used to help detect mobile banking fraud. This aligns with the results of previous studies, which found that language complexity is significantly related to the likelihood of fraud in financial reports. Therefore, this study, along with previous research, demonstrates the effectiveness of Natural Language Processing in identifying linguistic anomalies.

Furthermore, according to Muktha & Manish (2025), research related to automated financial reporting also emphasizes that Natural Language Processing can be used to conduct sentiment analysis and detect anomalies in a company's financial report narrative. This research further strengthens the findings, which show that Natural Language Processing models can outperform traditional audit methods in detecting fraud patterns in financial reports.

The study titled "Intelligent Fraud Detection in Financial Statement Literature Review" [26] conducted a systematic review of more than 150 article that examined the application of machine learning (ML), and data mining (DM) algorithms in detecting financial fraud. This study results indicate that supervised learning approaches, particularly Random Forest, SVM, and XGBoost, are the most common methods with high accuracy rates in identifying fraud indicators.

5. CONCLUSION AND SUGGESTION

This study developed and evaluated a Natural Language Processing (NLP) framework for detecting financial statement fraud in Indonesian public companies, addressing a significant gap in fraud detection research for emerging markets. Using 72 financial reports (20 fraudulent, 52 non-fraudulent) from Indonesia Stock Exchange-listed companies, we integrated TF-IDF-based text representation with linguistic complexity features to train machine learning classifiers.

This research shows that NLP-based models significantly outperformed traditional rule-based audit methods, achieving 77.3% accuracy compared to 40.9% for keyword-based approaches—an 88.9% improvement. This validates NLP as a viable complement to traditional audit procedures. Also, the research results are a little bit contrary to Western fraud literature, fraudulent Indonesian financial statements exhibited simpler, not more

complex, language (ambiguity score: 0.69 vs. 0.98, $p=0.005$). However, fraudulent reports used significantly more hedging language ($p=0.033$), suggesting fraud strategies vary by cultural and regulatory context. Lastly, this research shows that word and phrase patterns (TF-IDF features) contributed 99.9% of predictive power, while linguistic complexity features added minimal value in this dataset. This suggests that lexical choice patterns, rather than sentence-level complexity, are more reliable fraud indicators.

REFERENCES

- [1] A. T. Oyewole *et al.*, "Automating financial reporting with Natural Language Processing: A review and case analysis," *World Journal of Advanced Research and Reviews*, vol. 21, no. 3, pp. 575–589, Mar. 2024. doi:10.30574/wjarr.2024.21.3.0688
- [2] M. A. Al-Faryan, "Agency theory, Corporate Governance and Corruption: An Integrative Literature Review Approach," *Cogent Social Sciences*, vol. 10, no. 1, Apr. 2024. doi:10.1080/23311886.2024.2337893
- [3] M. Altalhan, A. Algarni, and M. Turki-Hadj Alouane, "Imbalanced Data Problem in Machine Learning: A Review," *IEEE Access*, vol. 13, pp. 13686–13699, 2025. doi:10.1109/access.2025.3531662
- [4] M. N. Ashtiani and B. Raahemi, "Intelligent fraud detection in financial statements using machine learning and Data Mining: A Systematic Literature Review," *IEEE Access*, vol. 10, pp. 72504–72525, 2022. doi:10.1109/access.2021.3096799
- [5] A. Badawi, "The effectiveness of Natural Language Processing (NLP) as a processing solution and semantic improvement," *International Journal of Economic, Technology and Social Sciences (Injects)*, vol. 2, no. 1, pp. 36–44, Apr. 2021. doi:10.53695/injects.v2i1.194
- [6] I. Bhattacharya and A. Mickovic, "Accounting fraud detection using contextual language learning," *International Journal of Accounting Information Systems*, vol. 53, Jun. 2024. doi:10.1016/j.accinf.2024.100682
- [7] P. Boulieris, J. Pavlopoulos, A. Xenos, and V. Vassalos, "Fraud detection with natural language processing," *Machine Learning*, vol. 113, no. 8, pp. 5087–5108, Jul. 2023. doi:10.1007/s10994-023-06354-5
- [8] C. Stryker and E. Kavlakoglu, "What is Artificial Intelligence (AI)?," IBM,

- <https://www.ibm.com/think/topics/artificial-intelligence>
- [9] P. M. DECHOW, W. GE, C. R. LARSON, and R. G. SLOAN, "Predicting material accounting misstatements," *Contemporary Accounting Research*, vol. 28, no. 1, pp. 17–82, Jan. 2011. doi:10.1111/j.1911-3846.2010.01041.x
- [10] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," *Procedia Computer Science*, vol. 17, pp. 26–32, 2013. doi:10.1016/j.procs.2013.05.005
- [11] R. B. Hadiprakoso, H. Setiawan, R. N. Yasa, and Girinoto, "Text preprocessing for Optimal Accuracy in Indonesian sentiment analysis using a deep learning model with word embedding," *AIP Conference Proceedings*, vol. 2879, 2023. doi:10.1063/5.0126116
- [12] T. D. Hardjanto, "Hedging through the use of modal auxiliaries in English academic discourse," *Jurnal Humaniora*, vol. 28, no. 1, pp. 37–50, May 2016. doi:10.22146/jh.v28i1.11412
- [13] J. C. Obi, "A comparative study of several classification metrics and their performances on data," *World Journal of Advanced Engineering Technology and Sciences*, vol. 8, no. 1, pp. 308–314, Feb. 2023. doi:10.30574/wjaets.2023.8.1.0054
- [14] B. Li, Y. Hou, and W. Che, "Data augmentation approaches in Natural Language Processing: A Survey," *AI Open*, vol. 3, pp. 71–90, 2022. doi:10.1016/j.aiopen.2022.03.001
- [15] S. Li, G. Wang, and Y. Luo, "Tone of language, financial disclosure, and Earnings Management: A textual analysis of form 20-F," *Financial Innovation*, vol. 8, no. 1, May 2022. doi:10.1186/s40854-022-00346-5
- [16] X. Liu, "Empirical analysis of financial statement fraud of listed companies based on logistic regression and Random Forest algorithm," *Journal of Mathematics*, vol. 2021, pp. 1–9, Dec. 2021. doi:10.1155/2021/9241338
- [17] Marketing Communications: Web | University of Notre Dame, "Loughran-McDonald master Dictionary W/ sentiment word lists," Software Repository for Accounting and Finance, https://sraf.nd.edu/loughranmcdonald-master-dictionary/?utm_source
- [18] A. A. Olayinka, "Financial statement analysis as a tool for investment decisions and assessment of companies' performance," *International Journal of Financial, Accounting, and Management*, vol. 4, no. 1, pp. 49–66, Jun. 2022. doi:10.35912/ijfam.v4i1.852
- [19] A. W. Pradana and M. Hayaty, "The effect of stemming and removal of stopwords on the accuracy of sentiment analysis on Indonesian-language texts," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, pp. 375–380, Oct. 2019. doi:10.22219/kinetik.v4i4.912
- [20] PricewaterhouseCoopers, "Global Economic Crime Survey 2024," PwC, <https://www.pwc.com/gx/en/services/forensics/economic-crime-survey.html>
- [21] J. Qiu, "An analysis of model evaluation with cross-validation: Techniques, applications, and recent advances," *Advances in Economics, Management and Political Sciences*, vol. 99, no. 1, pp. 69–72, Sep. 2024. doi:10.54254/2754-1169/99/2024ox0213
- [22] H. Sheikh, C. Prins, and E. Schrijvers, "Mission ai," *Research for Policy*, 2023. doi:10.1007/978-3-031-21448-6
- [23] G. M. M. Sujeewa, M. S. A. Yajid, A. Khatibi, S. M. F. Azam, and I. Dharmaratne, "THE NEW FRAUD TRIANGLE THEORY - INTEGRATING ETHICAL VALUES OF EMPLOYEES," *International Journal of Business, Economics and Law*, vol. 16, no. 5, pp. 52–57, Aug. 2018.
- [24] "UU No. 27 Tahun 2022," Database Peraturan | JDIH BPK, <https://peraturan.bpk.go.id/Details/229798/u-u-no-27-tahun-2022>
- [25] Z. J. Yu, "Financial report readability and accounting conservatism," *Journal of Risk and Financial Management*, vol. 15, no. 10, Oct. 2022. doi:10.3390/jrfm15100454
- [26] M. N. Ashtiani and B. Raahemi, "Intelligent fraud detection in financial statements using machine learning and Data Mining: A Systematic Literature Review," *IEEE Access*, vol. 10, pp. 72504-72525, 2022. doi:10.1109/access.2021.3096799