Scientax
Jurnal Kajian Ilmiah Perpajakan Indonesia

djp

VOLUME 7 NO. 2 | APRIL 2026 Journal Page: ejurnal.pajak.go.id ISSN 2686-5718

Utilizing Natural Language Processing and Logistic Regression Model for Automated Detection and Classification of Tax Objects in Tax Audit Processes

Bagas Dwi Suryo Wibowo[a], Wishnu Kusumo Agung Erlangga[b*]
[a] Directorate General of Taxes, Jakarta, Indonesia. Email: bagas.dwisuryowibowo@pajak.go.id
[b] Directorate General of Taxes, Jakarta, Indonesia. Email: wishnu.e@pajak.go.id

* Corresponding author: wishnu.e@pajak.go.id

ABSTRACT

The Directorate General of Taxes (DGT) plays a key role in collecting state revenue in Indonesia, with tax audits being one of its core functions. However, the sheer number of taxpayers far exceeds the number of tax auditors. Despite the use of recognized accounting standards in the preparation of financial statements, the selection of account names in taxpayers' financial statements often poses a unique challenge for tax auditors. Each company may have different standards when naming its accounts, even when referring to the same type of transaction or object. This is where Natural Language Processing (NLP) can help detect and classify tax objects from the general ledger. In this research, we developed and compared machine learning models to automatically classify tax objects and fiscal corrections. This research used real data consisting of 461,776 rows of general ledger entries and was processed using quantitative methods. By using real data, the developed model has an advantage over models trained on artificial data. We compared results from Logistic Regression, K-Nearest Neighbors, and Naïve-Bayes algorithms and found that the first-mentioned algorithm suits the best metrics. The Logistic Regression model achieved a precision level of 99% in detecting both types of tax objects and fiscal corrections from financial statements. The findings of this research are expected to assist tax authorities in detecting the presence or absence of tax objects and fiscal corrections in financial statements, thereby enabling various functions within DGT to operate more efficiently and effectively.

Keywords: natural language processing, logistic regression, tax income, tax audit, machine learning, web application

1. INTRODUCTION

State revenue can originate from various sources, including taxes and other forms of income. The proportion of tax revenue compared to non-tax revenue varies across countries. However, numerous studies suggest that taxes constitute a significant portion of state revenue in many countries, including Indonesia (Purwowidhu, 2023; Cox et al., 2024; Morar, 2015). The growth of state evenue from this sector has consistently increased over time.

Consequently, the volume of data that tax-collecting institutions, such as the Directorate General of Taxes, must handle continues to grow. This data, both structured and unstructured, accumulates in large quantities, forming what we know as big data.

DOI: 10.52869/t7ygg336
Received: January 16, 2025; Revised: March 31, 2026; Accepted: March 31, 2026; Published: April 30, 2026

Scientax: Jurnal Kajian Ilmiah Perpajakan Indonesia is Sinta 3 Journal (https://sinta.kemdikbud.go.id/journals/profile/9121)
How to Cite:
Wibowo, B. D. S., & Erlangga, W. K. A. (2026). Utilizing natural language processing and logistic regression model for automated detection and classification of tax objects in tax audit processes. *Scientax: Jurnal Kajian Ilmiah Perpajakan Indonesia, 7*(2), 188-207. https://doi.org/10.52869/t7ygg336

1.1 Institutional Settings

Taxes were not the primary source of state revenue before 2000. The sale of oil and natural gas was still the dominant source of revenue until the Indonesian government initiated tax reforms in 2000, which continued until 2007. Since then, the share of revenue from the tax sector has steadily increased, although achieving the optimal level of tax revenue has taken time. One of the reasons for this shortfall is the lack of integration among the functions performed by the Directorate General of Taxes (DGT), including service, supervision, auditing, collection, and law enforcement functions.

This condition is supported by the Theory of Fiscal Capacity, which posits that the ability to mobilize revenue is not solely determined by a country's economic base but also by the institutional quality and integration of its tax administration. The lack of coordination among key functions, such as service, supervision, auditing, collection, and enforcement, can hinder the efficiency and effectiveness of revenue collection, thereby limiting fiscal capacity (Besley et al., 2010).

Figure 1 illustrates that Indonesia's tax ratio has not shown significant improvement. From the onset of tax reforms in 2007 to 2022, Indonesia's tax ratio has remained around 12%. The lowest percentage (10.1%) was recorded in 2020, while the highest figure (13.0%) was achieved in 2008. Nevertheless, this average tax ratio is still much lower than the average tax ratio in Asia-Pacific countries, which was around 34% during the same period and peaked at 44% in 2019 (Organisation for Economic Co-operation and Development [OECD], 2024).

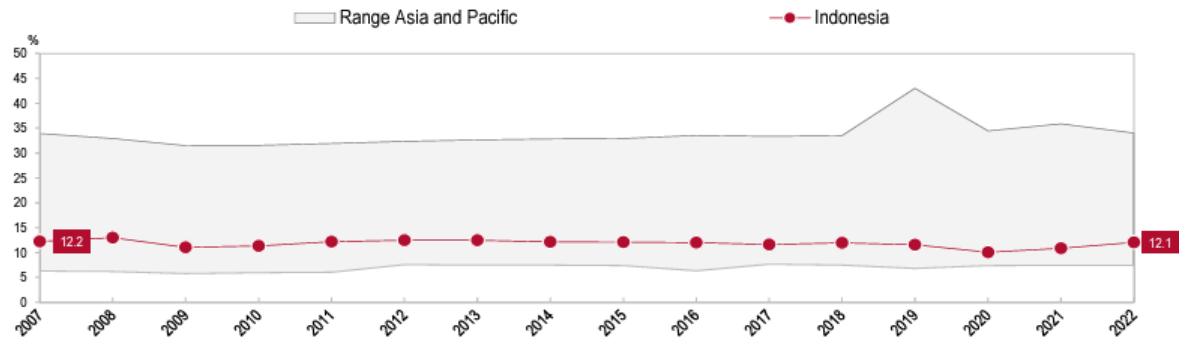
Based on Figure 2, we can compare Indonesia's tax structure with other regions, such as Asia-Pacific, Africa, Latin America and the Caribbean (LAC), and OECD countries. Several key points can be highlighted from the figure.

Unlike Asia-Pacific countries that rely on goods and services tax revenue as their primary source (25%), as well as LAC (28%) and African countries (28%), Indonesia relies on corporate income tax (CIT) as its main source (29%). This figure is significantly higher than the average in Asia-Pacific (21%), Africa (19%), LAC (19%), and OECD countries (10%) (OECD, 2024).

Indonesia's tax structure shows a high dependence on corporate income tax compared to other regions. This is due to the dominance of the corporate sector in Indonesia's economy. Conversely, the contribution of personal income tax and taxes on goods and services remains relatively low. To increase tax revenue, Indonesia needs to consider diversifying its tax structure by enhancing the efficiency of tax collection.

An examination of taxpayer conditions raises questions about whether the DGT has sufficient personnel to perform its functions effectively, or if alternative solutions are necessary

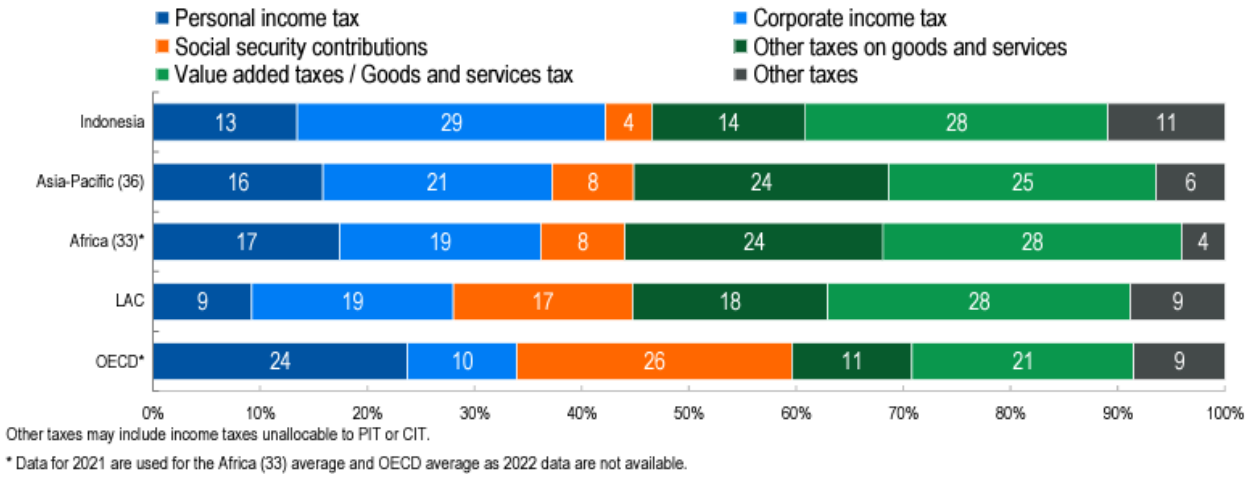
Figure 1
Tax-to-GDP Over Time



Note. Source: OECD (2024)

Figure 2

Tax Structure Compared to The Regional Averages



Note. Source: OECD (2024)

regardless of the number of employees. For instance, the target for formal taxpayer compliance in 2024 is 19,273,374 annual tax returns, with the number of annual tax returns filed by individual and corporate taxpayers shown in Table I: Formal Compliance Rate of Taxpayers.

The formal compliance rate in Table 1 indicates the percentage of taxpayers who submit their annual tax returns on time. A higher compliance rate is often indicative of a more effective and efficient tax administration system. However, achieving high compliance rates is challenging and requires substantial administrative capacity, effective enforcement, and taxpayer education.

As of April 30, 2024, the formal compliance rate reached only 73.58%, while the formal compliance target set by the Directorate General of Taxes (DGT) is 83.22%. Currently, the number of Account Representatives (AR) in the DGT is 11,028 employees, whereas the number of tax auditors is 6,107 employees. Therefore, the ratio between the Annual Tax Returns reported and ARs is (1:95), and for Tax Auditors is (1:171). This assumes that each employee works individually rather than in a team. This ratio is considered challenging to implement each year effectively.

The DGT has not officially implemented Artificial Intelligence (AI) in its business processes and related functions. However, the emphasis on the importance of data in decision-making has become a priority, as reflected in the Data-Driven-Organization framework (DGT, 2022). Essentially, the DGT will prioritize decision-making and the execution of business processes based on valid and high-quality data to avoid errors and biased results.

AI is increasingly expected to support operational activities across various sectors by improving efficiency, productivity, and decision-making (Plathottam et al., 2023), including the taxation sector. AI has become a reliable tool in tax administration. Tax authorities around the world benefit from the reduction of lengthy and

Table 1

Formal Compliance Rate of Taxpayers

Types of Tax Returns	2023	2024	YoY
Corporate Income Tax Return	944.264	1.044.911	10,7%
Individual Income Tax Return	12.295.752	13.141.719	6,9%
Total	13.240.016	14.186.630	7,1%

Note. Source: Kementerian Keuangan Republik Indonesia (2024)

complicated processes in their daily work through AI technology implementation. The costs incurred by these tax authorities can also be significantly reduced. Therefore, tax authorities can rely on AI, particularly to assist in the analysis process, considering the big data that emerges from taxpayer returns when compared with external data (Shakil & Tasnia, 2022).

For example, E-filing is an integrated service with DGT Online designed to facilitate taxpayers in submitting their Annual Individual Income Tax Returns or Corporate Income Tax Returns. E-filing has been widely implemented by tax institutions in various countries. The goal is to provide convenience in reporting so that taxpayers are more motivated to meet their formal tax compliance correctly.

In theory, lower costs for taxpayers to fulfill their tax obligations will encourage them to become compliant taxpayers. This system is also expected to reduce physical interaction between tax officials and taxpayers, potentially decreasing the likelihood of corruption offenses by tax officials (Kochanova et al., 2020).

Beyond the E-filing system, various countries have also used AI to facilitate taxpayers in meeting their obligations. The Singapore Inland Revenue Service (IRAS) uses "Joanna," a chatbot application designed to answer user questions. Additionally, Joanna can detect the type of tax service requested by taxpayers and assist in resolving related applications.

In the private sector, KPMG has launched "Tax Service," an AI-based tax service product that helps companies in China automatically resolve various tax compliance issues.

The application of AI in various tax functions can enhance tax management, leading to the transformation of complex business processes into more efficient ones. This can raise awareness that good tax services can positively impact taxpayers' trust in the integrity of a country's tax institution (Li, 2022).

All of this can be achieved if AI developers can understand the correct context. To understand the correct context, given the diversity of account names recorded in taxpayers' financial statements, a sophisticated NLP model is needed. This research was conducted using two large datasets. This research focuses on augmenting the NLP model with two main aspects in the tax return and financial statements: fiscal corrections and tax objects.

1.2 Issues in Tax Object Identification

1.2.1 Language Use in Financial Reports

The language used in financial reports can significantly impact the accuracy of tax object identification. For instance, reports written in Indonesian may differ in structure and terminology from those in other languages, potentially leading to discrepancies in tax object classification.

1.2.2 Manual Coding Risks

Developing program code manually is prone to errors, which can affect the accuracy of tax object identification. Automated solutions and machine learning models can help mitigate these risks by providing more consistent and reliable classifications.

1.2.3 Unique Characteristics of Taxpayers

The financial characteristics of individual and corporate taxpayers differ significantly. These differences must be carefully considered to avoid misinterpretation and ensure the accuracy of tax object identification. The model should be robust enough to handle the unique attributes of each taxpayer type, minimizing false positives and false negatives.

By addressing these factors, the efficiency and effectiveness of tax object identification can be significantly improved, aiding in better tax administration and compliance.

1.3 The Need for AI Adoption

Proper management of big data can provide valuable knowledge for many entities or organizations, including DGT. This knowledge can help entities extract meaningful insights that are useful for decision-making. Decision-making processes are greatly enhanced by the application of Artificial Intelligence (AI) (Alasmri & Basahel, 2022), particularly in supervisory and audit functions.

Many tasks performed by tax officials in these functions are repetitive, time-consuming, and labor-intensive (Veenendaal, 2023). This is where AI plays a crucial role in assisting humans, making complex and repetitive tasks more efficient. The adoption of AI has become quite prevalent, with applications such as Apple's Siri, IoT devices, adaptive Google Search, and even self-driving cars becoming increasingly common. This trend is also evident in the current and future functions of taxation. AI is predicted to be widely used to estimate audit risks, detect errors in reporting, and potentially classify transactions automatically based on their type and tax object (Supriadi, 2024). Two key aspects with significant potential for improvement are efficiency in terms of labor, time, and cost, and the effectiveness of the activities performed.

Generally, AI applications integrate technologies that mimic human capabilities. AI encompasses various types and levels, ranging from Machine Learning (ML) to Deep Learning (DL), all designed to provide computers with the ability to learn, perceive, recommend, or make decisions similar to humans. Thus, advanced computer analysis can be achieved with AI technologies, including the application of Natural Language Processing (NLP).

The preparation of financial reports and tax detection processes is highly complex, while AI, particularly NLP, is highly technical. The use of NLP to achieve the desired levels of efficiency and effectiveness, which are the two main goals of AI development, is essential for classifying data into the correct categories.

Currently, there is a significant gap within the Directorate General of Taxes (DGT), as AI-based modeling has not yet been widely implemented in tax supervision or tax audit processes. The urgency of implementing AI in various functions and tasks of the Directorate General of Taxes (DGT) is justified. The use of English in taxpayers' financial reports often poses a challenge for tax auditors, particularly in the context of NLP. Many multinational companies operating in Indonesia use English as the primary language in their financial reports. This condition may create challenges in understanding and identifying tax objects due to the differences in terminology and the structure of the language used.

Our research introduces a novel approach by integrating both methods to detect and classify tax objects from unstructured audit texts. This study uniquely addresses a practical gap by developing a model that is not only effective but also user-friendly for non-tech-savvy tax auditors, enabling broader adoption in real-world tax audit environments.

This study aims to demonstrate how an NLP model can be used to detect tax objects accurately and instantly based on diverse account names in taxpayers' financial reports, thereby significantly improving efficiency and effectiveness in one of the tax functions, namely tax auditing. With high accuracy in detecting tax fraud by DGT, the potential for tax losses can be minimized.

This research aims to study the patterns of word usage in taxpayers' financial reports. This study contributes to AI-based tax analysis by enabling the detection of tax objects and fiscal corrections in financial statements. After completing all processes, the developed model can extract various tax objects with minimal false positives and false negatives. The model development offers at least two advantages. Firstly, this research uses real data from the DGT, providing an accurate depiction of taxpayers' conditions in Indonesia. Secondly, at a practical level, the model development scheme can be applied to other functions in the DGT, providing actionable insights based on the tasks and authorities of each function.

In our opinion, the planned implementation of Core Tax by DGT also requires

real-time data sharing between tax officials and taxpayers. Transitioning to real-time reporting will enable DGT to monitor transactions more closely and respond quickly to discrepancies or suspicious activities. Moreover, as Pratama and Darono (2022) stated, combining customs and taxation data using a text analytics approach can be implemented in relevant use cases.

This paper consists of six main sections. Section 2 explains the theoretical framework. Section 3 describes the data approach and analysis used. Section 4 presents the research results and discussion. Section 5 provides the conclusion. Finally, Section 6 highlights the practical implications and limitations faced by the authors.

The lack of standardized naming conventions for accounts in taxpayers' financial reports adds to the complexity of identifying tax objects. Each company may have its own way of naming accounts, which can differ even for similar transactions or objects. This inconsistency necessitates a robust NLP model to accurately map and classify accounts without overfitting. Therefore, optimizing NLP to handle variations in account names in financial reports is crucial, including processes for text translation and normalization to ensure the analysis system can recognize the data effectively.

NLP has shown significant potential in transforming tax administration processes. For example, NLP enables tax auditors to quickly identify compliant or non-compliant sections of taxpayers' returns and ensure their adherence to existing regulations. With sufficient data from taxpayers, it is conceivable that the developed model could predict taxable objects and match them against taxpayers' returns.

2. THEORETICAL FRAMEWORK

Several factors must be considered when extracting words from taxpayers' financial statements. First, the use of Indonesian or other languages can yield different results in tax object identification. Second, manual coding programs are highly prone to errors. Third, the unique financial characteristics of individual and corporate taxpayers are susceptible to misinterpretation. Therefore, the model must carefully consider the potential for false positives and false negatives.

2.1 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a deep learning framework used to analyze text with the aid of computers (Jurafsky & Martin, 2008). The objective of this study is to identify the appropriate framework and model for accurately and automatically classifying tax objects amidst the diverse and non-standardized naming conventions used in financial statements.

According to Lozano and Ippolito (2024), Marinho (2023) conducted a study using 10,000 invoices from a government agency to calculate similarity levels between invoice goods descriptions and Mercosur item nomenclature. The results indicated low similarity and inconsistencies in the item naming conventions. These complex findings can be effectively addressed using NLP technology. Manual classification was also performed by Santos (2022) and Darrazão et al. (2023), using classification algorithms to group invoices. The Ministry of Public Administration in Paraiba provided 30,000 invoices, which were manually processed for classification purposes.

All these tasks were performed using NLP techniques, which involve the use of human language and can be applied in various fields such as semantic analysis and speech recognition (Steedman, 1996). One important step in text processing is feature extraction, where raw text is transformed into numerical representations that can be processed by machine learning algorithms. A commonly used technique in text feature extraction is the Bag of Words model, though it has limitations, as it only counts word occurrences without considering their context within a document.

2.1.1 Text Pre-Processing

The text pre-processing stage involves several steps to ensure the quality of the data used. High-quality data will improve the performance of the deep learning model, increasing the likelihood of extracting valuable insights.

Tokenization is the initial step in NLP for understanding the content of taxpayers' financial statements. Tokenization divides a sentence into individual words. The next step in text processing is normalization (stemming), where prefixes and suffixes are removed. Finally, lemmatization is applied to reduce words to their base form, while stop words are removed separately.

2.1.2 Feature Extraction

Text pre-processing results in data consisting of frequently occurring words in taxpayers' financial statements. This data can be accurately understood when placed in the appropriate context. Therefore, the context used to avoid significant errors in utilizing these words depends on the type of the taxpayers' business.

The extracted text, which has undergone pre-processing, is grouped based on the type of the taxpayer's business. This produces a collection of words representing each group.

Term Frequency–Inverse Document Frequency (TF-IDF) is a numerical statistic that reflects the importance of a word in a collection of documents or a corpus. Unlike the simple Bag of Words model, which only counts word occurrences, TF-IDF evaluates a word's importance by considering its frequency in a specific document (Term Frequency) and its frequency across all documents in the corpus (Inverse Document Frequency). The Term Frequency (TF) measures how frequently a term appears in a document, normalized by the total number of terms in that document. The Inverse Document Frequency (IDF) component reduces the weight of terms that frequently appear across many documents and increases the weight of terms that appear less frequently. The TF-IDF value is calculated by multiplying the TF and IDF values, giving higher scores to words that are important to specific documents but not common across all documents (Ambi, 2022).

This method of feature extraction is particularly effective for text classification tasks because it highlights the most relevant words and reduces the impact of common, less informative words. By applying TF-IDF, we convert text data into numerical features that can be used by machine learning algorithms, enhancing their ability to distinguish between different classes based on textual content. For the classification tasks using Logistic Regression, K-Nearest Neighbors (KNN), and Naive Bayes, the TF-IDF representation helps improve model performance by providing a more nuanced understanding of the text compared to a simple Bag of Words approach.

2.2 Logistic Regression

Based on the theory and the advantages and disadvantages of various classification models discussed by Wendler & Grottrup (2021), logistic regression was chosen for this study. This algorithm is widely used in many fields and was awarded the Nobel Prize in 2000.

Logistic regression differs from linear regression as it transforms the projection from a linear combination of independent variables into specific dependent variable values. Logistic regression uses a set of independent variables to predict binary outcomes (0 or 1). A value of 0 indicates that the dependent event does not occur (see Figure 3).

The logistic regression model estimates the probability that an observation belongs to a given class. The logistic function is expressed as follows:

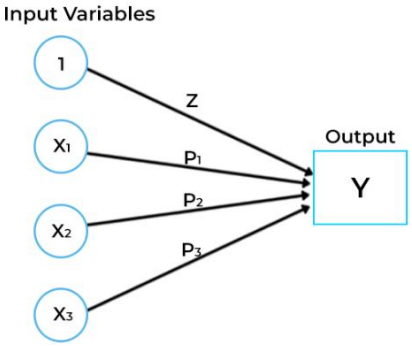
$$P(Y=1|X)=\frac{e\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta n\,xn}{1+e\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta nxn} \quad (1)$$

where P(Y=1|X) denotes the probability that an observation belongs to a particular class, x1,x2,...,xn represent the predictor variables, β0 is the intercept, and β1,β2,...,βn are the regression coefficients. In linear regression, the response variable is continuous, while the dependent variable is binary in logistic regression. Metrics such as mean squared error or R-squared are commonly used to evaluate linear regression models' performance. In contrast, accuracy, precision, and recall are typically used to measure logistic regression models' performance.

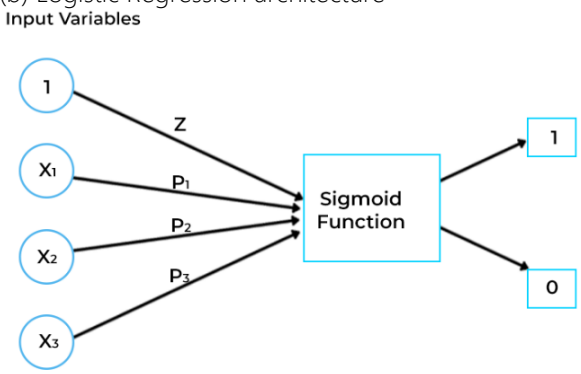
Figure 3

Comparison of linear regression and logistic regression architectures

(a) Linear Regression architecture



(b) Logistic Regression architecture



Note. Source: Kanade (2022)

This study employs Multinomial Logistic Regression. Multinomial logistic regression is used when the dependent variable has more than two categories without a clear order. This means the categories have no inherent order and are mutually exclusive.

Logistic regression produces an S-shaped curve known as the Sigmoid curve, generated by the Sigmoid function. The Sigmoid function is a mathematical function that converts input values to values between 0 and 1, making it useful for binary classification and logistic regression.

The logistic regression model has advantages over the other two models as it does not require normally distributed variables (Glantz & Slinker, 2001; Lemeshow & Hosmer, 1982). Harris (2021) explained that the binary logistic regression model relies on assumptions, including independent observations, no perfect multicollinearity, and linearity. Binary logistic regression relies on three underlying assumptions:

1) The observations must be independent.
2) There must be no perfect multicollinearity among independent variables.
3) Continuous predictors are linearly related to a transformed version of the outcome (linearity).

2.3 Limitations of TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a popular method for feature extraction in text analysis and natural language processing. However, despite its advantages, TF-IDF has several limitations:

2.3.1 Simplicity and Context Ignorance

TF-IDF treats words as independent entities, ignoring the context in which they appear. It does not consider the semantic relationships between words, which can be crucial for understanding the meaning of a text. Therefore, this algorithm lacks contextual understanding.

Moreover, TF-IDF does not account for the order of words in a document. For example, the phrases "New York" and "York New" would have the same representation despite having different meanings.

2.3.2 Sparse Representations

With large vocabulary sizes, TF-IDF results in high-dimensional vectors, especially when dealing with large vocabularies. This can lead to sparse representations where many entries are zero, making the vectors inefficient to process.

Calculating TF-IDF for a large corporation can be computationally expensive, both in terms of time and memory usage. This can be a limitation when scaling to very large datasets.

2.3.3 Static Weights

Once TF-IDF weights are calculated for a corpus, they remain static. This means that any new document added to the corpus will not affect the existing TF-IDF weights, potentially reducing the adaptability of the model.

Over time, the importance of certain terms may change, but TF-IDF does not dynamically update weights to reflect these changes. This can lead to outdated representations if the corpus evolves.

2.3.4 Handling Rare and Common Words

Regarding common and rare words, while TF-IDF aims to reduce the influence of very common words, it can also overemphasize rare words that may not be relevant to the document's overall meaning. Common words that are important for certain contexts might be underrepresented if their document frequency is high across the corpus. This can lead to the loss of valuable information.

2.3.5 Document Length Bias

In terms of length normalization, TF-IDF does not inherently normalize for document length. Longer documents tend to have higher term frequencies, which can skew the TF-IDF values if not properly normalized. Without normalization, longer documents may appear more significant than shorter ones, even if the shorter documents are more relevant to the query or analysis. Thus, this results in a bias toward longer documents.

2.3.6 Limited Handling of Polysemy and Synonymy

TF-IDF cannot distinguish between different meanings of the same word (polysemy). For example, the word "bank" in "river bank" and "savings bank" would be treated the same, leading to potential misinterpretation.

Unfortunately, TF-IDF does not recognize different words with similar meanings (synonyms). For instance, "car" and "automobile" would be treated as completely unrelated terms, missing the semantic similarity between them.

2.3.7 Dependence on Preprocessing

The effectiveness of TF-IDF heavily relies on the quality of text preprocessing steps such as tokenization, stemming, and stopword removal. Inadequate preprocessing can result in suboptimal feature representations. The quality of the preprocessing steps must be carefully considered.

While TF-IDF remains a valuable tool for many text analysis tasks, these limitations highlight the need for more sophisticated techniques, such as word embeddings (e.g., Word2Vec, GloVe) or transformer-based models (e.g., BERT), which can capture richer semantic information and contextual relationships between words.

3. RESEARCH METHODOLOGY

This study adopts a quantitative research approach. One of the most notable advantages of this research is the utilization of a large dataset, consisting of 461,776 raw data rows from a single column in the taxpayers' general ledger. This data underwent an initial data preparation stage, which included data cleansing to ensure its usability throughout the entire process. The data cleansing process involved removing stopwords, deleting empty rows, and eliminating duplicate data. As a result, 66,732 rows of clean data with one column were obtained, representing 14.45% of the total raw data.

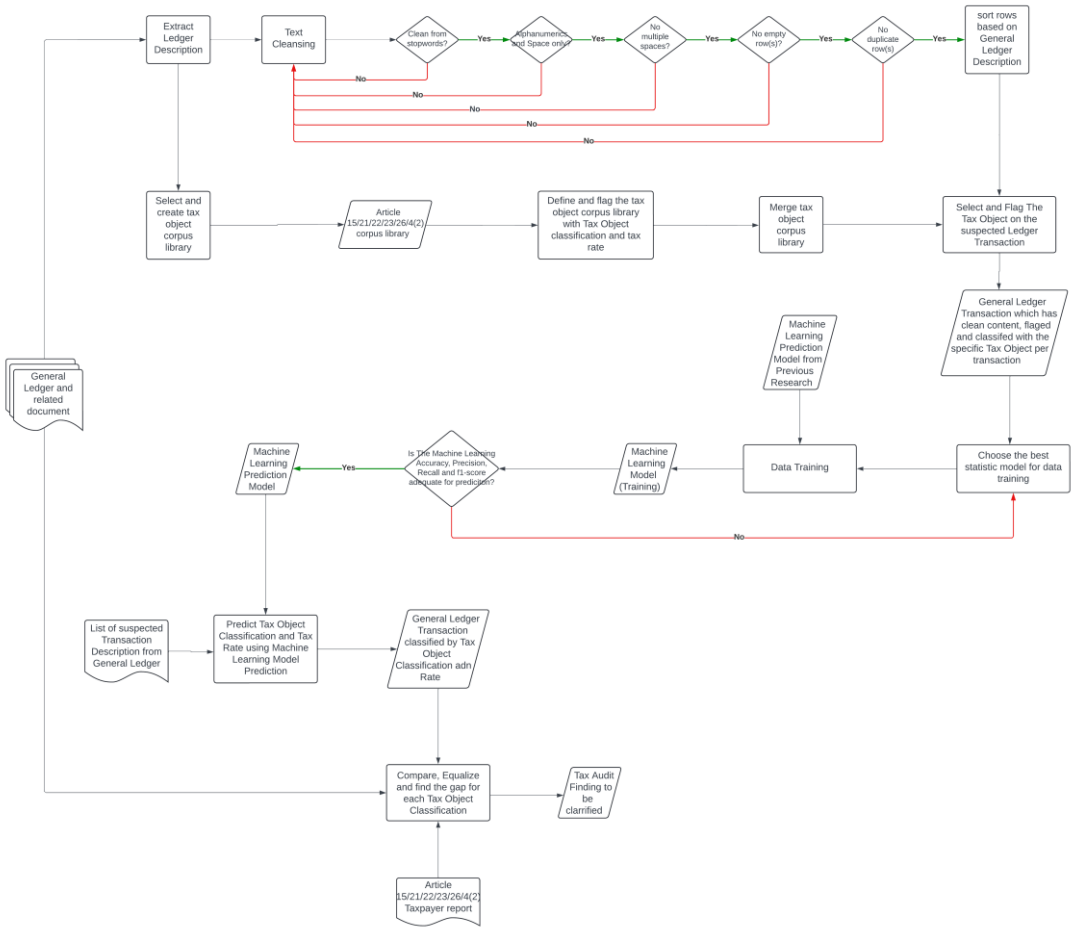
The schema for developing the Tax Object Finder algorithm in this research is illustrated in Figure 4, Tax Object Finder Algorithm. The Tax Object Finder algorithm is designed to detect and classify tax objects within general ledger transactions using Natural Language Processing (NLP) techniques.

3.1 *Data Processing Techniques*

This research was conducted using Python programming language version 3.11. The chosen framework for this study is the Cross Industry Standard Process for Data Mining (CRISP-DM). This

Figure 4

Proposed Detection Machine Development Framework



Note. Source: Developed by the authors

framework was selected because it provides a balanced approach between the variability in account names found in actual financial reports and its adaptability across various industry sectors.

The research is currently in the model-building stage with a linear process. The authors focused on data understanding and continuous model development. The data processing techniques in this study consisted of seven stages: data collection, data extraction, data cleansing, data training, data validation, data analysis, and tax finding.

1. Data collection
   At this stage, the study collected general ledger data and related supporting documents containing taxpayer transaction records. These data served as the primary source for identifying transaction descriptions relevant to tax object classification.
2. Data extraction
   The textual descriptions of transactions were extracted from the general ledger to isolate the information required for natural language processing and classification.
3. Data cleansing
   The extracted text data were preprocessed to improve consistency and analytical quality. This stage included stopword removal, elimination of non-alphanumeric characters except spaces, removal of multiple spaces, deletion of empty rows, removal of duplicate records, and sorting of entries based on general ledger descriptions.
4. Data training
   A tax object corpus library was constructed from the cleaned transaction descriptions and labeled according to tax object categories and applicable tax rates. The textual data were then transformed into numerical representations

using text classification techniques, including Bag of Words and TF-IDF. Based on these representations, the most appropriate statistical model was selected and trained to classify tax objects from ledger descriptions.

5. Data validation
   The trained model was evaluated using accuracy, precision, recall, and F1-score to assess its classification performance. A minimum threshold of 0.80 was used as the acceptance criterion to determine whether the model was sufficiently reliable for predictive use.
6. Data analysis
   After validation, the model was applied to classify tax objects and applicable tax rates in general ledger transactions. The prediction output was then compiled into a list of transactions requiring further analytical review.
7. Tax finding
   In the final stage, the model predictions were compared with taxpayer-reported information to identify inconsistencies, classification gaps, and potential audit findings. These results were used as analytical support in the tax audit process.

3.2 Logistic Regression Model

This study utilised a quantitative text classification methodology to identify tax objects and fiscal adjustments from general ledger transaction descriptions. After preprocessing the text and extracting features, the cleaned transaction descriptions were turned into numerical vectors using the TF-IDF representation. The classification model used these vectors as predictor variables and the predefined class label as the target variable. The predefined class label could be either the tax object category or the fiscal correction category.

We chose Logistic Regression as the main classification algorithm because the dependent variable in this study is categorical, not continuous. Logistic regression, on the other hand, tries to figure out the chance that an observation belongs to a certain class based on a set of predictor variables. Linear regression, on the other hand, tries to predict continuous outcomes. The model was implemented as a multinomial logistic regression because this study had more than two class labels.

The logistic regression model calculates the likelihood of class membership mathematically

$$P(Y=1|X)=\frac{1}{1+e^{-(\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta nxn)}} \quad (2)$$

as follows:

where P(Y=1|X) denotes the probability of an observation belonging to a given class, x1, x2,…,xn represent the predictor variables derived from the TF-IDF features, β0 is the intercept, and β1, β2,…,βn are the model coefficients estimated during training.

In this research, the model learned the relationship between textual patterns in general ledger descriptions and the corresponding class labels. Since this study involved more than two categories, logistic regression was implemented in a multinomial setting, and the final predicted class was determined based on the highest estimated probability.

The model figures out the chance that a transaction description belongs to each class. The final predicted label is based on the class with the highest chance. This is how the model learns how to connect the patterns in general ledger descriptions to the right tax object or fiscal correction labels.

The dataset utilised in this study comprised 66,732 sanitised records derived from the preprocessing of 461,776 raw entries. The dataset that was ready was split into training and testing data in a 70:30 ratio. We used the training set to figure out the parameters for logistic regression and the testing set to see how well the model worked.

3.3 Analytical Approach

NLP considers the frequency of word occurrences in the input text. The importance of these words relative to the entire text is also measured based on the appropriate context, in this case, tax object classification and fiscal corrections. Although the

language used in financial reports is not always linear due to complex structures and strong context dependence, proper extraction techniques can address this issue. To ensure that context and linearity are connected in the classification process, the target variables must be distinguishable through linear regression.

Linear regression focuses on the existence of numerical values or continuous target variables and attempts to estimate the functional relationship between predictors and the target variable. However, in classification models, the target variable is categorical, making linear regression unsuitable. The main issue lies in predicting the relationship between predictor words and target categories rather than producing the target itself. Therefore, logistic regression is chosen as the algorithm for this research (Wendler & Grottrup, 2021).

Logistic regression is preferred due to its capability in handling binary outcomes effectively and its mathematical approach to modeling the probability of categorical outcomes. This choice aligns with the goal of accurately classifying tax objects and ensuring robust predictions in various tax-related scenarios.

By following this framework, the research aims to develop a reliable and effective machine learning model that enhances the classification of tax objects in financial transactions from many sources (general ledgers, income statements, etc.) and many datatypes, as long as it is recognizable by the Pandas library in Python. Therefore, the results may ultimately contribute to more efficient and accurate tax administration processes.

4. RESULTS AND DISCUSSIONS

4.1 Logistic Regression

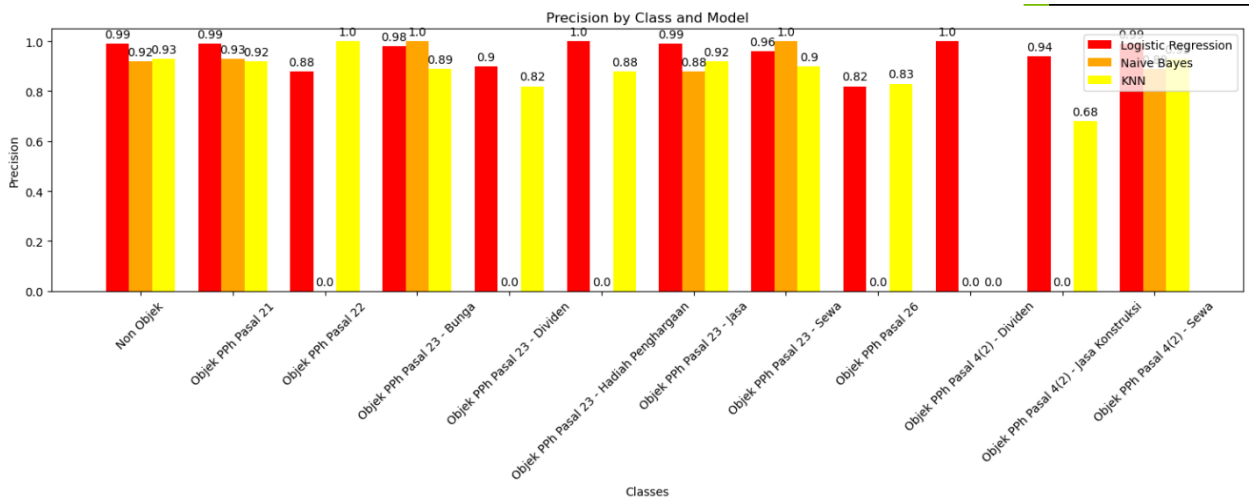
Figure 5 presents evaluation metrics for the developed model, covering the training and testing phases. The prepared dataset was divided into two parts, with 70% used for training data and the remaining 30% for testing data.

The classification report provides detailed evaluation metrics for each class present in the dataset. This report helps evaluate how well the model performs in classifying each tax object category. For instance, if there are several types of tax objects, the report will show precision, recall, and F1-score for each tax object type separately, allowing the identification of classes that may require further model improvement.

Figures 5 and 6 show the precision and F1-scores for the three models derived from each algorithm. These labeled charts clearly show which models perform best for each class, highlighting the strengths of Logistic Regression in handling both frequent and infrequent classes effectively. Naive Bayes and KNN show more variability,

Figure 5

Precission by Class and Model



Note. Source: Developed by the authors

particularly for classes with low support, leading to lower overall performance.

The Logistic Regression model applied to this dataset shows excellent results with the following evaluation metrics:

1. Accuracy: The model achieved an accuracy of 0.92 or 92%, indicating that it correctly predicted 92% of the total test data.
2. Precision: The model's precision is 0.91, meaning that 91% of the predicted positive cases are correct. High precision indicates that the model accurately identifies the correct account names corresponding to the predicted tax object types, thus reducing the possibility of false positives.
3. Recall: The model shows a recall value of 0.90, indicating that 90% of all actual positive cases were correctly identified by the model. This metric shows that the model correctly identified about 90% of the true positive cases (account names that indeed fall into a specific tax object type). High recall demonstrates the model's ability to detect most of the positive cases in the dataset, reducing the possibility of false negatives.
4. F1-score: The F1-score of the model is 0.90, which is the harmonic mean of precision and recall. This value reflects the balance between precision and recall, providing a comprehensive picture of the model's performance. This value indicates that the model has a balanced and good performance in identifying tax objects while minimizing prediction errors, both positive and negative.

The resulting Logistic Regression model demonstrates overall good performance in classifying tax object types from account names in financial reports.

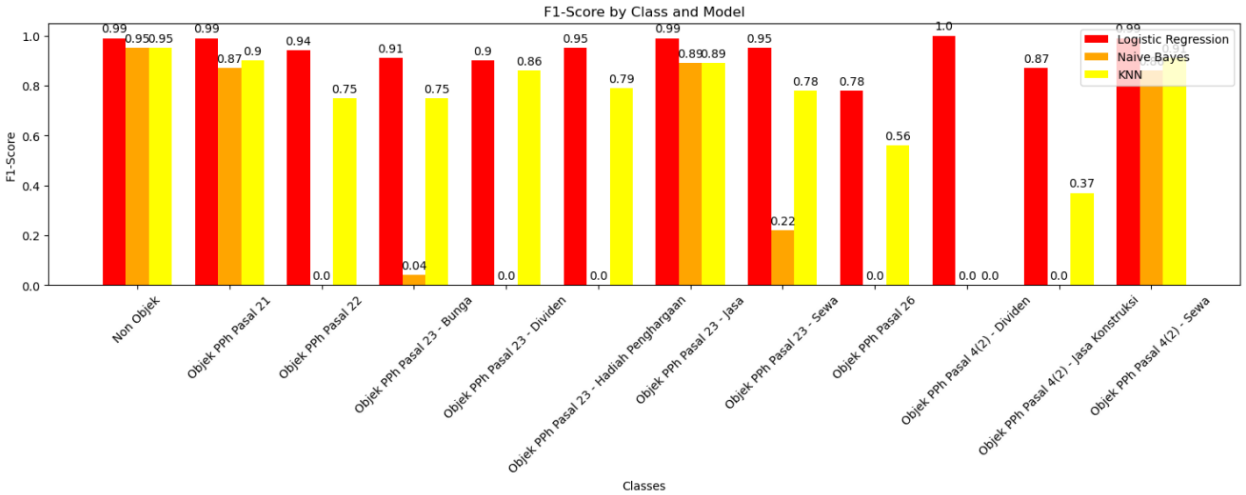
4.2 Naive Bayes and K-Nearest Neighbor (KNN)

Logistic Regression was selected for this study for several reasons. Naive Bayes is a probabilistic classifier based on Bayes' theorem with the assumption of feature independence. Naive Bayes can perform well for certain types of problems, but has the following limitations in this study:

- Lower Accuracy: The accuracy of Naive Bayes is 91.11%, significantly lower than that of Logistic Regression.
- Undefined Metrics: The presence of undefined metrics indicates that the Naive Bayes algorithm struggles with some classes, causing precision and recall to be zero for some classes, such as "*Objek PPh Pasal 22*" and "*Objek PPh Pasal 26.*"
- Performance Variability: Naive Bayes shows greater variability in precision and recall across different classes, making it less reliable for classification problems. Similarly, a model was developed using the KNN algorithm. KNN is a non-parametric method used for classification and regression.

Figure 6

F1-Score by Class and Model



Note. Source: Developed by the authors

- Moderate Accuracy: KNN achieved an accuracy of 92.72%, better than Naive Bayes algorithm struggles with some classes, causing precision and recall to be zero for some classes, such as "*Objek PPh Pasal 22*" and "*Objek PPh Pasal 26*."
- Sensitivity to Data Distribution: KNN's performance is highly dependent on data distribution and the choice of 'k'. This algorithm also tends to require substantial computational resources for relatively large datasets.
- Lower F1-Score for Classes with Fewer Samples: KNN struggles with training and testing classes that have fewer samples, such as "*Objek PPh Pasal 23 - Sewa*" and "*Objek PPh Pasal 26*," resulting in lower F1-scores for these classes.

4.3 Comparative Analysis

Logistic Regression outperforms Naive Bayes and KNN in terms of precision and recall for the three classes included in this study. This finding ensures balanced performance for the Logistic Regression model across various tax objects.

The F1-score, which considers precision and recall, is consistently high for Logistic Regression. This finding highlights its strength as an algorithm for the developed model.

Handling Imbalanced Data: The Logistic Regression model maintains high scores even for classes with fewer samples, such as "*Objek PPh Pasal 4(2) - Jasa Konstruksi*." This finding demonstrates its ability to handle imbalanced data.

4.4 Case Study and Implementation

In this study, we tested the developed model to assess its accuracy in predicting tax objects from real-world general ledger data. Two datasets with different sizes and account naming characteristics were used. The first dataset is 692 MB with a size of 2,835,472 rows x 32 columns. The second

Table 2
Model Accuracy Test in Predicting Tax Object Types

Number	Transactions	Income Tax Article	Classification
1	*beli fresh milk*	Article 22 (PPh Pasal 22)	In accordance with the purchase of goods
2	*sewa Gudang*	Article 4 (2) – Rent (PPh Pasal 4 ayat (2) – Sewa)	In accordance with property rental transactions
3	*sewa bangunan*	Article 4 (2) – Rent (PPh Pasal 4 ayat (2) – Sewa)	In accordance with property rental transactions
4	*sewa kendaraan*	Article 23 – Rent (PPh Pasal 23 – Sewa)	In accordance with vehicle rental transactions
5	*biaya untuk sewa motor*	Article 23 – Rent (PPh Pasal 23 – Sewa)	In accordance with vehicle rental transactions
6	*pembelian fresh milk*	Article 22 (PPh Pasal 22)	In accordance with the purchase of goods
7	*membayar pegawai lembur*	Article 21 (PPh Pasal 21)	In accordance with overtime wage payments
8	*membayar lembur buruh*	Article 21 (PPh Pasal 21)	In accordance with overtime wage payments
9	*bayar upah pegawai*	Article 21 (PPh Pasal 21)	In accordance with wage payments
10	*biaya jasa perbaikan*	Article 23 – Service (PPh Pasal 23 – Jasa)	In accordance with service fees

Note. Source: Developed by the authors

dataset is 78 MB with a size of 491,881 rows x 21 columns.

Using the trained Logistic Regression model, predictions were made on several example transactions shown in Table 2: Model Accuracy Test in Predicting Tax Object Types.

Based on the prediction results above, the Logistic Regression model can identify tax object types from transaction texts quite well. The following are key points from the predictions.

4.4.1 Transactions Related to "Rent"

Predictions for transactions like "*sewa gudang*" and "*sewa bangunan*" are "*Objek PPh Pasal 4(2) - Sewa,*" which is appropriate considering the nature of these transactions.

Transactions categorized as "*sewa kendaraan*", "*sewa kendaraan*", and "*biaya untuk sewa motor*" are predicted as "*Objek PPh Pasal 23 - Sewa,*" which is also appropriate for these transaction types.

4.4.2 Purchase and Payment Transactions

Transactions recorded as "*beli fresh milk*" and "*pembelian fresh milk*" are predicted as "*Objek PPh Pasal 22,*" which reflects the purchase of goods.

Payment transactions such as "*membayar pegawai lembur*," "*membayar lembur buruh*," and "*bayar upah pegawai*" are predicted as "*Objek PPh Pasal 21,*" which reflect wage or salary payments.

4.4.3 Service Fees

Transactions like "*biaya jasa perbaikan*" are predicted as "*Objek PPh Pasal 23 - Jasa,*" which is appropriate for service-related transactions.

The prediction results show that the Logistic Regression model is capable of classifying tax object types from transactions effectively. Additionally, this model can identify tax objects based on the transaction texts presented in the general ledger.

4.4.4 Implications

In this research, the proposed framework has been tested not only to detect the presence of tax objects and identify their types but also to identify potential fiscal corrections, including positive corrections, negative corrections, or no corrections. Given that the Pandas library is used, the input data is of a type that can be read by this library.

To test the success of the classification model related to tax objects and fiscal corrections, we developed a web application. A general ledger

Table 3
Classification Results of Transactions and Estimated Tax Potential

Tax Object	Transaction Value	Estimated Tax Potential
Non Objek	34,881,830,000.00	0.00
Objek PPh Pasal 21	6,649,932,000.00	404,142,900.00
Objek PPh Pasal 23 - Bunga	730,819,000.00	109,622,850.00
Objek PPh Pasal 23 - Jasa	7,377,203,000.00	217,159,320.00
Objek PPh Pasal 23 - Sewa	369,861,000.00	55,479,150.00
Objek PPh Pasal 4(2) - Sewa	2,468,776,000.00	246,877,600.00
Objek PPh Pasal 23 - Royalti	117,579,000.00	17,636,850.00
Objek PPh Pasal 4(2) - Bunga	828,113,000.00	165,622,600.00
Objek PPh Pasal 23 - Hadiah Penghargaan	634,125,000.00	95,118,750.00
Grand Total	54,058,238,000.00	1,311,660,020.00

Note. Source: Developed by the authors

with 100 rows and 7 columns was used as input data. These seven columns include *Nomor_Transaksi, Nama_Akun, Nomor_Akun, Tanggal_Transaksi, Nomor_Dokumen, Deskripsi_Transaksi, Nilai,* as shown in Appendix 1, Dummy Raw General Ledger As Input.

These seven columns were combined into a single *.txt file, which was then input into our web application to be processed. The result included 5 additional columns showing the outcomes of data preparation (column transaction_text_clean), the classification model for "*objek pajak penghasilan*", "*tarif pajak*", "*koreksi fiskal*", and "*potensi pajak*" as shown in Appendix 2 Processed Data Output.

The results of using the model in the application to detect tax objects and potential tax values can be seen in Table 3, Classification Results of Transactions and Estimated Tax Potential.

The results of using the model in the application to detect tax objects and potential tax values can be seen in Table 4 titled Transaction Value by Fiscal Correction Position.

The operation of the proposed web application can be reviewed in Appendix 3, Tax Object and Fiscal Correction Web Application.

Table 4

Transaction Value by Fiscal Correction Position

Position	Transaction Value
Negative Fiscal Correction	828,113,000.00
Positive Fiscal Correction	1,748,710,000.00
No Fiscal Correction	51,481,415,000.00
Grand Total	54,058,238,000.00

Note. Source: Developed by the authors

5. CONCLUSION

Based on the findings, some issues related to the performance of the developed model can be highlighted. This research utilized the logistic regression algorithm to predict income tax objects from accounts in taxpayers' financial statements.

Logistic Regression was chosen as the most suitable algorithm for detecting income tax objects from general ledger data due to its superior accuracy, balanced performance across all classes, and resistance to overfitting. While Naive Bayes and KNN have their respective advantages, they fall short in overall accuracy and consistent performance. These factors make Logistic Regression the best machine-learning model for this application.

As is known, the presentation and disclosure of accounts in financial statements follow accounting standards. There are no rigid rules that constrain financial statement preparers in choosing account names.

This study shows that the logistic regression model can achieve a very satisfactory accuracy level (an average of 90%) in predicting tax objects from account names in the general ledger. Additionally, the results from training and testing the model demonstrate that NLP technology offers various advantages over conventional methods in understanding taxpayers' financial statements, particularly in business processes that require a comprehensive understanding of these financial statements, such as tax audits. By leveraging the developed model, tax objects and fiscal corrections can be automatically detected and classified from general ledger data, enabling the Directorate General of Taxes to carry out audit-related functions more efficiently and effectively.

Implementing this model in a tax audit system will be extremely beneficial for automating the classification of transaction findings based on tax context into appropriate tax objects. This is expected to help tax auditors perform more efficient and accurate tax audits, reduce human error, and potentially increase compliance with tax regulations. With a continually updated and refined model, tax authorities can manage financial data more effectively and make more informed decisions based on the information generated by this model.

However, further research is still needed to explore the possibility of building more advanced models. For instance, using Artificial Neural Network (ANN) algorithms to predict tax objects and fiscal corrections while being more resilient against overfitting.

The findings discussed suggest that it is possible to minimize the time and cost required by the Directorate General of Taxes (DGT) in performing its various functions. Given the current

resources of the DGT, relying solely on human labor without implementing NLP technology cannot optimize the various functions to operate more efficiently and effectively. At least with the application of NLP technology and accurate models, repetitive yet highly subjective tasks can be better managed. This is especially relevant if applied on a larger scale, such as using big data as more robust modeling material before deployment.

Even though our work is still in the developmental stage, the proposed findings can add to decision-making processes and be implemented to assist tax auditors in performing various functions. By using the model built in this research, accompanied by tech-savvy auditors, it is expected that this research can help the DGT detect tax objects and identify potential errors in the tax returns submitted by taxpayers more efficiently and effectively.

From a theoretical perspective, this study contributes to the growing body of research on the application of natural language processing and machine learning in tax administration, particularly in the automated classification of tax objects from textual taxpayers' data.

This research contributes to the growing body of literature on the application of machine learning in public sector decision-making, particularly in tax administration. It provides theoretical support for the use of natural language processing and classification models in transforming unstructured accounting text into structured audit-relevant information. It also extends the discussion on how machine learning can enhance analytical capacity, consistency, and objectivity in complex tax audit processes.

In the context of enhancing the current model and considering future development, it is worth exploring the potential integration of smoothing techniques that trace back to specific identifiers or withholding tax receipt numbers reported by taxpayers. Even though the current model employs fuzzy matching or record linkage based on description, date, or amount, it often encounters challenges due to the absence of counterparty names in ledger descriptions.

This approach would allow for a more precise and verifiable connection between transactions and their corresponding tax records, thereby enhancing the overall effectiveness of the system. Future research and development could focus on implementing and testing such techniques to address these challenges and further streamline the tax reporting and compliance process.

REFERENCES

Alasmri, N., & Basahel, S. (2022). Linking artificial intelligence use to improved decision-making, individual and organizational outcomes. *International Business Research, 15*(10), 1–15. https://doi.org/10.5539/ibr.v15n10p1

Ambi, C. (2022). Text classification using bag of words and TF-IDF with TensorFlow. Python Simplified. https://pythonsimplified.com/text-classification-using-bag-of-words-and-tf-idf-with-tensorflow/

Besley, T., Persson, T., & Sturm, D. M. (2010). Political competition, policy and growth: Theory and evidence from the US. *The Review of Economic Studies, 77*(4), 1329–1352. https://doi.org/10.1111/j.1467-937X.2010.00606.x

Cox, M. S., Neumark, F., & McLure, C. E. (2024). Taxation. In *Encyclopaedia Britannica*. https://www.britannica.com/money/taxation

Darrazão, E., Amorim, V., Oliveira, K., & Gomes-Jr, L. (2023). Engineering and evaluation of features for information extraction in invoices. In *Annals of the XVIII Regional Database School* (pp. 80–89). SBC.

DGT. (2022). *Konsisten pengembangan BDA, DJP ulas penerapan CRMBI via webinar*. https://www.pajak.go.id/id/berita/konsisten-pengembangan-bda-djp-ulas-penerapan-crmbi-webinar

Glantz, S. A., & Slinker, B. K. (2001). *Primer of applied regression and analysis of variance* (2nd ed.). McGraw-Hill.

Harris, J. K. (2021). Binary logistic regression. *Family Medicine and Community Health, 9*(Suppl 1), Article e001290. https://doi.org/10.1136/fmch-2021-001290

Kementerian Keuangan Republik Indonesia. (2024). *Kinerja APBN terjaga di tengah risiko global yang dinamis (APBN KiTa Edisi Mei 2024)*. https://media.kemenkeu.go.id/getmedia/dfa30e00-df85-4ce4-8a9c-0c58c7b1e9f7/APBNKITA-Ed-Mei-2024.pdf?ext=.pdf

Kochanova, A., Hasnain, Z., & Larson, B. (2020). Does e-government improve government capacity? Evidence from tax compliance costs, tax revenue, and public procurement competitiveness. *The World Bank Economic Review*, *34*(1), 101–120. https://doi.org/10.1093/wber/lzy012

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing: An introduction to speech recognition, computational linguistics, and natural language processing*. Prentice Hall.

Kanade, V. (2022). Linear regression vs. logistic regression: Understanding 13 key differences. Spiceworks. https://www.spiceworks.com/tech/artificial-intelligence/articles/linear-regression-vs-logistic-regression/

Lemeshow, S., & Hosmer, D. W. (1982). A review of goodness-of-fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology*, *115*(1), 92–106. https://doi.org/10.1093/oxfordjournals.aje.a113284

Li, H. (2022). Application analysis of AI technology in tax collection and administration in China. In *Proceedings of the 2022 Chinese Control and Decision Conference (CCDC)* (pp. 5843–5846). IEEE. https://doi.org/10.1109/CCDC55256.2022.10033590

Lozano, A. C. G., & Ippolito, A. (2024). *Natural language processing in the detection of fraud in invoices of the municipality of São Paulo (Part 1)*. Inter-American Center of Tax Administrations. https://www.ciat.org/natural-language-processing-in-the-detection-of-fraud-in-invoices-of-the-municipality-of-sao-paulo-part-1/?lang=en

Morar, I. D. (2015). Taxation: Effects and influences. *Procedia Economics and Finance*, *32*, 1622–1627. https://doi.org/10.1016/S2212-5671(15)01488-4

Organisation for Economic Co-operation and Development. (2024). *Revenue statistics in Asia and the Pacific 2024*. https://doi.org/10.1787/e4681bfa-en

Plathottam, S. J., Shrinidhi, K. N., & Hesketh, R. P. (2023). A review of artificial intelligence applications in manufacturing operations. *AIChE Journal*, *69*(5), Article e18065. https://doi.org/10.1002/aic.18065

Pratama, A., & Darono, A. (2022). Exploring Indonesian customs and tax data using a text analytics approach. In *Proceedings of the International Conference on Custom and Tax Cooperation*.

Purwowidhu, C. S. (2023). Perkuat reformasi, capai target pajak. Media Keuangan MK+. https://mediakeuangan.kemenkeu.go.id/article/show/perkuat-reformasi-capai-target-pajak

Santos, M. T. M. (2022). *Classification of products on electronic invoices using unstructured textual descriptions* [Undergraduate thesis, Universidade Federal de Alagoas].

Shakil, M., & Tasnia, M. (2022). Artificial intelligence and tax administration in Asia and the Pacific. In *Artificial intelligence and tax administration in Asia and the Pacific*. Routledge. https://doi.org/10.4324/9781003196020-4

Steedman, M. (1996). Natural language processing. In M. A. Boden (Ed.), *Artificial intelligence* (pp. 229–266). Academic Press. https://doi.org/10.1016/B978-012161964-0/50010-8

Supriadi, I. (2024). The audit revolution: Integrating artificial intelligence in detecting accounting fraud. *Akuntansi dan Teknologi Informasi*, *17*(1), 48–61. https://doi.org/10.24123/jati.v17i1.6279

Veenendaal, A. (2023). The benefits of back-office automation. Blue Prism. https://www.blueprism.com/resources/blog/the-benefits-of-back-office-automation/

Wendler, T., & Gröttrup, S. (2021). *Data mining with SPSS modeler: Theory, exercises, and solutions*. Springer. https://doi.org/10.1007/978-3-030-54338-9

APPENDICES

Appendix A
Dummy Raw General Ledger as Input

Nomor_Transaksi	Nama_Akun	Nomor_Akun	Tanggal_Transaksi	Nomor_Dokumen	Deskripsi_Transaksi	Nilai_Transaksi
TRX0001	Sales	301	2024-06-20	DOC6390	Penjualan sisa stok barang yang hampir kedaluwarsa	535502000.0
TRX0002	Inventory	103	2024-06-26	DOC5426	Penerimaan pendapatan dari kontrak jangka panjang	477143000.0
TRX0003	Inventory	103	2024-07-24	DOC2685	Pembayaran gaji karyawan untuk bulan Juni 2024	838744000.0
TRX0004	Accounts Receivable	102	2024-07-06	DOC6311	Pembayaran tagihan listrik untuk kantor pusat	205872000.0
TRX0005	Sales	301	2024-07-22	DOC9666	Pembelian peralatan medis untuk klinik karyawan	968026000.0
TRX0006	Rent Expense	502	2024-07-28	DOC1189	Penerimaan pendapatan dari penyewaan alat	711242000.0
TRX0007	Accounts Payable	201	2024-07-23	DOC2528	Pembelian barang modal untuk produksi	200307000.0
TRX0008	Salaries Expense	501	2024-06-09	DOC6393	Penerimaan pendapatan dari lisensi produk	736511000.0
TRX0009	Cash	101	2024-07-26	DOC8513	Pengeluaran untuk CSR dan kegiatan sosial	530310000.0
TRX0010	Salaries Expense	501	2024-06-28	DOC1775	Pengeluaran untuk kampanye pemasaran digital	707523000.0
TRX0011	Rent Expense	502	2024-07-18	DOC8555	Pendapatan sewa dari properti komersial	768012000.0
TRX0012	Cost of Goods Sold	401	2024-07-04	DOC7873	Penerimaan pendapatan dari penjualan proyek	88203000.0
TRX0013	Salaries Expense	501	2024-07-14	DOC8629	Penerimaan pembayaran dari kerjasama bisnis	506598000.0
TRX0014	Rent Expense	502	2024-07-21	DOC8916	Pembayaran kontrak outsourcing tenaga kerja	932082000.0
TRX0015	Accounts Payable	201	2024-07-25	DOC7331	Penjualan produk premium dengan harga tinggi	321322000.0
TRX0016	Cash	101	2024-06-13	DOC9006	Penjualan karya seni dan koleksi	594289000.0
TRX0017	Salaries Expense	501	2024-07-24	DOC6258	Pembayaran sewa kendaraan operasional untuk bulan Juli 2024	369861000.0
TRX0018	Inventory	103	2024-06-01	DOC6892	Penjualan sisa stok barang yang hampir kedaluwarsa	454814000.0
TRX0019	Cash	101	2024-07-07	DOC4104	Pendapatan sewa dari properti komersial	549055000.0
TRX0020	Inventory	103	2024-07-05	DOC3731	Pembelian perangkat medis untuk klinik perusahaan	549373000.0
TRX0021	Sales	301	2024-06-03	DOC9110	Pengeluaran untuk CSR dan kegiatan sosial	202529000.0
TRX0022	Salaries Expense	501	2024-07-23	DOC7938	Penerimaan piutang dari pelanggan Z	684887000.0
TRX0023	Sales	301	2024-07-28	DOC8858	Pembayaran sewa bangunan kantor untuk bulan Juni 2024	88780000.0
TRX0024	Cost of Goods Sold	401	2024-07-04	DOC8574	Pembelian peralatan keamanan untuk kantor	139686000.0
TRX0025	Sales	301	2024-06-04	DOC4157	Diskon penjualan diberikan kepada pelanggan setia	3708000.0
TRX0026	Sales	301	2024-07-25	DOC3693	Pembayaran royalti kepada pemegang hak cipta	117579000.0
TRX0027	Rent Expense	502	2024-07-11	DOC6592	Pembelian peralatan medis untuk klinik karyawan	473693000.0
TRX0028	Cost of Goods Sold	401	2024-06-27	DOC7776	Penerimaan pendapatan dari penjualan online	606496000.0
TRX0029	Rent Expense	502	2024-06-06	DOC6530	Pembelian barang impor untuk dijual kembali	794495000.0
TRX0030	Accounts Payable	201	2024-06-10	DOC1663	Pembelian perlengkapan keselamatan kerja	107593000.0
TRX0031	Cash	101	2024-07-24	DOC4763	Pengeluaran untuk pengembangan aplikasi mobile	850877000.0
TRX0032	Cost of Goods Sold	401	2024-07-12	DOC8554	Penerimaan pembayaran dari kerjasama bisnis	746229000.0
TRX0033	Cash	101	2024-06-27	DOC5737	Pembelian perangkat lunak keamanan siber	409110000.0
TRX0034	Cash	101	2024-07-01	DOC6249	Penerimaan pendapatan dari penjualan saham	933005000.0
TRX0035	Accounts Payable	201	2024-07-23	DOC4510	Pembayaran biaya perizinan dan registrasi	990939000.0
TRX0036	Rent Expense	502	2024-06-06	DOC9004	Diskon pembelian dari pemasok karena volume pembelian tinggi	205797000.0
TRX0037	Rent Expense	502	2024-06-16	DOC9800	Pengeluaran untuk kampanye pemasaran digital	379849000.0
TRX0038	Accounts Payable	201	2024-07-03	DOC3811	Penerimaan pendapatan dari penjualan proyek	926523000.0
TRX0039	Accounts Payable	201	2024-06-19	DOC3911	Pembayaran biaya keamanan dan kebersihan	721875000.0
TRX0040	Accounts Receivable	102	2024-07-07	DOC5736	Pembayaran biaya pemasaran dan iklan	49047000.0
TRX0041	Salaries Expense	501	2024-07-23	DOC5061	Diskon pembelian dari pemasok karena volume pembelian tinggi	781733000.0
TRX0042	Rent Expense	502	2024-07-26	DOC3049	Pendapatan bunga dari rekening tabungan perusahaan	828113000.0
TRX0043	Rent Expense	502	2024-06-05	DOC8158	Pembelian hak siar acara televisi	750751000.0
TRX0044	Inventory	103	2024-07-16	DOC5499	Pembayaran biaya pemasaran dan iklan	799738000.0
TRX0045	Inventory	103	2024-07-20	DOC2648	Penjualan properti komersial	825308000.0
TRX0046	Rent Expense	502	2024-07-03	DOC3557	Pembelian inventaris berupa komputer dan peralatan IT	187219000.0
TRX0047	Accounts Receivable	102	2024-07-14	DOC3869	Pembayaran biaya pelatihan eksekutif	236456000.0
TRX0048	Accounts Payable	201	2024-07-01	DOC7944	Diskon penjualan diberikan kepada pelanggan setia	634125000.0
TRX0049	Salaries Expense	501	2024-07-08	DOC2218	Penerimaan pendapatan dari acara amal	907958000.0
TRX0050	Inventory	103	2024-07-27	DOC4446	Pengeluaran untuk kampanye pemasaran digital	316880000.0

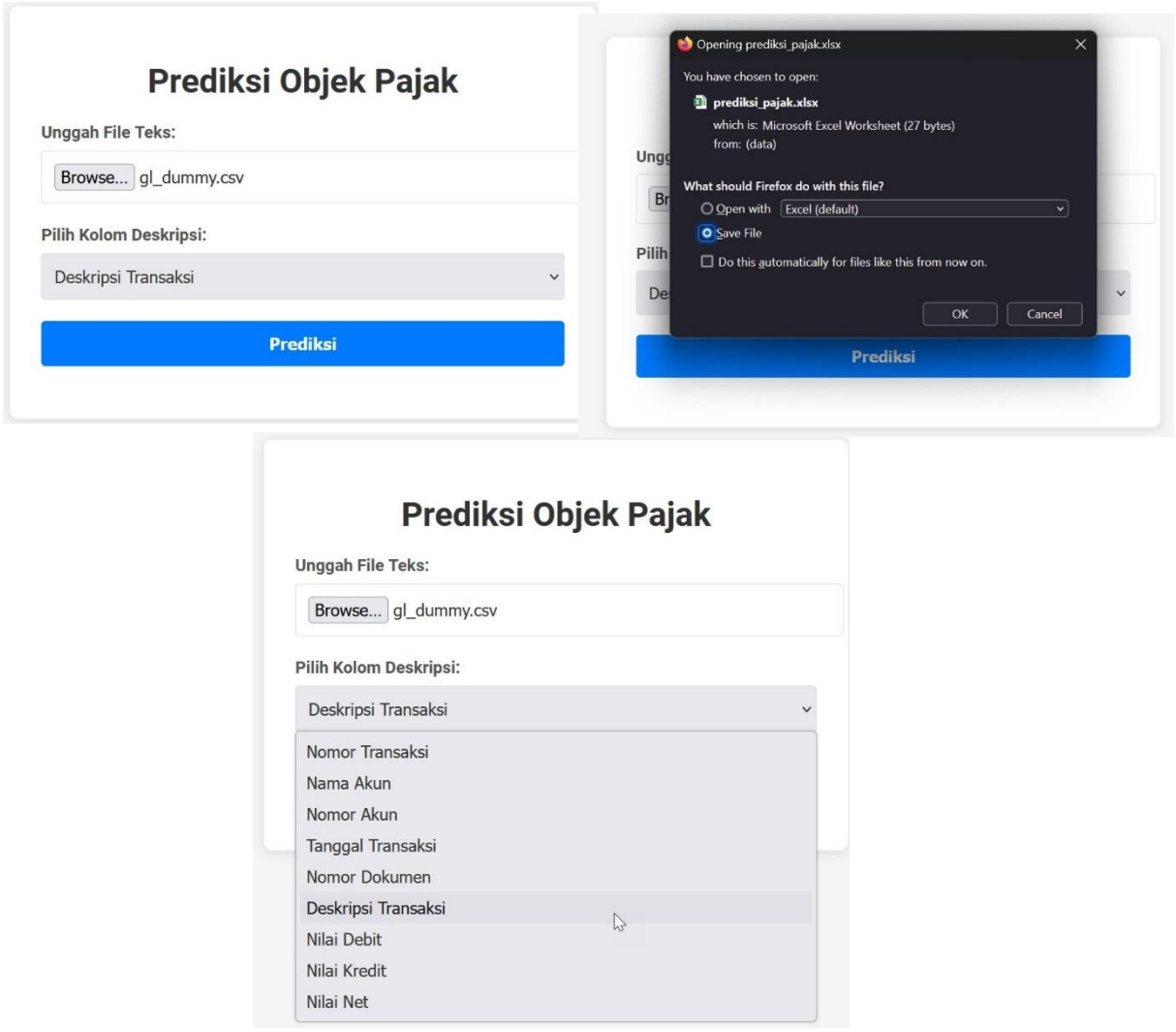
Note. Source: Processed by Author

Appendix B
Processed Data Output

Nomor_Transaksi	Nama_Akun	Nomor_Akun	Tanggal_Transaksi	Nomor_Dokumen	Deskripsi_Transaksi	Nilai	transaction_text_clean	Objek_PPh	Tarif_Minimal	Koreksi_Fiskal	Potensi Pajak
TRX0001	Sales	301	6/20/2024	DOC6390	Penjualan sisa stok barang yang hampir kedaluwarsa	535502000	Penjualan sisa stok barang hampir kedaluwarsa	Objek PPh Pasal 23 - Jasa	15%	Tidak ada koreksi fiskal	80,325,300
TRX0002	Inventory	103	6/26/2024	DOC5426	Penerimaan pendapatan dari kontrak jangka panjang	477143000	Penerimaan pendapatan kontrak jangka panjang	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0003	Inventory	103	7/24/2024	DOC2685	Pembayaran gaji karyawan untuk bulan Juni 2024	838744000	Pembayaran gaji karyawan bulan	Objek PPh Pasal 21	5%	Tidak ada koreksi fiskal	41,937,200
TRX0004	Accounts Receivable	102	7/6/2024	DOC6311	Pembayaran tagihan listrik untuk kantor pusat	205872000	Pembayaran tagihan listrik kantor pusat	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0005	Sales	301	7/22/2024	DOC9666	Pembelian peralatan medis untuk klinik karyawan	968026000	Pembelian peralatan medis klinik karyawan	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0006	Rent Expense	502	7/28/2024	DOC1189	Penerimaan pendapatan dari penyewaan alat	711242000	Penerimaan pendapatan penyewaan alat	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0007	Accounts Payable	201	7/23/2024	DOC2528	Pembelian barang modal untuk produksi	200307000	Pembelian barang modal produksi	Objek PPh Pasal 21	5%	Koreksi Fiskal Positif	10,015,350
TRX0008	Salaries Expense	501	6/9/2024	DOC6393	Penerimaan pendapatan dari lisensi produk	736511000	Penerimaan pendapatan lisensi produk	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0009	Cash	101	7/26/2024	DOC8513	Pengeluaran untuk CSR dan kegiatan sosial	530310000	Pengeluaran CSR kegiatan sosial	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0010	Salaries Expense	501	6/28/2024	DOC1775	Pengeluaran untuk kampanye pemasaran digital	707523000	Pengeluaran kampanye pemasaran digital	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0011	Rent Expense	502	7/18/2024	DOC8555	Pendapatan sewa dari properti komersial	768012000	Pendapatan sewa properti komersial	Objek PPh Pasal 4(2) - Sewa	10%	Tidak ada koreksi fiskal	76,801,200
TRX0012	Cost of Goods Sold	401	7/4/2024	DOC7873	Penerimaan pendapatan dari penjualan proyek	88203000	Penerimaan pendapatan penjualan proyek	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0013	Salaries Expense	501	7/14/2024	DOC8629	Penerimaan pembayaran dari kerjasama bisnis	506598000	Penerimaan pembayaran kerjasama bisnis	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0014	Rent Expense	502	7/21/2024	DOC8916	Pembayaran kontrak outsourcing tenaga kerja	932082000	Pembayaran kontrak outsourcing tenaga kerja	Objek PPh Pasal 23 - Jasa	2%	Tidak ada koreksi fiskal	18,641,640
TRX0015	Accounts Payable	201	7/25/2024	DOC7331	Penjualan produk premium dengan harga tinggi	321322000	Penjualan produk premium harga tinggi	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0016	Cash	101	6/13/2024	DOC9006	Penjualan karya seni dan koleksi	594289000	Penjualan karya seni koleksi	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0017	Salaries Expense	501	7/24/2024	DOC6258	Pembayaran sewa kendaraan operasional untuk bulan Juli 2024	369861000	Pembayaran sewa kendaraan operasional bulan	Objek PPh Pasal 23 - Sewa	15%	Tidak ada koreksi fiskal	55,479,150
TRX0018	Inventory	103	6/1/2024	DOC6892	Penjualan sisa stok barang yang hampir kedaluwarsa	454814000	Penjualan sisa stok barang hampir kedaluwarsa	Objek PPh Pasal 23 - Jasa	2%	Tidak ada koreksi fiskal	9,096,280
TRX0019	Cash	101	7/7/2024	DOC4104	Pendapatan sewa dari properti komersial	549055000	Pendapatan sewa properti komersial	Objek PPh Pasal 4(2) - Sewa	10%	Tidak ada koreksi fiskal	54,905,500
TRX0020	Inventory	103	7/5/2024	DOC3731	Pembelian perangkat medis untuk klinik perusahaan	549373000	Pembelian perangkat medis klinik perusahaan	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0021	Sales	301	6/3/2024	DOC9110	Pengeluaran untuk CSR dan kegiatan sosial	202529000	Pengeluaran CSR kegiatan sosial	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0022	Salaries Expense	501	7/23/2024	DOC7938	Penerimaan piutang dari pelanggan Z	684887000	Penerimaan pelanggan	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0023	Sales	301	7/28/2024	DOC8858	Pembayaran sewa bangunan kantor untuk bulan Juni 2024	88780000	Pembayaran sewa bangunan kantor bulan	Objek PPh Pasal 4(2) - Sewa	10%	Tidak ada koreksi fiskal	8,878,000
TRX0024	Cost of Goods Sold	401	7/4/2024	DOC8574	Pembelian peralatan keamanan untuk kantor	139686000	Pembelian peralatan keamanan kantor	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0025	Sales	301	6/4/2024	DOC4157	Diskon penjualan diberikan kepada pelanggan setia	3708000	Diskon penjualan diberikan pelanggan setia	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0026	Sales	301	7/25/2024	DOC3693	Pembayaran royalti kepada pemegang hak cipta	117579000	Pembayaran royalti pemegang hak cipta	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0027	Rent Expense	502	7/11/2024	DOC6592	Pembelian peralatan medis untuk klinik karyawan	473693000	Pembelian peralatan medis klinik karyawan	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0028	Cost of Goods Sold	401	6/27/2024	DOC7776	Penerimaan pendapatan dari penjualan online	606496000	Penerimaan pendapatan penjualan online	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0029	Rent Expense	502	6/6/2024	DOC6530	Pembelian barang impor untuk dijual kembali	794495000	Pembelian barang impor dijual	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0030	Accounts Payable	201	6/10/2024	DOC1663	Pembelian perlengkapan keselamatan kerja	107593000	Pembelian perlengkapan keselamatan kerja	Objek PPh Pasal 23 - Jasa	2%	Tidak ada koreksi fiskal	2,151,860
TRX0031	Cash	101	7/24/2024	DOC4763	Pengeluaran untuk pengembangan aplikasi mobile	850877000	Pengeluaran pengembangan aplikasi mobile	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0032	Cost of Goods Sold	401	7/12/2024	DOC8554	Penerimaan pembayaran dari kerjasama bisnis	746229000	Penerimaan pembayaran kerjasama bisnis	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0033	Cash	101	6/27/2024	DOC5737	Pembelian perangkat lunak keamanan siber	409110000	Pembelian perangkat lunak keamanan siber	Objek PPh Pasal 23 - Jasa	2%	Tidak ada koreksi fiskal	8,182,200
TRX0034	Cash	101	7/1/2024	DOC6249	Penerimaan pendapatan dari penjualan saham	933005000	Penerimaan pendapatan penjualan saham	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0035	Accounts Payable	201	7/23/2024	DOC4510	Pembayaran biaya perizinan dan registrasi	990939000	Pembayaran perizinan registrasi	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0036	Rent Expense	502	6/6/2024	DOC9004	Diskon pembelian dari pemasok karena volume pembelian tinggi	205797000	Diskon pembelian pemasok volume pembelian tinggi	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0037	Rent Expense	502	6/16/2024	DOC9800	Pengeluaran untuk kampanye pemasaran digital	379849000	Pengeluaran kampanye pemasaran digital	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0038	Accounts Payable	201	7/3/2024	DOC3811	Penerimaan pendapatan dari penjualan proyek	926523000	Penerimaan pendapatan penjualan proyek	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0039	Accounts Payable	201	6/19/2024	DOC3911	Pembayaran biaya keamanan dan kebersihan	721875000	Pembayaran keamanan kebersihan	Objek PPh Pasal 23 - Jasa	2%	Tidak ada koreksi fiskal	14,437,500
TRX0040	Accounts Receivable	102	7/7/2024	DOC5736	Pembayaran biaya pemasaran dan iklan	49047000	Pembayaran pemasaran iklan	Objek PPh Pasal 23 - Jasa	2%	Tidak ada koreksi fiskal	980,940
TRX0041	Salaries Expense	501	7/23/2024	DOC5061	Diskon pembelian dari pemasok karena volume pembelian tinggi	781733000	Diskon pembelian dari pemasok volume pembelian tinggi	Non Objek	0%	Tidak ada koreksi fiskal	-
TRX0042	Rent Expense	502	7/26/2024	DOC3049	Pendapatan bunga dari rekening tabungan perusahaan	828113000	Pendapatan bunga rekening tabungan perusahaan	Objek PPh Pasal 23 - Bunga	15%	Koreksi Fiskal Negatif	124,216,950

Note. Source: Processed by Author

Appendix C
Tax Object and Fiscal Correction Web Application



Note. Source: Processed by Author