Data Science : Journal Of Computing And Applied Informatics
Journal homepage: https://jocai.usu.ac.id

# Precision Document Transaction Type Classifier Using Machine Learning Techniques

Jay Carlou C. Sabado*[1], Sheena I. Sapuay-Guillen[2]
[1] *City Government of San Fernando, La Union, City of San Fernando, La Union, 2500, Philippines*
[2] *Don Mariano Marcos Memorial State University - MLUC, City of San Fernando, La Union, 2500, Philippines*
*Corresponding Author: jsabado0012@student.dmmmsu.edu.ph

**ARTICLE INFO**

**Article history:**
Received 10 November 2024
Revised 07 Oktober 2024
Accepted 20 Desember 2024
Available online 31 January 2025

E-ISSN: 2580-829X
P-ISSN: 2580-6769

**How to cite:**
Jay Carlou C. Sabado and Dr. Sheena I. Sapuay-Guillen, "Precision Document Transaction Type Classifier Using Machine Learning Techniques," Journal of Computing and Applied Informatics, vol. V9, no.1, Jan. 2025, doi: 10.32734/jocai.v9.i1-19945

http://doi.org/10.32734/jocai.v9.i1-19945

## ABSTRACT

This paper aimed to develop a Precision Document Transaction Type Classifier using machine learning to identify transaction types, aligning with the Ease of Doing Business Law (RA 11032), which aims to streamline government services and improve service delivery. With the use of existing government documents, a dataset was created and processed for the training and evaluation of models, including Naïve Bayes, Bidirectional Long Short-Term Memory (Bi-LSTM), and Bidirectional Encoder Representations from Transformer (BERT). The BERT Model was the most accurate, efficient, and precise among other models. For the development of the software application Agile Methodology was used to ensure iterative progress and adaptability during the development phase. For the software quality evaluation, it was assessed using ISO/IEC 25010:2011, achieving a general high score mean of 4.25 corresponding to a descriptive equivalent of Excellent covering various software quality metrics demonstrating reliability, efficiency and overall performance.

**Keyword:** Precision Document Transaction Type Classifier, Agile Methodology, Software Quality, Machine Learning, Ease of Doing Business Law, Republic Act 11032

## ABSTRAK

Penelitian ini bertujuan untuk mengembangkan Pengklasifikasi Jenis Transaksi Dokumen Presisi menggunakan pembelajaran mesin untuk mengidentifikasi jenis transaksi, sejalan dengan Undang-Undang Kemudahan Berusaha (RA 11032), yang bertujuan untuk merampingkan layanan pemerintah dan meningkatkan pemberian layanan. Dengan menggunakan dokumen pemerintah yang ada, sebuah dataset dibuat dan diproses untuk pelatihan dan evaluasi model, termasuk Naïve Bayes, Bidirectional Long Short-Term Memory (Bi-LSTM), dan Bidirectional Encoder Representations from Transformer (BERT). Model BERT adalah yang paling akurat, efisien, dan tepat di antara model-model lainnya. Untuk pengembangan aplikasi perangkat lunak, Metodologi Agile digunakan untuk memastikan kemajuan berulang dan kemampuan beradaptasi selama fase pengembangan. Untuk evaluasi kualitas perangkat lunak, evaluasi dilakukan dengan menggunakan ISO/IEC 25010: 2011, mencapai rata-rata skor tinggi secara umum sebesar 4,25 yang sesuai dengan deskriptif yang setara dengan Sangat Baik yang mencakup berbagai metrik kualitas perangkat lunak yang menunjukkan keandalan, efisiensi, dan kinerja secara keseluruhan.

**Keyword:** Rprime RSA, Extended Tiny Encryption Algorithm, Kriptografi, Kriptografi Hibrida, Pesan Cepat

## 1. Introduction

Electronic governance or e-governance was defined as a modern approach to delivering government services to the people through ICT technologies. E-governance has been observed to significantly improve the efficiency, transparency, and accountability of government services [1]. E-governance systems were widely adopted in the middle of the 2000s due to their benefits and advancements over the traditional means of service delivery [2].

The Government of Estonia is a pioneer in implementing e-governance programs; it started in 1996 when e-banking was introduced, encouraging people to shift to online banking. Year after year, new e-governance programs, such as the E-Tax Board, were introduced, where Estonians can settle their taxes online within three (3) minutes. E-ID and Digital Signatures are the primary identification systems for Estonian residents. I-Voting is the first electronic election platform where people can cast their votes securely over the Internet. E-Health is an integrated platform for healthcare services [3]. Singapore uses e-governance programs to track the processing of land title registration applications throughout the entire process, from the initial submission of the application to the issuance of the new land title. This has helped to improve the transparency, efficiency, and accuracy of land title registration in Singapore [4].

More governments and entities worldwide use e-governance to manage their affairs and respective services. According to [5], government institutions focus on creating a livable environment for its people. Using the Naive Bayes algorithm for data analysis and other technologies, agencies will have a complete understanding of factors affecting public health in the environment; applying technologies to the systems being used and managed by the government leads to significant improvements in the evaluation and architectural design of public places. Moreover, according to [6], to help improve traffic safety, they used Bidirectional Encoder Representations forms Transformer (BERT) in their study to classify injury types. By using the tool, safety engineers, analysts, and other experts can uncover more factors contributing to traffic incidents and reveal more opportunities for safety countermeasures. In another study by [7], using Bi-directional Long Short Term Memory (Bi-LSTM), an e-governance initiative was created that categorizes public messages into topics, to better understand public opinions, suggestions, and feedback, helping the government to become more direct, efficient and cost-effective in providing services.

According to [8], governments are pressured to provide new public services using advanced technologies to modernize government practice in service delivery. One example is the license plate recognition system, which uses image and text recognition for toll payment collection, parking management, and illegal activity detection, making the government more efficient in its governance.

In the Philippines, the government is fully committed to achieving the United Nations Sustainable Development Goals (SDGs). SDG is a holistic and inclusive approach to addressing global challenges covering all sectors of society. SDG 9 focuses on Industry, Innovation, and Infrastructure. In line with this goal, the House of Representatives passed a bill called the E-governance Act for faster services to the public establish government services through the use of digital platforms for easier access for the people [9] reducing inequalities in accordance to the SDG 10. This demonstrates that the Philippine government as a whole prioritizes leveraging technology and innovation to improve the lives of the people.

The City Government of San Fernando, La Union, is mandated to provide different services to its citizens. These services include processing permits, memorandums, certifications, orders, and other government-related documents. One of the e-governance programs currently being used by the City Government is the Document Tracking System. The system can track and monitor the progress of every transaction in the city and measure the performance of every office and personnel processing the transactions. The system was designed to comply with the Republic Act No. 11032, commonly known as the Ease of Doing Business and Efficient Government Service Delivery Act of 2018. It was mandated that all government institutions streamline the practice, processes, and services of their respective agencies, making it more effective and efficient for service delivery to the public, promoting transparency, accountability, trust, and the prevention of corrupt practices in the government service. (RA). Under sections 4 and 9 of the Republic Act, there are three (3) transaction types mentioned. First, the Simple transactions require ministerial action that can be handled by an employee within three (3) working days. Second, complex transactions involve an evaluation of an officer before taking action, and they must be completed within seven (7) working days. Lastly, highly technical applications demand technical or specialized skills to ensure the accurate processing of a transaction and that it is completed within 20 working days.

The City Government is using the Document Tracking System, which was designed to comply with the said processing standards by the law. Still, despite its strength in tracking the status of transactions and notification features for the agency to comply with the said law, some features need to be improved. One of which is the manual identification of transaction type. When a client goes to the Records Office, the person in charge will categorize each document by analyzing its content through a detailed reading process. Documents containing keywords such as request, invitation, or ordinary communication were tagged as simple transactions. Documents associated with more intricate processes, such as project proposals, financial assistance requests, or programs of work, were labeled as complex. Furthermore, documents like ordinances, resolutions, and other similarly structured materials were designated as highly technical due to the technicalities of the content. However, it was observed that this transaction type classification process could be confusing, a document can initially appear as a simple communication that could contain complex or highly technical information upon further inspection and vice versa. In identifying transaction types, there are challenges for the employees handling the system to correctly identify the proper transaction type based on the set standards. The following are the following: First, there is an absence of trained personnel to classify documents correctly. According to the records officer, it will take 10 minutes or longer for the untrained personnel and three (3) to five (5) minutes for trained personnel to accurately identify the proper transaction type. Second, even if there is a trained or seasoned employee around, there is a chance that some of the documents can be incorrectly classified due to the simplicity and complexity of the content. Third, offices in the City Government have their own interpretation of classifying documents. Lastly, due to the voluminous incoming documents, with an average of 218 documents per day based on the November 2023 report, the officer in charge doesn't have enough time to read the whole document for proper classification. As a result, simple transactions can have a lengthy processing time, and complex and highly technical transactions will not have enough processing time. This will lead to possible complaints, dissatisfaction, and distrust of clients towards the government employees handling their respective transactions. According to [10], failure to process within the prescribed processing time will have heavy penalties, six (6) months of suspension for the first offense and imprisonment for up to six (6) years, disqualification of government service with up to two million pesos of penalty. With those challenges, the researcher was prompted to develop a Document Transaction Type Classifier (DTTC). This application uses machine learning techniques to identify transaction types accurately. With this document, transactions in the government can be more compliant with the Ease of Doing Business Law, guiding government workers by reducing unnecessary delays and errors. Client transacting their business with the government will have quicker turn-around time. Given the standard in place as a guide for government transactions, the developed application can be replicated in all government agencies that cater to front-line services.

The primary objective of this study is to develop a Precision Document Transaction Classifier by applying machine learning techniques to accurately identify document transaction types in alignment with the provision of the Ease of Doing Business Law. This study aims to accomplish the following objectives:

1. To build the dataset for classifying City Government documents
2. To train the document classifier using machine-learning techniques
   - Naïve Bayes
   - Bi-LSTM
   - BERT
3. To evaluate the best algorithm for document classification
4. To develop an application for classifying document transaction types
5. To Evaluate the software quality using ISO/IEC 25010:2011

## 2. Method

This study used the Design Science Research Method (DSRM). According to [11], the objective of Design Science Research is the implementation of innovative solutions to address real-world problems and scenarios using problem-solving. In this methodology, information technology provides tools, platforms, and frameworks for developing solutions, contributing to future technological advancements. In the study of Guntara et al. (2023), DSRM was effective for developing information systems due to a problem-solving approach involving problem identification, designing, and creating the best application as a solution.

Six (6) phases of DSRM need to be considered for the project to attain its goals. (1) Problem Identification and Motivations. According to [12], this phase involves awareness of the problems, gaps, and challenges that require specific solutions. Conversely, motivation is the desire to address the problems identified, seeing the future impact, results, and relevance in a real-world setting. (2) Objectives of the Solution. In this phase, goals and objectives will be clearly stated; the researcher needs to specify the criteria for the evaluation of the objectives, which can be qualitative or quantitative, depending on the stated problems. (3) Design and Development. [13] mentioned that by using the collected information and the analysis of data from the previous phase, this crucial information must be translated into the design and development of the solution with the use of appropriate technologies. (4) Demonstration. In this phase, the researcher will conduct a presentation of the designed solution. It was mentioned by [14] that a developed solution must be evaluated by its utility, fitness, and usefulness while considering the given evaluations. (5) Evaluation. In this phase, the designed solution's usability and appropriateness to the objectives will be assessed. This is where the researcher will gather user feedback and modify the solution based on the evaluation results. (6) Communication of Results. In this phase, the researcher will briefly present the overall results of the research, which includes the discussions and decisions made in the design of solutions and the evaluation outcomes.

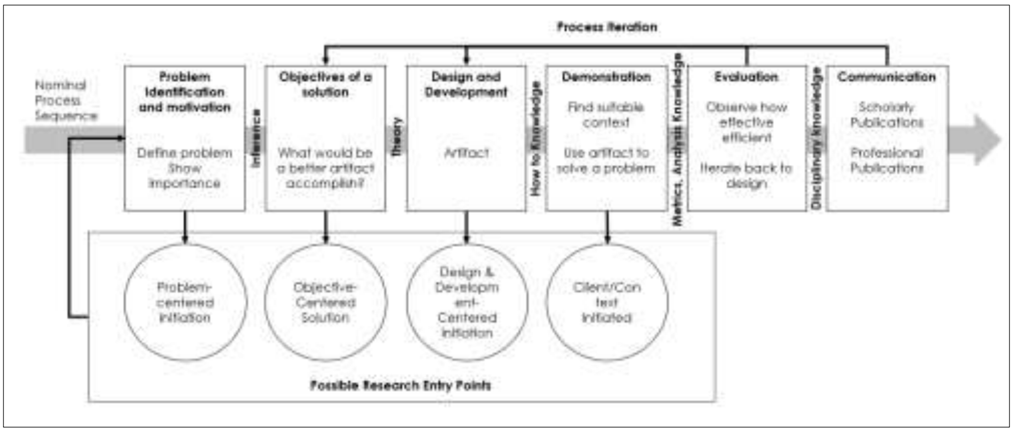

Process Iteration
Nominal Process Sequence
Problem Identification and motivation
Define problem Show importance
Inference
Objectives of a solution
What would be a better artifact accomplish?
Theory
Design and Development
Artifact
How to Knowledge
Demonstration
Find suitable context
Use artifact to solve a problem
Metrics, Analysis Knowledge
Evaluation
Observe how effective efficient
Iterate back to design
Disciplinary Knowledge
Communication
Scholarly Publications
Professional Publications
Problem-centered initiation
Objective-Centered Solution
Design & Development-Centered initiation
Client/Context initiated
Possible Research Entry Points

Figure 1: DSRM

All phases of Design Science Research Methodology (DSRM) were used in the design and development of the Document Transaction Type Classifier (DTTC), ensuring a structured, iterative approach that addressed the need and inputs of the stakeholders for the accurate transaction type classification of documents, compliant with the Ease of Doing Business Law. The process began with Problem identification and motivation, where the significant challenges faced by the City Government's Records Office, such as inconsistent transaction type classification, time-consuming processes to analyze the correct transaction type, and high-volume daily incoming documents for processing that need to be classified correctly. Incorrect classification of document transactions can lead to client dissatisfaction, leading to penalties established in the law, which also serves as an additional motivation for the study. The definition of objectives phase followed, establishing clear goals that aimed to set direction and targets to complete the study, such as building the dataset for model training, training of classifier models, evaluation of the best classification model, design and development of document type classifier application and the assessment of the application using ISO standards. The design and development phase involved creating the machine learning-powered software application, where inputs from the stakeholders were translated into actionable features and functionalities of a software application. The demonstration phase will test the software application in a controlled environment with the stakeholders, highlighting the ability to identify transaction types in accordance with the law correctly and, thereby, supporting the stakeholders in improving and standardizing their service to the public. In the evaluation phase, the software application was assessed using ISO/IEC 25010:2011, which provides quality metrics for software quality evaluation. Finally, the communication phase is where the developed software application is presented and demonstrated to the stakeholders, highlighting how it effectively addresses the identified challenges.

### *2.1 Materials and Procedures*

For the first objective, the researcher used Data Understanding, Data Preparation, and Modeling of the CRoss Industry Standard Process for Data Mining (CRISP-DM) to build a data set for text classification. The CRISP-DM is a 4th Industrial Revolution methodology effective for developing predictive models [15]. According to [16]. This model is popular among data science teams executing projects due to the straightforward and easy-to-understand workflow.

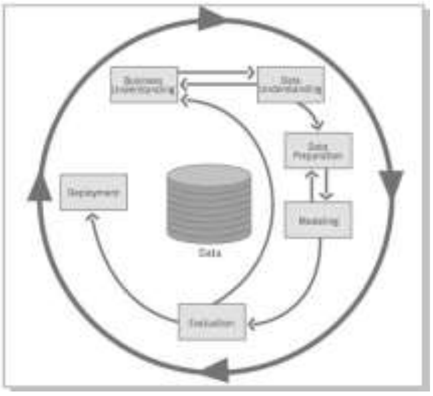


Figure 2: CRISP-DM

For the second objective, the researcher implemented the data mining phase, which involves selecting and applying modeling techniques. There were four (4) tasks involved in completing this phase: (1) selecting the modeling technique, (2) generating the test design, (3) building the model, and (4) assessing the model.

### *2.2 Training Model*

The researcher identified three (3) models for text classification. Naïve Bayes, Bidirectional Long Short-Term Memory (Bi-LSTM), and Bidirectional Encoder Representations form Transformer (BERT) as text classifiers for the study. According to [17], Naïve Bayes is a supervised learning technique that uses probability to make predictions and is used for categorizing textual data. According to [18], using Bi-LSTM has a promising result for automating abstract text screening classification due to its forward and backward processing. According to [19], BERT can provide cutting-edge performance in natural language processing tasks, specifically in understanding contextual nuances, outperforming traditional models.

The second task was to generate a test design. This involves splitting the available dataset, 80% for training and 20% for the test sets, and this will gauge the quality and accuracy of the used model. The third task was to build the model. This task requires running the prepared dataset using data mining tools such as Scikit-Learn, TensorFlow, Keras, PyTorch, and Transformers. Using these data mining tools, the researcher fine-tuned the parameters to improve the model's performance. Lastly, evaluation of the model involves training validations and back-testing to ensure that the created model will meet the desired outcome and objectives of the study.

For the third objective, the researcher would determine the best algorithm for text classification by comprehensively assessing the trained model's overall performance by evaluating key metrics such as accuracy, precision, recall, and F1 Score to ensure a thorough selection of the trained model. To accomplish this objective, the researcher implemented the Evaluation phase of CRISP-DM. This phase involves assessing the quality and performance of the predictive model that is being developed in terms of accuracy, precision, recall, and f1 score. According to [20], the performance of classifier models can be evaluated using the confusion matrix. Confusion Matrix consists of four outcomes. (1) True Positive are instances that are correctly predicted; (2) False Positive instances that are incorrectly predicted as positive; (3) True Negative instances that are correctly predicted as negative; (4) False Negative instances that are incorrectly predicted as negative. F1 Score, Precision, and Recall are essential metrics in evaluating the created model's performance.

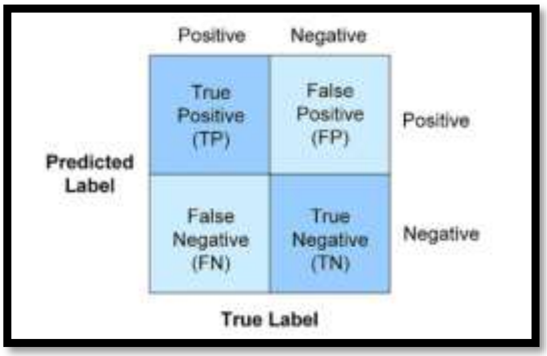


Figure 3: Confusion Matrix

For the fourth objective, the researcher developed an application capable of classifying documents using the extracted textual information in the document. It was mentioned by [21] that Agile software development is a goal-oriented method that considers user experiences that satisfy the client's demands.

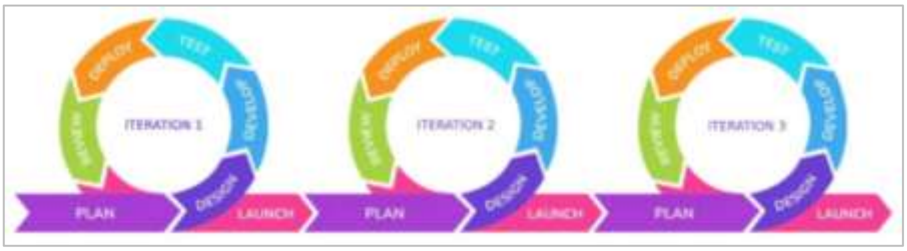


Figure 4: Agile Model

The development model also relies on an iterative development approach and constant delivery of software outputs until the customer objectives are met [22].

For the last objective, the developed software was evaluated using ISO/IEC 25010:2011. According to [23], the ISO standard provides a comprehensive framework to evaluate software's general quality characteristics, including suitability, performance, compatibility, usability, reliability, security, maintainability, and portability. These criteria ensure an overall quality assessment of the software product.

### *2.3 Data Gathered*

For the first objective of this study, the researcher focused on gathering textual information from a total of 3,000 scanned documents consisting of communications, memorandums, resolutions, the program of works, and other government documents, with 1,000 documents representing each transaction type within the City Government, it involves extracting text from scanned documents, using Optical Character Recognition (OCR), once important content was extracted, appropriate transaction type was be assigned for each record. The collected data was processed using data preparation under the CRISP-DM methodology before the training and creation of the text classification model.

For the fifth objective, the researcher employed a total enumeration to collect responses from all individuals within the Records Section. A questionnaire based on the ISO/IEC 25010:2011 standard was used to assess the software quality of the developed application. The questionnaire is provided in Appendix A in various areas of software quality, such as functionality, performance efficiency, compatibility, usability, reliability, security, maintainability, and portability. Through this method, the researcher will assess the software quality from the users' perspective. The data to be collected through the questionnaire was quantitative data derived from scaled responses, allowing for the measurement of different quality attributes mentioned.

### *2.4 Data Analysis*

For the third objective, the researcher compared the trained models regarding accuracy, precision, recall, and f1 score.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Accuracy will calculate the overall accuracy of a prediction model. True Positive (TP) and True Negative (TN) prediction divided by the sum of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Getting a high calculation indicates that the prediction model is correct.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Precision True Positive (TP) divided by the sum of True Positive (TP) and False Positive (FP). The calculation of this formula will show how well the model performs in producing positive and accurate predictions.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Recall is also referred to as True Positive Rate (TPR). True Positive (TP) divided by the sum of True Positive (TP) and False Negative (FN). This will evaluate the created model's ability to identify and remember instances in a dataset.

$$\text{F1} = \frac{\text{2 x Precision x Recall}}{\text{Precision} + \text{Recall}}$$

The F1 score is a performance metric that provides a balanced assessment of the model’s accuracy. It is computed by taking twice the product of precision and the result of recall divided by the sum of precision and recall.

In evaluating the accuracy, precision, recall, and f1 score of the trained models, it is crucial to consider the impact of overfitting. Overfitting occurs when a model is excessively trained on training data, capturing noise and irrelevant patterns that lead to poor generalization of new data. It is essential to address and mitigate overfitting to ensure that the performance metrics of the trained models accurately reflect the efficacy across diverse datasets and real-world scenarios. The researcher conducted cross-validation; according to [24] ,cross-validation will help assess and enhance the machine learning model's performance. By incorporating cross-validation, it will minimize the risk of overfitting or biases in the trained models.

For the last objective to determine the software quality of the developed application, the Likert scale will be applied to the ISO/IEC 25010:2011 questionnaire. According to [25], the Likert scale is frequently utilized in questionnaire-based research to assess survey answers. This approach will ensure a systematic and standardized assessment and facilitate a comprehensive analysis of key attributes such as suitability, performance, compatibility, usability, reliability, security, maintainability, and portability. Using the Likert scale enables respondents to express their degree of agreement or disagreement on a set of standards. The frequency count and the mean will be used for the data treatment. The frequency count shows the number of times a particular value occurs in a data set. The mean quantitatively measures central tendency in statistics.

| Numerical Equivalent | Statistical Range | Descriptive Equivalent |
|---|---|---|
| 1 | 1.00 - 1.79 | Poor |
| 2 | 1.80 - 2.59 | Fair |
| 3 | 2.60 - 3.39 | Good |
| 4 | 3.40 - 4.19 | Very Good |
| 5 | 4.20 - 5.00 | Excellent |

## 3. Results and Discussions

### *3.1 Building Datasets*

In gathering the necessary documents, the researcher requested that the City Government of San Fernando conduct research. In the data-gathering process, the Office of the City Administrator has assigned the Records Section to choose 1,000 scanned documents for each transaction type. A member of the City Committee Anti-Red Tape carefully examined these documents to guarantee that all documents were suitable for data gathering and that all documents were classified correctly according to the compliance of RA11032. To avoid any issues, the researcher and the agency agreed to carefully choose the text that will be included for the processing and training of the models.

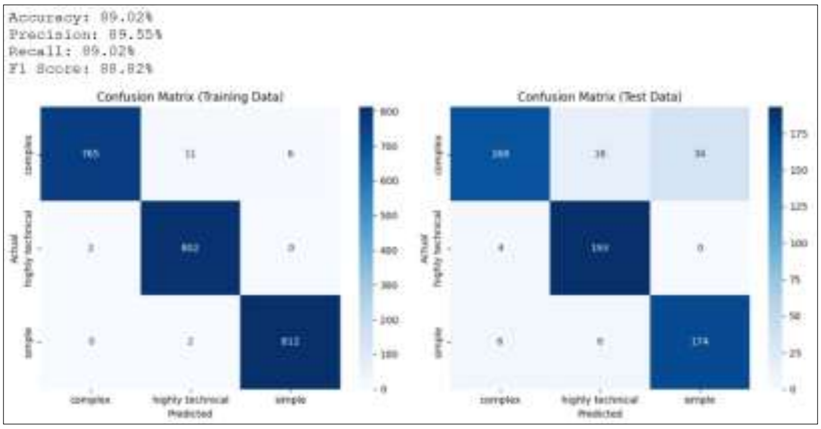
For the first objective, Data Understanding of the CRISP-DM was employed collecting relevant data from various sources, conducting examination, characterization, and verifying the quality is crucial to this stage. [26]. The objective of this phase of the study is for the researcher to gather and understand the content of documents being transacted by using optical character recognition to extract textual information and then classify each document according to their respective transaction types based on the legal requirements. In this process, the tagging of documents into the correct transaction types is essential to ensure successful data processing and model training, and the tagging of documents to their proper transaction type must be compliant with RA 11032. The selection and tagging process involved several steps. The Records Section conducted a preliminary assessment to identify 1,000 documents for each transaction type. Following this step, a City Committee Anti-Red Tape member carefully examined the documents to guarantee that documents that would be released to the researcher were suitable for data gathering and that all documents were classified correctly. For the documents tagged as Simple Transactions, the criteria must require minimal processing time and steps such as requests, standard communication, certificates, and clearances. Documents under Complex Transactions involve a verification process and multiple approvals and require additional documents. Documents with Highly Technical Transactions will demand in-depth assessment, reviews, solutions, and significant inter-departmental collaboration. An essential consideration throughout the data gathering process is data privacy; given that some of the documents contain sensitive information, some data in the text will not be included; only the necessary and permissible text was retained for the data gathering, processing, and model training.

For Data Preparation under CRISP-DM, the researcher prepared the collected data, which will be used as a data set for analysis. Data cleansing is important in the data preparation stage to achieve a high accuracy percentage in testing the data set. The following preparation methods are being performed. The researcher used Optical Character Recognition (OCR) to extract textual information from the scanned documents. All textual information gathered was placed into a CSV file for the data preparation stage.

With Python as the programming language, Pandas library is used for data manipulation and analysis, and Natural Language Toolkit is used for the preparation. The data must undergo different stages of processing. First, Lowercasing is a method used to standardize text words and follow a specific format for a more accurate data reduction process. Converting all text to lowercase will treat all words the same, reducing vocabulary size and enhancing uniformity and efficiency for text processing tasks. Second, removal of stop-words. This process automatically removes words with no significance or value [27]; this process will help improve the efficiency and accuracy of training a model. Stop words in the text that carry little information should be removed for analytical purposes. Eliminating these words, according to the study by [28], will increase the accuracy by 3% and 31% of the Mean Average Precision for the text classification model. Third removal of Punctuation Marks, Extra Spaces, Numbers, and Special Characters, Removing punctuation marks is a text processing technique that will make all words in the text equal and simplified to make it tokenized in preparation for the data reduction process. Eliminating these will help the model training to focus more on core words that carry meaning, reducing data that often do not contribute to text classification because it produces noise and inconsistencies, potentially leading to incorrect tokenization and reduced accuracy in understanding the meaning of the text. Extra spaces, if not removed, can result in fragmented tokens. Eliminating these data becomes more uniform, making the training focused more on meaningful content. Fourth, the researcher implements data reduction. The process consists of aggregating text to reduce the size and complexity of the data sets, making the data manageable during a model's training. According to [29], using text records to reduce the volume of data; will help to manage computational resources efficiently, reducing training time. The next step was lemmatization. It is a natural language processing technique similar to stemming. It reduces a word into a base form called "lemma." This technique considers the context of the word in a text, ensuring the reduced word is valid. According to [30], this technique involves organizing words into groups so that they can be analyzed as a single unit. Lemmatization helps create consistent data analysis for text classification, improving results' quality and interpretability. The last step is tokenization. Tokenization it is the process of breaking text into smaller pieces or units called tokens. It converts text into a structured format that can be utilized in language processing. According to [31], the last step was tokenization, an essential data preparation stage for text classification. This will convert text into smaller manageable pieces that models can analyze, and it also provides structure to the text into a sequence of tokens for the model to process and learn effectively. For Naïve Bayes and BI-LSTM, tokenization was done using NLTK; for the BERT model, tokenization must be processed using BertTokenizer, a specialized tool for tokenization that converts raw text into a format suitable for the BERT model. Using the BertTokenizer increases the accuracy of the trained model from 87% to 91%. According to [32], BertTokenizer uses word-piece embeddings and attention processes to boost the model's capability to quickly identify complex patterns and relationships in a text input, which produces increased accuracy and performance of the trained model.

### *3.2 Train and Evaluate the Classifier Model*

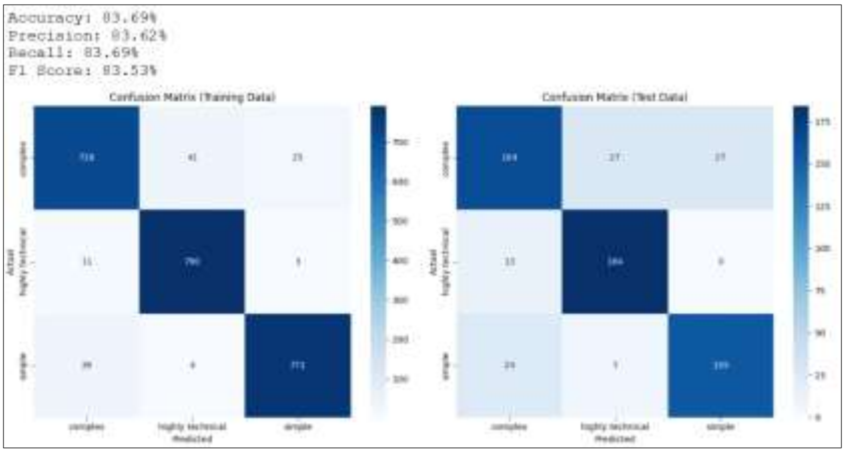
The second and third objectives are to train and evaluate text classifiers using three (3) models: Naïve Bayes, Bi-LSTM (Bidirectional Long Short-Term Memory), and BERT (Bidirectional Encoder Representations from Transformers). Each technique offers advantages and challenges in text classification. For the Naïve Bayes, according to [33], it is a popular and widely used text classification model for research due to a probabilistic and supervised machine learning classifier based on the Bayesian theorem, known for its effectiveness in categorizing text based on the contents. Before optimization of the Naïve Bayes, it achieved an accuracy of 84.86%, precision of 86.45%, recall of 84.86, and an F1 Score of 84.86%. To optimize the model, it was mentioned by [34] that, researchers utilized Naïve Bayes with Term Frequency-Inverse Document Frequency (TF-IDF) extraction technique to improve the accuracy of the classification task. To further enhance the model, the researcher applied a smoothing technique by adjusting the 'alpha' parameter or the smoothing parameter of the model. [35] mentioned that Laplace Smoothing is a technique that increases accuracy for classification tasks by adding a tiny positive value to the estimates to prevent zero values in the model. For the said model, the dataset was split into an 80:20 ratio. 80% was used for training the model, and 20% was used for testing the performance. Figure 4 shows the result of the Naïve Bayes model. The model achieved an accuracy of 89.02%, indicating a high percentage of accuracy in the test instances. For precision, the model is predicted to be 89.55% correct of the time. For recall, the model shows correctly identified 89.02% of all actual instances. And for the f1 score of 88.82% provides a balanced measure of precision and recall, making the model consistent across metrics.



```
Accuracy: 89.02%
Precision: 89.55%
Recall: 89.02%
F1 Score: 88.82%
```

Figure 5: Naïve Bayes Result

The following model is Bi-LSTM. According to [36], this model can understand the context of text by processing text sequences in forward and backward directions to capture contextual information to enhance the model's performance. Before optimization, the model achieved an accuracy of 48.59%, a precision of 36.47%, a recall of 48.59%, and an F1 score of 37.97%. Despite Bi-LSTM's architecture and use of (Adaptive Moment Estimation) Adam as an optimizer, it was outperformed by the Naïve Bayes model classifier. The model achieved an accuracy of 82.20% in epoch 3, which improved to 83.69% in epoch 4, with a slight decrease to 81.03% in epoch 5. The reduction in accuracy results from 4 to 5 suggests overfitting, wherein the model starts to memorize training data rather than learning the patterns. [37] mentioned that overfitting occurs when a model's training results are too close; this results in poor performance as the model captures noise and patterns that do not translate well to new data. Overfitting in this context is indicated by the lack of significant progress in the performance despite the additional training; this suggests that the model becomes more accurate on training but does not perform well on the new data. This highlights the importance of monitoring model performance across epochs to avoid overfitting in training the model.



```
Accuracy: 83.69%
Precision: 83.62%
Recall: 83.69%
F1 Score: 83.53%
```

Figure 6: Bi-LSTM Result

For the last model, BERT takes text classification further by using transformer architecture to provide deep bidirectional representations. According to [38], BERT achieves state-of-the-art performance in a diverse natural language processing task by capturing contextual relationships within text. Before fine-tuning with an optimizer, BERT attained an accuracy of 53%, precision of 34%, recall of 53%, and an F1 Score of 41% by setting hyperparameters, such as Adaptive Moment Estimation with Weight decay (AdamW) optimizer, batch size, and number of epochs. According to [39], AdamW is used to effectively fine-tune the BERT model in various natural language processing applications. With this tuning, BERT outperformed both Naïve Bayes and Bi-LSTM models; BERT achieved an accuracy, precision, and recall of 89% and 88% on the f1 score at epoch 3, which increased to 91% score in all metrics at epoch 4, demonstrating the ability to learn from the data. At epoch 5, BERT's accuracy remains consistent at a 91% score in all metrics, indicating stability in learning. According to [40], there can be instances that the performance can slightly improve and be consistent in the

BERT model due to fine-tuning strategies. The slight movement in BERT's metrics can be attributed to the model's comprehensive learning capacity, where initial epochs capture extensive patterns and succeeding epochs fine-tune the model, leading to a balanced learning and generalization of text contents. Even though BERT's model is computationally intensive and time-consuming, it is suitable for complex text classification.

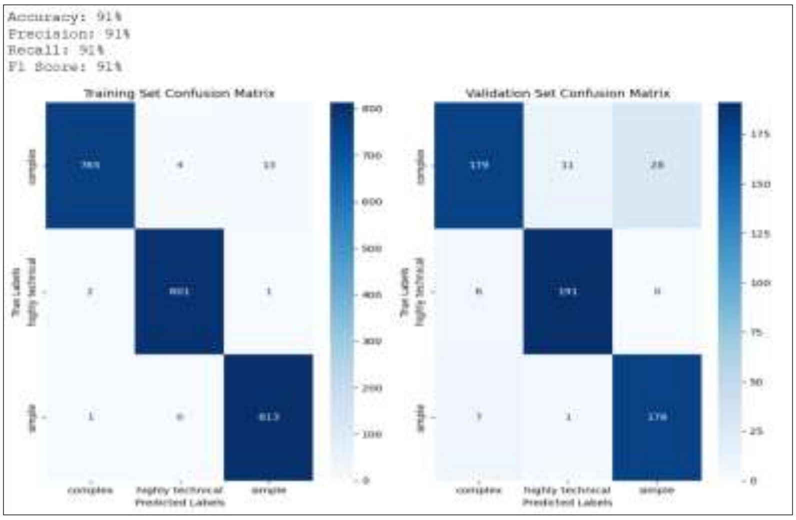


Figure 7: BERT Result

In evaluating the three (3) models for text classification, Naïve Bayes, Bi-LSTM, and BERT. Each model demonstrated strengths and weaknesses. The Naïve Bayes uses the Bayesian theorem to effectively classify text and achieve an Accuracy of 89.02%, Precision of 89.55%, Recall of 89.02%, and F1 Score of 88.82% to achieve these results. The model was enhanced using TF-IDF and Laplace Smoothing, resulting in a balanced precision and recall, making the model reliable for text classification. The Bi-LSTM, on the other hand, uses text sequences in both forward and backward directions to capture and understand textual information, and with Adam as an optimizer, the model only achieved an Accuracy of 83.69%, Precision of 83.62%, Recall of 83.69%, and F1 Score of 83.53% for Epoch 4. The Bi-LSTM model displayed the sign of overfitting from Epoch 5 due to the decrease in score in all metrics. This suggests that the model can capture context effectively but requires careful monitoring to avoid overfitting and ensure generalization to new data. For BERT, it outperformed both Naïve Bayes and Bi-LSTM. It achieved the highest result on Epoch 4 with 91% across all performance matrices, and it maintained stability at 91% across the matrix in the succeeding epoch.

The BERT model can understand and determine the three (3) transaction types based on Contextual Relationships. According to [41], the BERT model can have a deeper understanding of text semantics by generating contextual embeddings that reflect the meaning of words based on the surrounding context, the presence of government terms such as "Agreement," "Resolution," "Legal Inquiry" and other words associated to Highly Technical transactions indicate that BERT model has already learned to identify the context of the words and how to relate to different elements within the text. The BERT model can also recognize complexity with the length and clauses in a text; Simple Transactions might be shorter and more straightforward in text content, and Complex and Highly Technical Transactions can be longer and contain more intricate language patterns. BERT model can also recognize attention weights based on context; the model can give particular attention to the words and ignore or consider surrounding details during inference; for Simple Transaction, BERT would focus on the word "Request" and ignore the surrounding information for the other transaction, the model pay attention to specific terms that signify difficulty to Complex and Highly Technical transactions each training data will provide the model examples and patterns to infer for each transaction types. Despite the BERT's computational intensity, its high results in Accuracy, Precision, recall, and f1 score, combined with balanced learning and generalization of the learned data, the said model able to capture complex contextual relationships in a text, makes the best model for text classification for this study.

### *3.3 Application Development*

For the fourth objective, the development of an application for document classification, the researcher used the Agile Methodology. According to [42], the said methodology in the development of software focuses on small iterations in development to allow changes according to the client's need; this process will enhance the software quality, mitigate risk, proactively address occurring challenges and ensuring the client satisfaction.

The project was divided into sprints, each focused on delivering a specific functionality. The first sprint focused on the creation and testing of the BERT model. After training, the model was saved as an HDF5 file. According to the study of [43], using an HDF5 is tailored to manage intricate datasets effectively; it also supports fast read/write operations for complex datasets and efficiently handles large machine learning models like BERT. Model testing begins by importing necessary libraries, such as NLTK and PyTorch, for natural language processing and Transformers for implementing BERT. A function within the codebase loads a pre-trained BERT model from a save state dictionary, enabling the use of the previously trained model without retraining to save computational resources. The text input will undergo pre-processing and standardization by converting text to lowercase and removing special symbols, numbers, extra spaces, repeated words, and stopwords. This step is critical for enhancing the accuracy and reliability of the model's predictions. Then, the pre-processed input text will be processed by the BERT model, and the predicted index will be into predefined labels.

For the second sprint, the goal is to develop a mobile application to enable real-time text classification from the documents; the researcher focused on integrating the feature that will allow users to scan text from documents using the device’s camera, process the text with the trained BERT model, and display the classification result on the screen. The researcher used Kotlin with Jetpack Compose to build the mobile application. According to [44], Kotlin facilitates effective cross-platform development by enabling code sharing between Android and iOS, minimizing codebase, and streamlining development procedures without sacrificing functionality. While Jetpack Compose was used as a toolkit for building user interfaces, [45] mentioned that Jetpack Compose is a declarative programming paradigm that makes user interface development straightforward. It was also mentioned that it is more effective in code readability and maintenance than XML-based UI development due to the imperative approach of development. For the back-end development of API and for the processing of the trained model, the researcher used the Flask framework. According to [46], Web application development using the Flask framework can be completed quickly and effectively due to its lightweight, adaptive, and flexible design. The back-end web application was hosted on a cloud server, making it accessible to anyone who wants to submit information and receive predictions from the trained model.

### *3.4 Software Quality Evaluation*

For the last objective, the evaluation of the text classification application was conducted using the ISO/IEC 25010:2011 standard. According to [47], it is a comprehensive framework for rating software quality through the eight characteristics: functionality suitability, performance efficiency, compatibility, usability, reliability, security, maintainability, and portability. Also, according to [48] it is used to assess the quality of different types of software, such as mobile applications. Also, it prioritizes stakeholders' input for product quality evaluation. Given the small size of the records section of the Office of the City Administrator with only three (3) individuals, one (1) from the City on Anti-Red Tape, and (1) from the City Information and Communications Technology Office, the total enumeration was conducted to ensure comprehensive feedback.

### *3.4.1 Functionality Suitability*

Functionality Suitability refers to how well the software provides functions that meet users' needs. According to [49], Functional Suitability evaluates whether the product satisfies stakeholders' standards. The evaluation of the first indicator was analyzed using an arithmetic mean statistical tool and a Likert scale.

Table 1 shows the result of Functionality Sustainability. The mean score was 4.50 for the three (3) indicators: Functional Completeness, Functional Correctness, and Functional Appropriateness. The application scored highly in overall functionality, with the respondents highlighting its accuracy in text classification. All the respondents agreed that the application's main feature met their needs in the identification of the correct transaction type, with a total mean score of 4.53 and a descriptive equivalent of Excellent. This indicates that the developed application's main objective is being successfully met.

Table 1. Functionality Suitability Data Result

| Functionality Suitability | Mean Score | Descriptive Equivalent |
|---|---|---|
| Functional Completeness | 4.60 | Excellent |
| Functional Correctness | 4.50 | Excellent |
| Functional Appropriateness | 4.60 | Excellent |
| **Mean Score** | **4.53** | **Excellent** |

### *3.4.2 Performance Efficiency*

For the performance efficiency in software quality, it focuses on how the software application utilizes resources while performing the task. There are three (3) sub indicators needs to be considered Time Behavior refers to response time and throughput of the system, Resource Utilization refers to effective use of resources, and Capacity is the ability to handle maximum load.

Table 2. Performance Efficiency

| Performance Efficiency | Mean Score | Descriptive Equivalent |
|---|---|---|
| Time Behavior | 4.60 | Excellent |
| Resource Utilization | 4.40 | Excellent |
| Capacity | 4.60 | Excellent |
| **Mean Score** | **4.53** | **Excellent** |

Table 2 shows the result of Performance Efficiency. The mean score of both time behavior and capacity is 4.60 with the descriptive equivalent of Excellent, and for Resource Utilization, the result is 4.40 with the descriptive equivalent of Excellent. With a total mean score of 4.53 and a descriptive equivalent of Excellent. The respondents noted that the software application has a quick processing time to capture the text in single and multiple pages and identify the correct transaction type; this indicates that the developed software application has efficiently performed its primary function, achieving a high rating in terms of processing speed and resource usage.

### 3.4.3. Compatibility

Compatibility is the degree to which a system, product, or component can share hardware and software environment and carry out necessary operation and exchange of information. According to [50], software must work properly in a variety of settings and with various applications, which is crucial for user satisfaction and operational efficiency.

Table 3. Compatibility

| Compatibility | Mean Score | Descriptive Equivalent |
|---|---|---|
| Co-existence | 4.40 | Excellent |
| Interoperability | 4.40 | Excellent |
| **Mean Score** | **4.40** | **Excellent** |

Table 3 shows the result of compatibility; both Co-existence and Interoperability got a mean score of 4.40 with a descriptive equivalent of Excellent. With a total mean of 4.40 with a descriptive equivalent of Excellent. The respondents noted that the application performed well across different android devices, making the app interoperable and with no interference with other applications, thereby contributing a high score in this area.

### *3.4.4. Usability*

For the area of usability, it focuses on how well the software application can be utilized and interact with the users to accomplish specific tasks with effectiveness, efficiency, and satisfaction.

Table 4. Usability

| Usability | Mean Score | Descriptive Equivalent |
|---|---|---|
| Appropriateness | 4.40 | Excellent |
| Learnability | 4.60 | Excellent |
| Operability | 4.60 | Excellent |
| User Error Protection | 4.00 | Very Good |
| User Interface Aesthetics | 4.40 | Excellent |
| Accessibility | 4.20 | Excellent |
| **Mean Score** | **4.37** | **Excellent** |

Table 4 shows the result of usability; both Appropriateness and User Interface Aesthetics got a mean score of 4.40 with a descriptive equivalent of Excellent, for Learnability and Operability got a mean score of 4.60

with a descriptive equivalent of Excellent, User Error Protection got a mean score of 4.00 with a descriptive equivalent of Very Good, and Accessibility got 4.20 score with a descriptive equivalent of Excellent, with the total mean score of 4.37 with the descriptive equivalent of Excellent. The respondents noted that the layout and placements of functionality contribute to the overall user-friendliness. This indicates that the software application is well-designed.

### *3.4.5 Reliability*

Reliability refers to the software application’s ability to perform consistently under various conditions. According to [51], it is the software application's ability to tolerate faults and bounce back from failures.

Table 5. Reliability

| **Reliability** | **Mean Score** | **Descriptive Equivalent** |
|---|---|---|
| Maturity | 4.40 | Excellent |
| Availability | 4.20 | Excellent |
| Fault Tolerance | 4.20 | Excellent |
| Recoverability | 4.20 | Excellent |
| **Mean Score** | **4.25** | Excellent |

Table 5 shows the result of Reliability; Maturity got a mean score of 4.40 with a descriptive equivalent of Excellent. For Availability, Fault Tolerance and Recoverability got a mean score of 4.20 with the descriptive equivalent of Excellent. Overall, the area of Reliability got a total mean score of 4.25 with the descriptive equivalent of Excellent. The respondents noted that the application has a stable performance and performs consistently under various conditions.

### *3.4.6. Security*

Security refers to the protection embedded in the application; this will prevent unauthorized access, modification, and destruction; this indicator ensures that the software application operates securely and information is protected

Table 6. Security

| **Security** | **Mean Score** | **Descriptive Equivalent** |
|---|---|---|
| Confidentiality | 4.40 | Excellent |
| Integrity | 4.40 | Excellent |
| Non-repudiation | 3.80 | Very Good |
| Authenticity | 4.00 | Very Good |
| Accountability | 4.00 | Very Good |
| **Mean Score** | **4.12** | **Very Good** |

Table 6 shows the result of Security for Confidentiality and Integrity got a mean score of 4.40 with a descriptive equivalent of Excellent, Non-repudiation got a mean score of 3.80 with a descriptive equivalent of Very Good; for Authenticity and Accountability got a mean score of 4.00 with a descriptive equivalent of 4.00. Overall, Security got a total mean score of 4.12, which is a descriptive equivalent of Very Good. The respondents noted that the application can be accessed using a username and password, and improvements to the login process can be made to enhance Security.

### *3.4.7. Maintainability*

Maintainability refers to how easily a software product can be modified to correct faults and improve system performance. This area ensures that the system can be efficiently maintained and updated over time, allowing improvement, troubleshooting, and integration when needed without compromising software stability.

Table 7. Maintainability

| Maintainability | Mean Score | Descriptive Equivalent |
|---|---|---|
| Modularity | 4.20 | Excellent |
| Reusability | 4.40 | Excellent |
| Analysability | 4.20 | Excellent |
| Modifiability | 4.40 | Excellent |
| Testability | 4.20 | Excellent |
| **Mean Score** | **4.28** | **Excellent** |

Table 7 shows the result of Maintainability. For the sub indicators Reusability and Modifiability, the mean score was 4.40, with a descriptive equivalent of Excellent, while Modularity, Analysability, and Testability got a mean score of 4.20, with a descriptive equivalent of Excellent. Overall, this area got a total mean score of 4.28, with a descriptive equivalent of Excellent. The respondents noted that the application components were designed to accommodate future changes and improvements.

### *3.4.8. Portability*

Portability describes the quality of software applications that can be transferred from one environment to another with minimal resources and effort. This area ensures the software application can operate effectively in different hardware devices and operating software versions.

Table 8. Portability

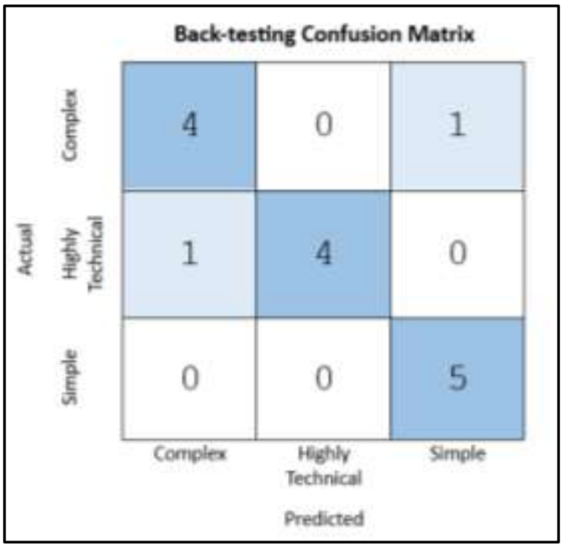
| Portability | Mean Score | Descriptive Equivalent |
|---|---|---|
| Adaptability | 4.40 | Excellent |
| Instability | 4.40 | Excellent |
| Replaceability | 4.20 | Excellent |
| **Mean Score** | **4.33** | **Excellent** |

Table 8 shows the result of Portability; both Adaptability and Instability got a mean score of 4.40 with the descriptive equivalent of Excellent and Replaceability of 4.20 with a descriptive equivalent of Excellent. Overall, Portability got a total mean score of 4.33 with a descriptive equivalent of Excellent. The respondents commended the software application's versatility to be installed in various mobile devices, regardless of brand or operating system, making it accessible and user-friendly for diverse users.

Overall, the stakeholders expressed satisfaction with the developed application, getting a grand mean of 4.35, which corresponds to a descriptive equivalent of Excellent, noting the application's reliability, usability, and efficiency in performing real-time classification of document transaction types. The survey indicates that the application not only meets the current needs of the stakeholders but also the software application and components have a potential for future enhancements and integration with other software applications.

### *3.5. Backtesting*

Back-testing refers to testing the software application to assess the performance and accuracy of output. According to [52], using back-testing can allow researchers to determine the effectiveness of a predicted model; it is an essential step to validate and refine the prediction results that utilize a machine learning technique. The researcher conducted back-testing for the developed software application by predicting 15 new documents not included in the data training. The documents used in this testing reflect potential real-world scenarios the software will encounter during the deployment.



Back-testing Confusion Matrix

| Actual \ Predicted | Complex | Highly Technical | Simple |
|---|---|---|---|
| Complex | 4 | 0 | 1 |
| Highly Technical | 1 | 4 | 0 |
| Simple | 0 | 0 | 5 |

Figure 8: Back-testing Result

Accuracy:

$$Accuracy = \frac{Total\,TP}{Total\,number\,of\,\,Samples} = \frac{4+4+5}{15} = \frac{13}{15} = 0.87$$

Precision:

$$Precision = \frac{TP}{TP+FP}$$

$$Complex\text{:}\ \frac{4}{4+1} = \frac{4}{5} = 0.8$$

$$Highly\,Technical\text{:}\ \frac{4}{4+0} = 1.0$$

$$Simple\text{:}\frac{5}{5+1} = 0.833$$

$$Average\ Precision = \frac{0.8+1.0+0.833}{3} = \frac{2.633}{3} = 0.878$$

Recall:

$$Recall = \frac{TP}{TP+FN}$$

$$Complex\text{:}\ \frac{4}{4+1} = \frac{4}{5} = 0.8$$

$$Highly\,Technical\text{:}\ \frac{4}{4+1} = \frac{4}{5} = 0.8$$

$$Simple\text{:}\frac{5}{5+0} = 1.0$$

$$Average\ Recall = \frac{0.8+0.8+1.0}{3} = \frac{2.6}{3} = 0.867$$

F1 Score:

$$F1\,Score = 2 * \frac{Precision * \ Recall}{Precision + Recall}$$

$$Complex: 2 * \frac{0.8 * 0.8}{0.8 + 0.8} = \frac{1.28}{1.6} = 0.8$$

$$Highly\,Technical: 2 * \frac{1.0 * 0.8}{1.0 + 0.8} = \frac{1.6}{1.8} = 0.889$$

$$Simple: 2 * \frac{0.833 * 1.0}{0.833 + 1.0} = 0.909$$

$$Average\,F1\,Score = \frac{0.8 + 0.889 + 0.909}{3} = \frac{2.598}{3} = 0.866$$

The back-testing analysis revealed that the software application achieved an accuracy of 87.8%, a precision of 87.8%, a recall of 86.7%, and an f1 score of 86.6%, indicating that the model correctly predicted the transaction types of 13 out of 15 documents; this shows that it is accurate in the identification of simple, complex, and highly technical transactions. The result of precision highlighted the model's reliability in reducing false positive results, while recall shows the model's effectiveness in capturing recent and relevant document transactions. The F1 Score, which balances the precision and recall, confirms that the model has a balanced performance. These results validate the software's capability to predict transaction types effectively.

## 4. Conclusion

Building a dataset for document transaction type classifier requires a thorough data pre-processing. This involves converting text to lowercase, removing stopwords, punctuation marks, extra spaces, special characters, and also performing data reduction, lemmatization and tokenization. These pre-processing are essential for enhancing the performance metrics during the training of the models.

In training the document transaction type classifier model, incorporating hyperparameter tuning on the training process, such as optimizers, smoothing techniques, vectorization, estimations, batch sizes, and number of epochs helps optimize the models performance leading to better accuracy and generalization of unseen data. Properly tuned hyperparameters can result faster trainings, reduce overfitting and ensure that the trained model will perform well.

In evaluating the best algorithms for document transaction type classifier, BERT emerged as the top performing model, excelling all metrics. Despite enhancements and tuning done with the Naïve Bayes and Bi-LSTM, BERT's ability to capture complex contextual relationships resulted higher number of results making it the best choice for classification task.

In the development of application, Agile methodology was used by the researcher to break down tasks into manageable sprints, achieving goals one at a time at the same time incorporating feedback per sprint to ensure a high-quality of outcomes.

The overall software quality achieved the total mean score of 4.25, indicating that the application is excellent meets the standards across various areas of software quality. This comprehensive evaluation standard ensures that the developed application will reflect the holistic approach to software quality assessment

## References

[1] S. B. Vuyokasi, "*A comparative analysis of the use of e-government services by small businesses"*. University of Johannesburg.

[2] D. MacLean, R. Titah, "A Systematic Literature review of Empirical Research on the Impacts of e-Government: A Public Value Perspective".

[3] *Story - e-Estonia*. (2023c, February 1). e-Estonia. https://e-estonia.com/story/

[4] *Singapore Land Authority*. (n.d.). https://www.sla.gov.sg/

[5] J. O. H. Engineering, "Retracted: Application of internet of things and naive bayes in public health environmental management of government institutions in China". *Journal of Healthcare Engineering*, vol. 1, 2023. https://doi.org/10.1155/2023/9815658

[6] A. H. Oliaee, S. Das, J. Liu, M. A. Rahman, "Using Bidirectional Encoder Representations from Transformers (BERT) to classify traffic crash severity types", *Natural Language Processing Journal*, vol. 3, pp. 100007, 2023. https://doi.org/10.1016/j.nlp.2023.100007

[7] P. P. Pan, C. Yijin, "Automatic subject classification of public messages in e-government affairs", *Data and Information Management*, vol. 5, no. 3, pp. 336–347, 2021. https://doi.org/10.2478/dim-2021-0004

[8] D. Jiapeng, G. Shuaiying, T. Yuan, Y. Tengyuan, "Enhancing the Governance Capabilities through Smart Technology: Scenario Application of Image Recognition and Its Effects in Chinese Local Governance. (n.d.-b). *Digital Object Identifier*, 2020. https://ieeexplore.ieee.org/document/9186691

[9] *House of Representatives press releases*. (n.d.-b). https://www.congress.gov.ph/press/details.php?pressid=12406

[10] *Republic Act No. 11032*. (n.d.). https://lawphil.net/statutes/repacts/ra2018/ra_11032_2018.html

[11] J. V. Brocke, A. R. Hevner, A. Maedche, "Introduction to Design Science Research", *In Progress in IS*, pp. 1–13, 2020. https://doi.org/10.1007/978-3-030-46781-4_1

[12] S. Mdletshe, M. Oliveira, B. Twala, "Enhancing medical radiation science education through a design science research methodology", *Journal of Medical Imaging and Radiation Sciences*, vol. 52, no. 2, pp. 172–178, 2021. https://doi.org/10.1016/j.jmir.2021.01.005

[13] M. Yazdani, M. Loosemore, M. Mojtahedi, D. Sanderson, M. Haghani, "An integration of operations research and design science research methodology: With an application in hospital disaster management", *Progress in Disaster Science*, pp. 100300, 2023. https://doi.org/10.1016/j.pdisas.2023.100300

[14] S. Mdletshe, O. S. Motshweneng, M. Oliveira, B. Twala, "Design science research application in medical radiation science education: A case study on the evaluation of a developed artifact", *Journal of Medical Imaging and Radiation Sciences*, vol. 54, no. 1, pp. 206–214, 2022. https://doi.org/10.1016/j.jmir.2022.11.007

[15] S. Maataoui, G. Bencheikh, G. Bencheikh, "Predictive Maintenance in the Industrial Sector: A CRISP-DM Approach for Developing Accurate Machine Failure Prediction Models", *IEEE*, 2023.

[16] J. Saltz, "CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps", *IEEE,* 2021.

[17] P. Sudhir, V. D. Suresh, "Comparative study of various approaches, applications and classifiers for sentiment analysis", *Global Transitions Proceedings*, vol. 2, no. 2, pp. 205–211, 2021. https://doi.org/10.1016/j.gltp.2021.08.004

[18] R. Ofori-Boateng, M. Aceves-Martins, C. Jayne, N. Wiratunga, C. F. Moreno-García, "Evaluation of Attention-Based LSTM and Bi-LSTM networks for abstract text classification in systematic literature review automation", *Procedia Computer Science,* vol. 222, pp. 114–126. https://doi.org/10.1016/j.procs.2023.08.149

[19] A. Turchin, S. Masharsky, M. Žitnik, "Comparison of BERT implementations for natural language processing of narrative medical documents", *Informatics in Medicine Unlocked*, vol. 36, pp. 101139. https://doi.org/10.1016/j.imu.2022.101139

[20] Salonso, D. *Valero-Carreras, J. Alcaraz, M. Landete, "Comparing two SVM models through different metrics based on the confusion matrix", Computers and Operations Research,* vol. 152, pp. 106131 – CIO, 2023. https://cio.umh.es/2023/01/12/valero-carreras-d-alcaraz-j-landete-m-2023-comparing-two-svm-models-through-different-metrics-based-on-the-confusion-matrix-computers-and-operations-research-152106131-2/

[21] A. Hinderks, F. J. D. Mayo, J. Thomaschewski, M. J. Escalona, "Approaches to manage the user experience process in Agile software development: A systematic literature review", *Information & Software Technology*, vol. 150, pp. 106957, 2022. https://doi.org/10.1016/j.infsof.2022.106957

[22] A. Alami, O. Krancher, M. Paasivaara, "The journey to technical excellence in agile software development", *Information & Software Technology*, vol. 150, pp. 106959, 2022. https://doi.org/10.1016/j.infsof.2022.106959

[23] M. Klima, M. Bures, K. Frajtak, V. Rechtberger, M. Trnka, X. Bellekens, T. Cerny, B. S Ahmed, B, "*Selected Code-Quality Characteristics and Metrics for Internet of Things Systems",* 2022. https://ieeexplore.ieee.org/document/9762941

[24] H. Salehinejad, A. M. Meehan, P. A. Rahman, M. A. Core, B. J. Borah, P. J. Caraballo, "Novel machine learning model to improve performance of an early warning system in hospitalized patients: a retrospective multisite cross-validation study", *EClinicalMedicine*, vol. 66, pp. 102312, 2023. https://doi.org/10.1016/j.eclinm.2023.102312

[25] K. Anjaria, "Knowledge derivation from the Likert scale using Z-numbers" *Information Sciences*, vol. 590, pp. 234–252, 2022. https://doi.org/10.1016/j.ins.2022.01.024

[26] C. Schröer, F. Kruse, J. M. Gómez, “A Systematic Literature Review on Applying CRISP-DM Process model”. *Procedia Computer Science*, vol. 181, pp. 526–534, 2021. https://doi.org/10.1016/j.procs.2021.01.199

[27] K. Madatov, S. Bekchanov, J. Vičič, “Dataset of Karakalpak language stop words”, *Data in Brief*, vol. 48, pp. 109111, 2023. https://doi.org/10.1016/j.dib.2023.109111

[28] D. J. Ladani and N. P. Desai, "Automatic stopword Identification Technique for Gujarati text," *2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)*, Gandhinagar, India, pp. 1-5, 2021. doi: 10.1109/AIMV53313.2021.9670968.

[29] A. AlKarawi, K, AlJanabi, “Data Reduction Techniques: A Comparative study”, *Journal of Kufa for Mathematics and Computer*, vol. 9, no. 2, pp. 1–17, 2022. https://doi.org/10.31642/jokmc/2018/090201

[30] M. Karwatowski, M. Pietron, “*Context based lemmatizer for Polish language*”, 2022, arXiv.org. https://arxiv.org/abs/2207.11565

[31] P. Prakrankamanant and E. Chuangsuwanich, "Tokenization-based data augmentation for text classification," *2022 19th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, Bangkok, Thailand, pp. 1-6, 2022. doi: 10.1109/JCSSE54890.2022.9836268.

[32] Y. Guo, Z. Xie, X. Chen, H. Chen, Wang, L., Du, H., Wei, S., Zhao, Y., Li, Q., & Wu, G. (2022, November 27). *ESIE-BERT: Enriching Sub-words Information Explicitly with BERT for Joint Intent Classification and SlotFilling*. arXiv.org. https://arxiv.org/abs/2211.14829

[33] I. Dawar and N. Kumar, "Text Categorization By Content using Naïve Bayes Approach," *2023 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON)*, Jaipur, India, pp. 1-6, 2023. doi: 10.1109/IEMECON56962.2023.10092372.

[34] Chingmuankim and R. Jindal, "Classification and Analysis of Textual data using Naive Bayes with TF-IDF," *2022 4th International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)*, KualaLumpur, Malaysia, pp. 1-9, 2022. doi: 10.1109/ICECIE55199.2022.10000309.

[35] A. P. Noto, D. R. S. Saputro, “Classification data mining with Laplacian Smoothing on Naïve Bayes method”. *AIP Conference Proceedings*, 2023. https://doi.org/10.1063/5.0116519

[36] H. Zhang, C. Ma, Z. Jiang, J. Lian, "Image Caption Generation Using Contextual Information Fusion With Bi-LSTM-s," in *IEEE Access*, vol. 11, pp. 134-143, 2023. doi: 10.1109/ACCESS.2022.3232508.

[37] J. Schmidt, “*Testing for overfitting*”, 2022. arXiv.org. https://arxiv.org/abs/2305.05792

[38] J. Sawicki, M. Ganzha, M. Paprzycki, “The state of the art of Natural Language Processing - a systematic automated review of NLP literature using NLP techniques. *Data Intelligence*, pp. 1–47, 2023. https://doi.org/10.1162/dint_a_00213

[39] M. U. Joseph, M. Jacob, "Developing a Real time model to Detect SMS Phishing Attacks in Edges using BERT," *2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*, Kochi, India, pp. 1-7, 2022. doi: 10.1109/IC3SIS54991.2022.9885427.

[40] S. Shen, Z. Dong, J. Ye, L. Ma, Z. Yao, A.Gholami, M. W. Mahoney, K. Keutzer, “Q-BERT: Hessian based Ultra Low precision Quantization of BERT”. *Proceedings of the . . . AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 8815–8821, 2022. https://doi.org/10.1609/aaai.v34i05.6409

[41] R. Dodda and S. B. Alladi, “BERT-based document clustering: unveiling semantic patterns in 20News Group, Reuters, and BBC Sports Corpora,” *Authorea (Authorea)*, 2024, doi: 10.22541/au.171506422.20645846/v1.

[42] M. H. Zahedi, A. R. Kashanaki, E. Farahani, “Risk management framework in Agile software development methodology”, *International Journal of Power Electronics and Drive Systems/International Journal of Electrical and Computer Engineering*, vol. 13, no. 4, pp. 4379, 2023. https://doi.org/10.11591/ijece.v13i4.pp4379-438

[43] S. Lee, K. Hou, K. Wang, S. Sehrish, M. Paterno, J. Kowalkowski, Q. Koziol, R. Ross, A. Agrawal, A. Choudhary, W. Liao, “*A case study on parallel HDF5 Dataset concatenation for High Energy Physics data analysis*”, 2023. arXiv.org. https://arxiv.org/abs/2205.01168

[44] I. Olenych, R. Korostenskyi, “Analysis of The Effectiveness of Using Kotlin Multiplatform Mobile Technology for Creating Cross-Platform Applications. *Elektronìka Ta Ìnformacìjnì Tehnologìï, 21*. https://doi.org/10.30970/eli.21.3

[45] S. Marchenko, “*Jetpack Compose: New Approaches To Android Ui Development”,* 2023. http://baltijapublishing.lv/omp/index.php/bp/catalog/view/291/8064/16856-1

[46] J. Xiao, L. Wang, Y. Cheng, J. Zhang, J. Hu, S. Tan, Y. Su, H. Zhou, "Web Front-end Development based on Flask Architecture for Image Recognition," *2023 IEEE 7th Information Technology and Mechatronics Engineering Conference (ITOEC)*, Chongqing, China, pp. 1964-1967, 2023. doi: 10.1109/ITOEC57671.2023.10291828.

[47] M. R. A. Assifa, F. Setiadi, R. G. Utomo, “Evaluation of Software Quality For I-Office Plus Applications Using Iso/Iec 25010 and Kano Model. *Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika*, vol. 8, no. 2, pp. 561–571. https://doi.org/10.29100/jipi.v8i2.3561

[48] Y. I. Irawan, E. S. Negara, “Evaluation of Software Quality Assurance Silampari Smart City of Lubuklinggau based on ISO/IEC 25010:2011 Analysis Model”. *2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, 2023. https://doi.org/10.1109/icimcis56303.2022.10017834

[49] K. Moumane, A. Idri, F. E. Aouni, J. Laghnimi, N. C. Benabdellah, O. Hamal, “ISO/IEC 25010-based quality evaluation of three mobile applications for reproductive health services in Morocco. *Clinical and Experimental Obstetrics & Gynecology*, vol. 51, no. 4, pp. 88, 2024. https://doi.org/10.31083/j.ceog5104088

[50] A. Tursia, D. Pernadi, “Pengukuran kualitas perangkat lunak Persona berdasarkan ISO/IEC 25010 menggunakan tingkat capaian responden (TCR)”, *Digital Transformation Technology*, vol. 3, no. 2, pp. 879–887, 2024. https://doi.org/10.47709/digitech.v3i2.3416

[51] B. I. Rumabar, E. Maria, “Evaluasi kualitas ShopeePay menggunakan ISO/IEC 25010. *JURNAL SISTEM INFORMASI BISNIS*, vol. 14, no. 1, pp.54–61, 2023. https://doi.org/10.21456/vol14iss1pp54-61

[52] I. Gasanov, A. Ereshko, “Computational experiments on Back-Testing Complex using the forecast of artificial neural network”, *2022 15th International Conference Management of Large-scale System Development (MLSD)*, pp. 1–4, 2022. https://doi.org/10.1109/mlsd55143.2022.9934431