International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I11P115

Volume 73 Issue 11, 193-207, November 2025

Original Article

Application of Text Mining in Categorizing Complaints Related to Teaching Materials at XYZ University

Hanson Geraldi Pardede[1], Tuga Mauritsius[2]

[1,2]*Information Systems Management Department, Bina Nusantara University Graduate Program - Master of Information Systems Management, Bina Nusantara University, Jakarta, Indonesia.*

[1]*Corresponding author : hanson.pardede@binus.ac.id*

Received: 25 July 2025 Revised: 30 October 2025 Accepted: 10 November 2025 Published: 25 November 2025

Abstract - *XYZ University is an institution that relies on Bahan Ajar (BA) as the primary learning medium, which is mandatory for every student. However, in its implementation, numerous complaints related to BA continue to be reported. Currently, complaint handling at XYZ University still involves manual categorization by the customer service team. This practice leads to several issues, such as delayed complaint resolution, inaccurate problem handling, and the potential degradation of the university's reputation. This research aims to design and evaluate a model that enables XYZ University to automatically categorize BA-related complaints from students. This study proposes a novel approach by using the CRISP-DM framework and integrating Indonesian Bidirectional Encoder Representations from Transformers (IndoBERT) with the Naive Bayes (NB) machine learning algorithm, as well as applying a combination of hyperparameter customization to Neural Network (NN) and Support Vector Machine (SVM) algorithms to categorize BA-related complaints. The results show that the NN algorithm, using a combination of hyperparameters consisting of four hidden layers with sequential neuron counts of 512, 256, 128, and 64; a dropout rate of 0.4 on each hidden layer; batch normalization applied to each layer; a learning rate of 0.0005; ReLU activation; softmax on the output layer; CrossEntropyLoss as the loss function; Adam optimizer; and 200 epochs, achieved the best performance. The model evaluation resulted in an accuracy of 0.9196, a precision of 0.9200, a recall of 0.9196, and an F1 score of 0.9196.*

Keywords - *Text mining, Machine Learning, Categorization, Hyperparameters, CRISP-DM.*

1. Introduction

XYZ University is an institution that organizes educational services with the Open and Distance Higher Education system, in which students are not required to attend campus in person. The learning media consists of teaching modules or Bahan Ajar (BA) in the form of printed materials, referred to as Bahan Ajar Cetak (BAC), and non-printed materials, such as Bahan Ajar Digital (BAD), online audio/video content, radio broadcasts, and television programs. BAC and BAD serve as the primary learning resources and are mandatory purchases for every student, which are then distributed directly to each student.In practice, there are numerous complaints related to BA.

These complaints must be handled appropriately to ensure they are resolved quickly and effectively. If these complaints are not handled appropriately, they will undermine the university's credibility and reputation. This will also impact prospective students considering continuing their studies. This will also impact the satisfaction and trust of current students, potentially reducing their motivation to continue their studies at the university.

To handle student complaints related to BA, XYZ University provides an application that can manage these complaints. Complaints are submitted through the application, which is managed by the BA Customer Service (CS) team, so that each complaint can be monitored and followed up according to procedures. In the application, the CS team is required to manually select the BA complaint category for each complaint received so that it can be resolved by the person responsible for each complaint, as seen in Figure 1. Based on the data obtained (Table 1), the number of BA complaints at XYZ University has increased along with the increase in student numbers. As the number of complaints increases, various issues arise due to the manual selection of complaint categories. One such issue is misclassification. This misclassification can lead to various negative consequences, such as slower complaint handling due to repetitive processes. While the Customer Service (CS) team should handle issues according to their categories, the team should review and ensure that complaints are handled by the appropriate division. This misclassification can also lead to inaccurate resolutions, as complaints may be handled by CS officers who are not responsible for their categories.

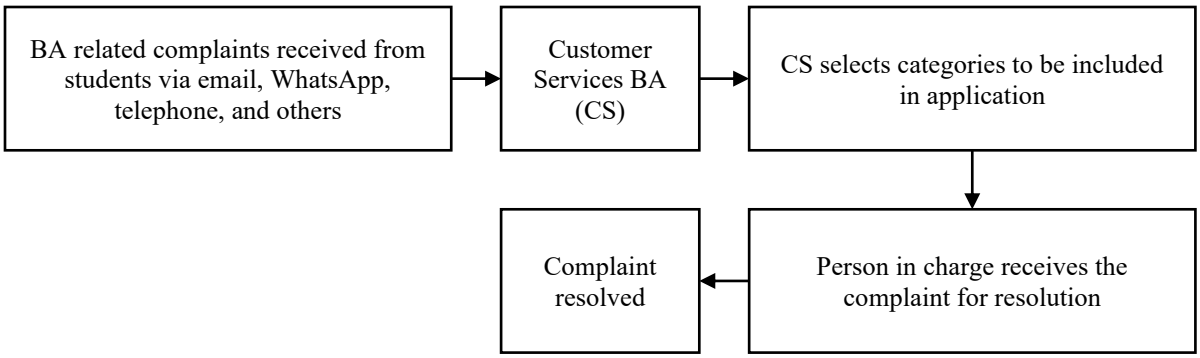

Fig. 1 Customer Service (CS) complaint related to the BA handling process

Table 1. Number of students and BA-related complaints

Semester	Number of Students	Number of BA-related Complaints
2022 even	439.311	3.154
2023 Odd	525.696	4.835
2023 even	549.553	4.981

Furthermore, the university's reputation can suffer if student complaints are not handled quickly and accurately. Decision-making by stakeholders aimed at improving BA services may also be compromised, as it relies on inaccurate category data caused by manual misclassification. On the other hand, if the classification process becomes more accurate, complaint handling will significantly improve. Stakeholders can develop strategies to optimize processes in specific problem categories that are still underperforming. Consequently, even though the number of students increases, the complaints can be effectively reduced because previously unidentified or mismanaged issues can now be addressed more efficiently.

Although the adoption of machine learning approaches in text mining and complaint classification is increasing, many studies conducted by researchers in Indonesia still rely on traditional vectorization techniques, such as TF–IDF, combined with algorithms like NB and SVM. These conventional methods are limited in their ability to capture contextual meaning and semantic overlap among complaint categories, which often results in misclassification and slower routing in practice. Transformer-based embeddings such as IndoBERT have the capacity to represent linguistic context more effectively; however, previous studies have typically implemented IndoBERT through end-to-end fine-tuning on a single deep-learning architecture, without assessing its performance as a unified embedding applied to multiple classifiers.

This study addresses this problem by creating a complaint classification model that uses semantic representations in the form of contextual embeddings from IndoBERT-Large P2. This representation was used for three algorithms: NB, SVM, and NN. Each algorithm was trained and evaluated. Parameter optimization was also performed on SVM and NN to improve model performance and stability. The objectives of this study were: (1) to compare the results of the three algorithms to find the most effective model, and (2) to analyze classification errors per category to identify the root causes of errors within each complaint category using a confusion matrix and error taxonomy. With this approach, this study is expected to identify the best model and provide a deeper understanding of the impact of automation in supporting more accurate and consistent complaint management at XYZ University.

1.1. Research gap

Much research on text mining and machine learning is used for automatic text classification. This research is often conducted in areas such as spam filtering, fake news detection, customer feedback analysis, and public service complaints. Many of these studies use traditional algorithms such as NB, SVM, and NN, which generally rely on vectorization methods like TF-IDF and Word2Vec. While some recent studies have incorporated pre-trained neural models and embeddings such as BERT to improve classification quality, most research remains focused on fine-tuning within a single deep learning architecture, rather than using these embeddings as a shared semantic representation across multiple classifiers.In research that uses the Indonesian language as data, existing studies [1] and [2] mostly use NB and SVM models with TF-IDF or Bag-of-Words features for their models. Meanwhile, some recent efforts in references [3, 4] utilize IndoBERT through direct fine-tuning to analyze sentiment or reviews.

However, these approaches have not explored the use of comparative IndoBERT embeddings as a unified feature representation, combined with specific hyperparameter optimization for SVM and Neural Network. Consequently, there remains a lack of empirical evidence testing the performance of IndoBERT contextual embeddings when applied uniformly across various paradigms, such as probabilistic, margin-based, and neural, in the same experimental setting, as shown in Table 2. Due to the increasing number of student complaints and the inefficiency of the manual categorization process at XYZ University, a solution is needed that can automate the process effectively and accurately. This study addresses this problem by using the IndoBERT-Large P2 vector representation, as well as a

comparative modeling approach based on Naive Bayes, SVM with adjusted parameters, and NN with optimized architecture. The data used comes from real complaint records filed by students at XYZ University. The data used in this study includes actual complaint records from students at XYZ University. Through this approach, the research seeks to improve categorization accuracy, reduce semantic overlap between complaint categories, and provide a scalable foundation for more effective complaint handling and decision support.

Table 2. Previous consolidated results

No	Author(s), Year	Domain / Dataset	Main Algorithm(s)	Language	Feature / Embedding	Evaluation Metrics	Limitation(s)
1	Surjandari et al., 2016 [6]	LAPOR! public complaint reports	SVM	Indonesian	TF–IDF	Accuracy	Manual categorization; non-contextual
2	Hermanto et al., 2020 [22]	University complaint service	NB, SVM	Indonesian	TF–IDF	Accuracy, CM	Classical features; no contextual embedding
3	Gozali et al., 2020 [1]	Student e-complaint	Naïve Bayes	Indonesian	TF–IDF	Accuracy, F1	Small dataset; no error analysis
4	Anwar et al., 2021 [2]	Public complaint portal	RF, GB	English	TF–IDF	Accuracy, P/R/F1	No embedding; feature-based only
5	Tejavath & Hirwarkar, 2020 [12]	Text mining benchmark	NB, SVM, DT	English	BOW	Accuracy	No contextual embedding
6	Penchala et al., 2024 [11]	Generic text classification	SVM, NN	English	TF–IDF, Word2Vec	Accuracy	No complaint data; benchmark only
7	Nissa & Yulianti, 2023 [3]	Customer reviews	IndoBERT	Indonesian	Contextual IndoBERT	Accuracy, F1	Single-model fine-tuning
8	Asri et al., 2025 [4]	PLN Mobile app reviews	IndoBERT	Indonesian	Contextual IndoBERT	Accuracy, P/R/F1	Sentiment only; not complaint domain
9	Çallı et al., 2022 [23]	Airline complaints	Topic Modeling (LDA)	English	Word-topic	Topic coherence	Qualitative; not classification
10	Bazzan et al., 2023 [24]	Real-estate complaint management	NLP + ML	Portuguese	Structured features	Accuracy	Process-level; no embeddings
11	This Study (2025)	University complaint	NB, SVM, NN	Indonesian	IndoBERT-Large P2 contextual	Accuracy, P/R/F1, CM	(proposed model) IndoBERT all model comparison with SVM and NN hyperparameter tuning

2. Literature Review

Manually classifying articles or texts requires significant time and effort. Humans need a lot of time and energy to categorize an article or text manually. To overcome this, many researchers have tried to conduct automation studies in document classifiers with text mining [5]. The work in [6] categorization of reports submitted through the LAPOR! The application was found to be ineffective, so it is necessary to use text mining with the Support Vector Machine (SVM) algorithm to facilitate analysis according to the problem groups that are often reported by the public. The paper [7] explores text mining in basic machine learning, already used to categorize messages into several categories. Furthermore, text mining techniques have been developed to be more advanced, so that predetermined classifications of text documents can automatically determine groups of similar documents. In another study, the Naive Bayes algorithm was used to classify spam messages received on mobile phones, demonstrating higher efficiency than both SVM and Random Forest algorithms in spam message classification tasks [8]. A recent study [9] categorized data using text mining with the help of Naive Bayes. Text mining itself is a technique used to

extract information from text by indexing each word in unstructured textual data. Through this process, information can be categorized according to the content of the text. The naive bayes algorithm is one that is known to be effective, efficient, and has good accuracy in producing categorization/classification of textual data. Text classifiers are related to each syllable on pages that are interrelated. Text classification can use various kinds of machine consider how often words appear.

The paper [11] assesses neural network algorithms that have demonstrated very high accuracy in classification tasks with model evaluation using accuracy, precision, and recall. In another study, the use of Bidirectional Encoder Representations from Transformers (BERT) vector representations (embeddings) was shown to have significant results in improving both accuracy and generalization. In another study [12], text classification was applied to analyze customer review perceptions of products offered by various industries.

The study used the IndoBERT encoder pre-trained on Indonesian. IndoBERT was combined with various machine learning algorithms, thereby increasing the effectiveness and accuracy of the model. Model evaluation for all models used in the study included F1 score, accuracy, and hamming loss. The paper in [12] proposed that the algorithms NB, Random Forest (RF), and SVM were compared to determine which algorithm was most effective in classifying text mining and categorizing text. To evaluate the models, researchers used F1 score, accuracy, recall, and precision. The results showed that naïve Bayes performed well on text data and features that were not strongly correlated. Despite advances in text mining applications, only a few studies have addressed the automatic classification of student complaints in the Indonesian higher education context using IndoBERT. This research aims to address this key gap.

3. Methodology

Figure 2 illustrates the methodology for this research, which is combined with the CRISP DM steps.

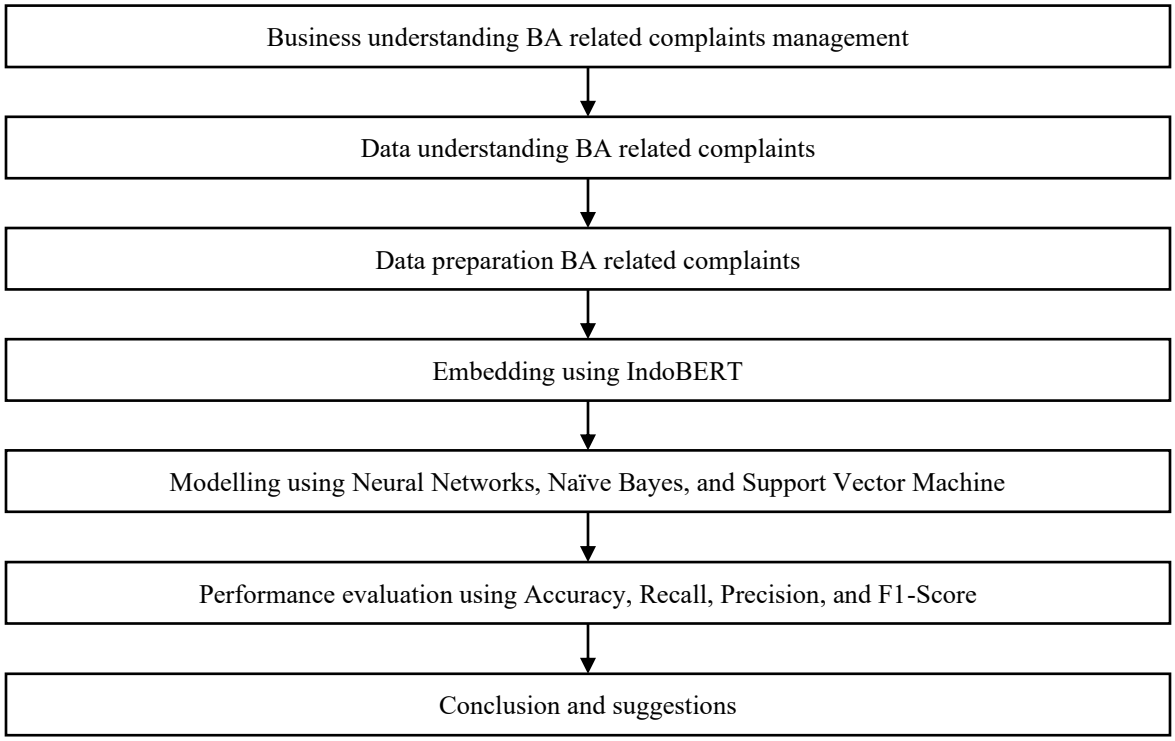

Fig. 2 Research process diagram

3.1. Business Understanding BA Related Complaints Management

This stage aims to understand the business processes at XYZ University related to handling student complaints related to the BA. The university provides educational services using the Open and Distance Learning system, in which the BA serves as the primary learning material that every student is required to possess. The BA consists of Bahan Ajar Cetak (BAC) or printed teaching materials, and Bahan Ajar Digital (BAD) or digital teaching materials. As the number of students at XYZ University increases, many complaints have been made regarding the learning process using BA.

3.2. Data Understanding BA Related Complaints

At this stage, the data used is BA-compliant data at XYZ University. The time range selected for analysis includes

complaints that were manually categorized between August 2023 and May 2024. The data were obtained from the university's internal unit database and extracted using HeidiSQL. The dataset structure includes the following fields: id_keluhan (complaints ID), no_tiket (ticketing complaints number), tgl_input (complaints date), thn_reg (registration complaints year), nim_input (student ID), nama_input (student's name), jenis_keluhan (complaints category), and keluhan (complaints). The complaints are categorized into five categories, as shown in Table 3.

Table 3. Complaint categories

Number	Categories
1	Informasi Paket BAC dan Kendala Pengirimannya
2	Kendala Isi Paket BAC
3	Tracking Status dan Data Tujuan Penerima
4	Kendala Akun dan Aplikasi BAD
5	Kendala Isi Aplikasi BAD

3.3. Data Preparation BA Related Complaints

At this stage, the data must be prepared to ensure its suitability for use in the modeling process. All steps follow the CRISP-DM framework and are implemented using Python 3.10, pandas 1.5, and scikit-learn 1.5. Data preparation steps include:

3.3.1. Data Validation

In the first stage, the complaint data obtained and categorized by the CS team consisted of 15 categories. After analysis, these 15 categories showed overlapping meanings between some labels, such as "delivery problem" and "late delivery." To prevent misclassification and redundancy, these categories were combined into five categories representing different responsible divisions (see Table 3). This merger ensured balanced class representation and reduced potential bias during model training.

3.3.2. Data Cleansing

After passing the data validation process, the data will undergo a cleaning phase to maintain good quality and be suitable for analysis. To be usable, the “jenis_keluhan,” which contains the complaints category, and the “keluhan,” which contains the complaint details, must be cleaned of incomplete or missing data. Text cleaning was performed using Python version 3.10 and pandas version 1.5. In addition, this study implemented several rules, such as:

- Conversion of all text to lowercase.
- Removal of URLs, emails, hashtags, punctuation, non-alphanumeric symbols, emoji, and non-standard Unicode characters.
- Compression of multiple spaces into a single space.

Additionally, custom stopwords are applied to eliminate non-informative words and reduce noise in the complaint text.

3.3.3. Encoding

After the data cleaning process is complete, the next step is encoding using LabelEncoder. This step converts the values in the "jenis_keluhan" (complaints category) into a numerical form so the machine learning algorithm can use them for training and testing.

3.3.4. Tokenization

Following the encoding process, the next step is tokenization. This process is essential to prepare the data for further analysis. Tokenization is applied to the “keluhan” (complaint detail) column, where the text written by students is broken down or parsed into smaller units, such as words or subwords (tokens), to enable more effective processing by Natural Language Processing (NLP) techniques. IndoBERT was chosen because it is trained on 220 million Indonesian tokens and performs reliably in public NLP benchmarks. The process rules included:

- WordPiece-based IndoBERT tokenizer with max_length = 512, padding = True, and truncation = True.
- Output type 32-bit floating-point dense vectors (float32) stored in NumPy arrays for compatibility with classical machine-learning algorithms.
- Compression of repeated blank spaces into a single space.

3.3.5. Split Data

In the final stage of data preparation, the dataset undergoes a data splitting process. The dataset, which consists of the “jenis_keluhan” (complaint category) and “keluhan” (complaint content) columns, is divided into two parts: 80% for training data and 20% for testing data. The training data is used to train the model, allowing it to learn the relationship between the complaint texts and their corresponding categories in order to classify new, unseen data.

The testing data is then used to evaluate the model's performance by comparing the predicted categories to the actual labels, based on the model trained using the training set. Standardization applied via *StandardScaler* to adjust feature variance before input to SVM; this step improves convergence stability, and SVM models were tuned with class_weight {None, 'balanced'} to compensate for minor label imbalance observed in the dataset.

3.4. Embedding Using IndoBERT

At this stage, embedding is used to turn the tokens created by the tokenizer into vector forms that the model can work with. This study used IndoBERT for embedding, which is a pretrained language model made for the Indonesian language.

3.5. Modelling Using Neural Network (NN), Naïve Bayes (NB), and Support Vector Machine (SVM)

At this stage, the complaint categorization model is developed using three machine learning algorithms: NN, NB, and SVM, implemented in Visual Studio Code. The available

dataset is divided into two subsets: the training and testing data. All model is trained using the data for training, and the resulting trained model is then applied to the testing data to evaluate its performance.

3.6. Performance Evaluation Using Accuracy, Recall, Precision, and F1 Score

The categorization results generated by the three models using the testing data were evaluated using accuracy, recall, precision, and F1 score. The results from the evaluation were compared to see which model scored the highest in categorizing complaints related to BA at XYZ University.

4. Results and Discussion

4.1. Business Understanding BA Related Complaints Management

Complaint management of BA-related complaints begins with the submission of complaints, which are entered into the system by the designated Customer Service (CS) personnel. They are responsible for categorizing each complaint so that it can be addressed by the appropriate category handler, as illustrated in Figure 3. As the number of incoming BA-related complaints continues to increase, various problems are found, primarily due to the manual process of complaint categorization.

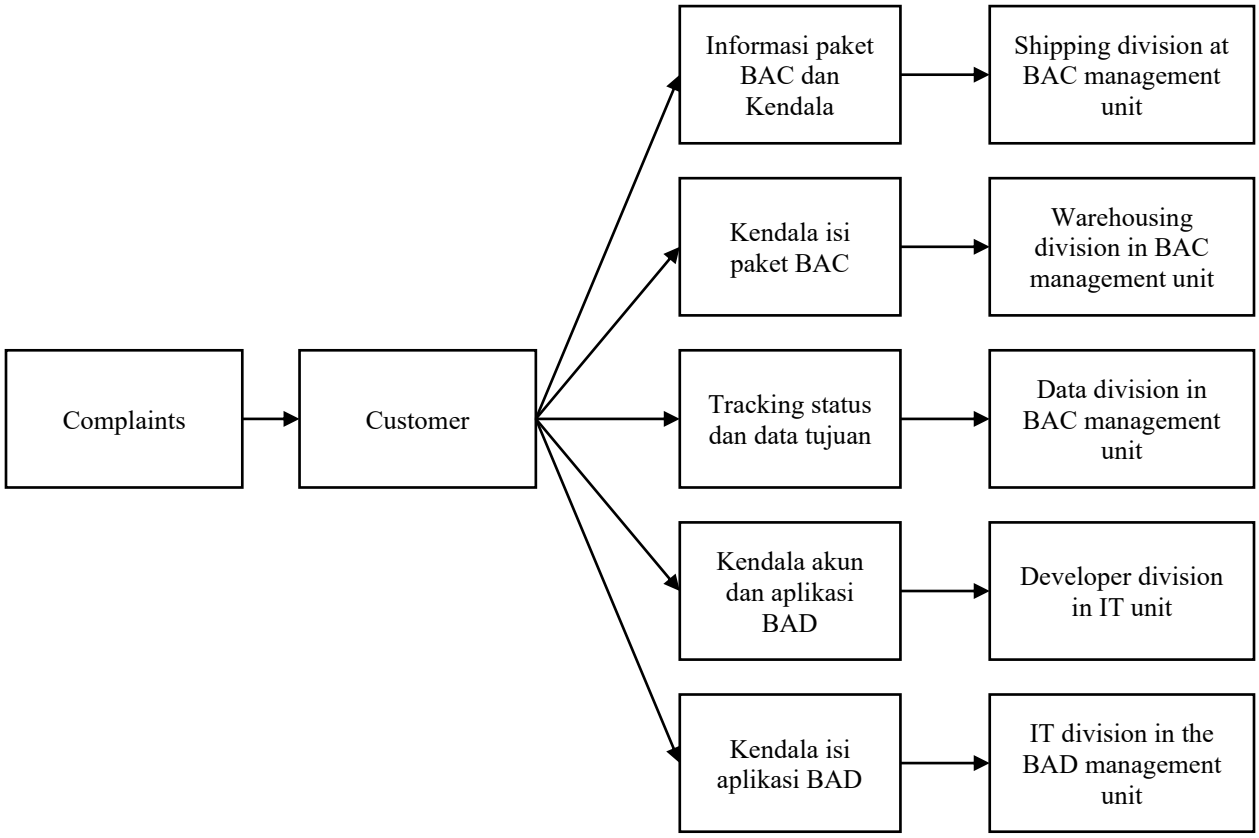

Fig. 3 The responsible division for each BA-related complaint

4.2. Data Understanding BA Related Complaints

The data used in this study is BA complaint data at XYZ University. The time range of the data to be analyzed is the complaint data that has been categorized manually from August 2023 to May 2024.

The data were retrieved from the unit database at XYZ University using HeidiSQL with the data structure comprising the following fields: id_keluhan (complaint ID), no_tiket (ticket number), tgl_input (input date), thn_reg (registration year), nim_input (student identification number), nama_input (student name), jenis_keluhan (complaint category), and keluhan (complaint details). However, for this study, only the jenis_keluhan field, which contains the BA complaint category, and the keluhan field, which contains the details of the complaints from students, are utilized. The total number of complaint records and their categories, used in this study, amounted to 8,152 records.

4.3. Data Preparation BA Related Complaints

4.3.1. Data Validation

In the initial stage, the original complaint data that has been obtained and categorized by the CS team includes 15 categories. However, because there are similarities between categories, the category will be biased during model learning, and the category also has the same party handling it; therefore, these categories were consolidated and optimized into 5 main categories (See Table 2).

4.3.2. Data Cleansing

The collected data were cleaned with the rules described in Section 3.3.2, resulting in 8,152 valid complaints. Figure 4 shows examples of custom stopwords eliminated from the dataset. The cleansing process preserved important numeric identifiers (e.g., shipment codes and student codes) that contribute to contextual interpretation by the embedding model. These steps were performed to ensure that the data were suitable for use.

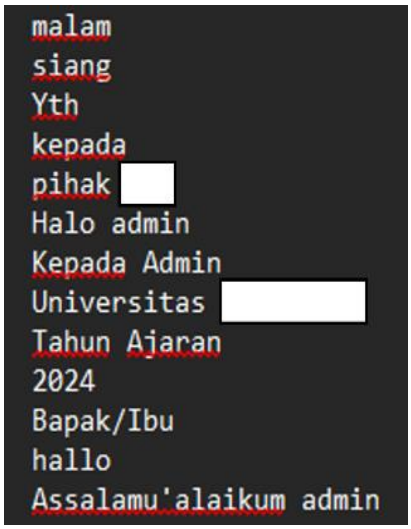

Fig. 4 Examples of words used in stopwords

4.3.3. Encoding

At this stage, encoding is applied to transform the data in the "jenis_keluhan" column into numerical labels (0 to 4) that can be utilized by a machine learning algorithm to ensure consistent label usage throughout all algorithms.

4.3.4. Tokenization

Tokenization using IndoBERT produced subword segments. The data in the "keluhan" column, which contains detailed descriptions of student complaints, is broken down or parsed into smaller word units, such as words, subwords, and made into a numeric token format. Each token sequence was encoded to capture contextual dependencies between words.

4.3.5. Split Data

The dataset, consisting of the "jenis_keluhan" (complaint category) and "keluhan" (complaint details) columns, was split into 80% (6,521) training and 20% (1,631) testing. This balance ensures unbiased model evaluation and reliable generalization performance for all models.

4.4. Embedding Using IndoBERT

This stage converts the tokens generated by the tokenizer into vectorial types that can be used by the model. One of the existing monolingual pretrained models for Indonesian is IndoBERT. IndoBERT has been pretrained for Indonesian with a total of more than 220 million words [13]. Embedding uses the IndoBERT Large P2 model.

The use of IndoBERT was selected because it is more suitable for processing complaint sentences written in Indonesian. The Large model was chosen due to its larger vocabulary and better performance compared to the Base model, while version P2 was selected because it offers improvements over P1. This decision aligns with the findings of another study that demonstrated that the IndoBERT Large model achieved the best evaluation results compared to the Base and Lite models. As this model has a greater number of parameters than the others, it is expected to capture and represent the data more accurately [12].

4.5. Modelling Using Neural Network, Naïve Bayes, and Support Vector Machine

At this stage, a complaint categorization model is developed using three machine learning algorithms: NN, NB, and SVM. Data that has been divided into training data is entered into each algorithm model to be trained for each model.

4.5.1. Neural Network (NN)

Neural networks are nonlinear regression techniques inspired by how the brain works. Similar to partial least squares, the results are modeled by a set of unseen intermediate variables called hidden variables. These hidden variables are linear combinations of some or all predictor variables and are not estimated hierarchically [14]. Several combinations of hyperparameter customization can improve the accuracy of this model.

The hyperparameters used include hidden layers, ReLU activation, dropout, batch normalization for all layers, softmax on the output layer, learning rate, CrossEntropyLoss as the loss function, Adam optimizer, and 200 epochs. The combination of hyperparameters used for this model can be seen in Table 4.

Table 4. NN hyperparameter combination for testing

Combination number	Layer 1	Layer 2	Layer 3	Layer 4	Dropout	Learning rate
1	512	256	128	64	0.4	0.0005
2	512	256	128	64	0.5	0.0005
3	512	256	128	64	0.4	0.00005
4	512	256	128	-	0.4	0.0005
5	512	256	128	-	0.4	0.001
6	512	256	64	-	0.4	0.001

4.5.2. Naïve Bayes (NB)

Naive Bayes is one of the supervised classification models whose construction is very easy. The model only requires a set of objects, each of which belongs to a known class, and each has a known vector of variables. This study aims to create rules that allow us to assign future objects to a class [15]. In this study, the Gaussian Naive Bayes model was applied, using the training data previously split during the earlier phase.

4.5.3. Support Vector Machine (SVM)

SVM is a discriminative classification method that is generally recognized for its high accuracy. SVM classifiers work by optimally partitioning the data space into different classes [16].

This study uses several combinations of hyperparameter customization to improve the accuracy of this model, including C, kernel, gamma, class_weight, maximum iterations, error tolerance, and standard scaler across all combinations.

In the initial stage, this study uses a grid search function to find the highest evaluation value of a hyperparameter combination. After getting the combination of grid search, this study also compared the results of the grid search combination placed on combination number 1 with other hyperparameter combinations that will be tested to get maximum results, as seen in Table 5.

4.6. Evaluation Results and Performance Comparison: Neural Network, Naïve Bayes, and Support Vector Machine

This section presents the evaluation results of each ML performance comparison. For the NN model, the evaluation was conducted for each combination of hyperparameters, with the results shown in Table 6. Based on the evaluation shown in Table 6, hyperparameter combination number 1 achieved the highest and most optimal performance among all combinations, with evaluation values of accuracy 0.9196, precision 0.9200, recall 0.9196, and F1 score 0.9196. The confusion matrix generated during the evaluation of the NN model, which provides deeper insight into the classification performance, is shown in Figure 5.

Table 5. Combination of SVM hyperparameters for testing

Combination number	C	Class_weight	Kernel	Gamma	Max_iter	error tolerance	Standard scaler
1	10	none	rbf	scale	1000	0.0001	Ya
2	100	none	rbf	scale	1000	0.0001	Ya
3	10	none	rbf	auto	5000	0.00001	Ya
4	10	balanced	rbf	scale	1000	0.00001	Ya
5	10	none	rbf	auto	1000	0.00001	Ya
6	10	none	linear	scale	5000	0.00001	Ya

Table 6. Evaluation results of each combination of NN hyperparameter models

Combination number	Accuracy	Precision	Recall	F1 Score
1	0.9196	0.9200	0.9196	0.9196
2	0.9166	0.9166	0.9166	0.9164
3	0.8730	0.8764	0.8736	0.8733
4	0.9117	0.9112	0.9117	0.9113
5	0.9123	0.9130	0.9123	0.9140
6	0.9110	0.9123	0.9110	0.9112

Based on that confusion matrix, categories 1 (Informasi Paket BAC dan Kendala Pengirimannya) and 5 (Tracking Status dan Data Tujuan Penerima) show a notable number of misclassifications compared to the other categories, with over 30 complaints misclassified in each direction. This indicates that the model has difficulty distinguishing between these two categories. This could be due to the high similarity in the textual content and phrases used in the complaints within these categories. Both categories may involve issues related to delivery or delivery status, which could result in overlapping vocabulary and sentence structure, making it difficult for the model to learn distinct features for each class.

For the NB model, the evaluation results showed an accuracy score of 0.7486, a precision score of 0.7601, a recall score of 0.7486, and an F1 score of 0.7445 (see Table 7). The confusion matrix generated during the evaluation of this NB model is shown in Figure 6. From the confusion matrix conducted, the NB model exhibits significant misclassification errors not only between category 1 (Informasi Paket BAC dan Kendala Pengirimannya) and 5 (Tracking Status dan Data Tujuan Penerima), but between categories 2 (Kendala Akun dan Aplikasi BAD) and 3 (Kendala isi aplikasi BAD), categories 3 and 4 (Kendala Isi Paket BAC), and between category 1 and 4. In several of these category pairs, the number of misclassified complaints exceeds 30 in both directions. This indicates that the model struggles to clearly distinguish between most complaint categories. These findings suggest that the Naive Bayes model may not be suitable for the task of categorizing BA-related complaints, particularly given the complexity of the language and expressions used by students.

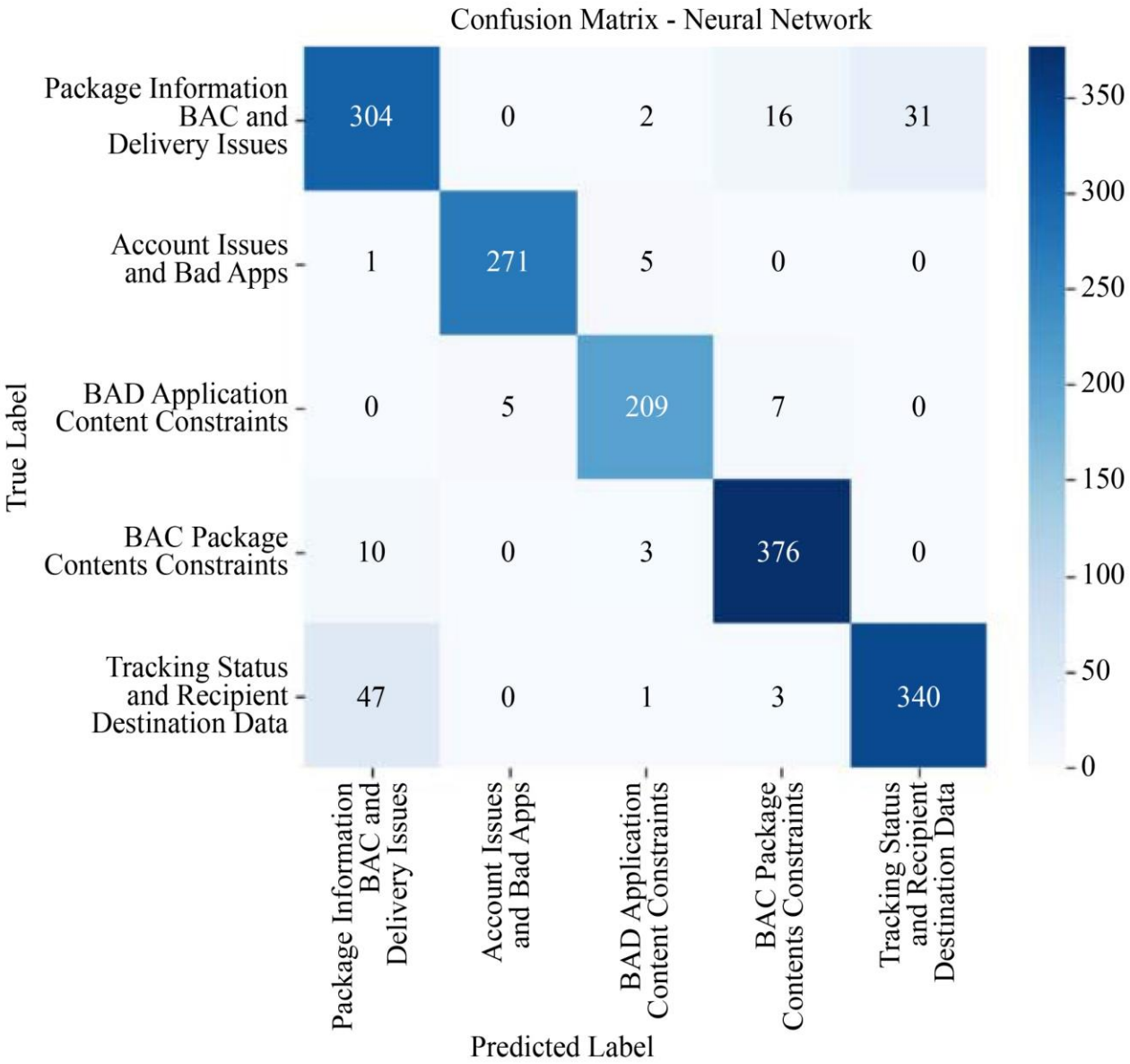

Fig. 5 Confusion matrix of NN model

Table 7. NB model evaluation results

Accuracy	Precision	Recall	F1 Score
0.7486	0.7601	0.7486	0.7445

For the SVM model, a grid search was conducted to identify the optimal combination of hyperparameters. The best performing combination from the grid search included: C = 10, class_weight = none, gamma = scale, kernel = rbf, maximum iterations = 1000, standard scaler applied, and error tolerance set to 0.0001 (1e-4). Following the grid search, this study also compared with other hyperparameter combinations. The best evaluation results for the SVM model were achieved with combination number 3. The evaluation of this combination resulted in Table 8, an accuracy score of 0.9104, a precision of 0.9121, a recall of 0.9104, and an F1 score of 0.9109. The hyperparameters used for this configuration were C = 10, class_weight = none, gamma = auto, kernel = rbf, maximum iterations = 5000, application of standard scaler, and an error tolerance of 0.00001 (1e-5), from the confusion matrix in Figure 7 generated for the evaluation SVM model.

Based on the confusion matrix, the SVM model shows misclassifications exceeding 30 instances in both directions between category 1 (Informasi Paket BAC dan Kendala Pengirimannya) and 5 (Tracking Status dan Data Tujuan Penerima), which indicates significant confusion. This likely occurred because the issues discussed in both categories were similar, namely, related to the delivery conditions of goods.

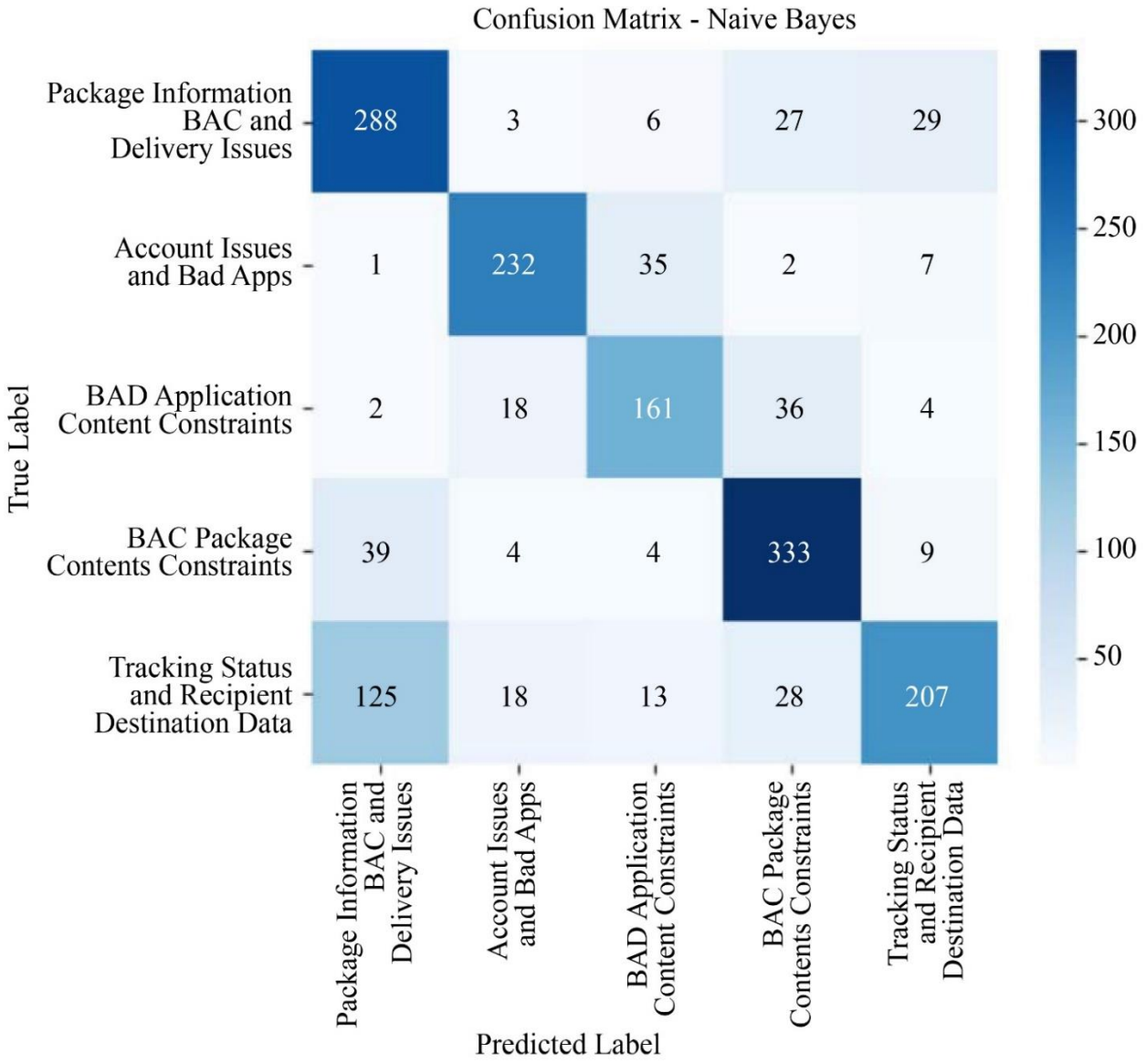

Fig. 6 Confusion matrix of NB model

Furthermore, there were numerous misclassifications between categories 1 (Informasi Paket BAC dan Kendala Pengirimannya) and 4 ("Kendala Isi Paket"), which relate to student complaints related to BAC. This is due to the fact that both categories involve the same BAC complaints, resulting in similar wording and sentence structure. This similarity can make it difficult for the model to distinguish between the different contexts of the two categories. Based on the overall evaluation results, the Artificial Neural Network (ANN) model is highly suitable for use in text mining to classify student complaints related to BAC, especially in diverse datasets with imbalances between categories. This finding aligns with [17], which highlights that NN have several advantages, such as high accuracy in modeling complex systems, the ability to process large amounts of data, the ability to capture nonlinear relationships, the ability to automatically extract features, and the ability to handle imbalanced data. However, this model also has several disadvantages, such as being sensitive to irrelevant or noisy input, being susceptible to overfitting, and the risk of performance degradation if trained improperly or for too long, or over-training. A similar conclusion was also expressed at [18], who applied text mining to classify whether sentences contain sarcasm in user comments on platforms such as YouTube, Facebook, and Blogs. In the study, the ANN model achieved the highest evaluation score. This indicates its ability to capture complex text patterns, with the score for accuracy 92.29%, precision 92.27%, and recall 92.29%.

Table 8. SVM model evaluation results for each of the best hyperparameter combinations

Combination number	Accuracy	Precision	Recall	F1-Score
1	0.9098	0.9107	0.9098	0.9101
2	0.9068	0.9084	0.9068	0.9072
3	0.9104	0.9120	0.9104	0.9108
4	0.9098	0.9111	0.9098	0.9102
5	0.9104	0.9112	0.9104	0.9107
6	0.8448	0.8491	0.8448	0.8471

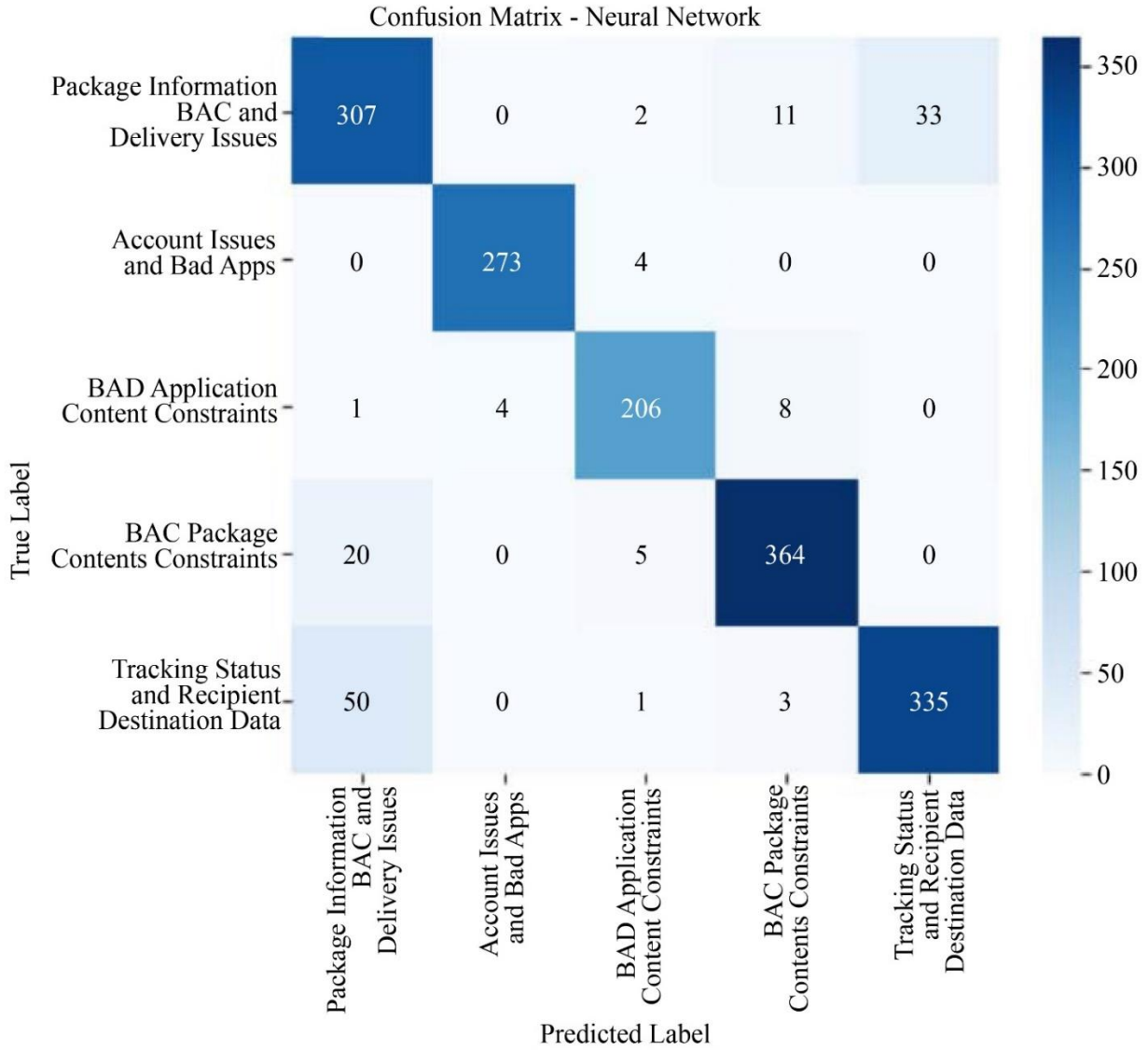

Fig. 7 Confusion matrix of the SVM model

4.7. Comprehensive Error Analysis

4.7.1. Class-Wise Performance

Regarding model performance by class, evaluation was conducted based on the complaint category to compare model performance in detail within each category. According to [16], standard metrics used in text mining evaluation are precision, recall, and F1 score, which are directly derived from the confusion matrix. These three metrics measure how well the model is able to recognize relevant text (recall) and how accurately it classifies that text (precision). The F1 score, the harmonic mean of precision and recall, is used in text classification to reduce errors caused by class imbalance. The NN and SVM models performed very well across all five complaint categories, with macro F1 scores of 0.926 and 0.918, respectively (see Table 9). Model NB performed significantly lower, with a macro F1 score of 0.801. This

indicates that the model's ability to understand contextual relationships between sentences is still limited. Evaluation by category showed that the NN and SVM models scored highest in the "*Kendala Akun dan Aplikasi BAD*" category, with F1 scores of 0.988 and 0.986. These categories are clearly defined, making them easier to distinguish.

Table 9. Class-wise performance

Complaint Category	Precision (NN)	Recall (NN)	F1-Score (NN)	Precision (SVM)	Recall (SVM)	F1-Score (SVM)	Precision (NB)	Recall (NB)	F1-Score (NB)
Informasi Paket BAC & Kendala Pengirimannya	0.840	0.861	0.851	0.851	0.872	0.861	0.762	0.808	0.784
Kendala Akun & Aplikasi BAD	0.996	0.981	0.988	0.985	0.987	0.986	0.862	0.816	0.839
Kendala Isi Aplikasi BAD	0.954	0.957	0.955	0.939	0.940	0.939	0.811	0.775	0.793
Kendala Isi Paket BAC	0.951	0.974	0.962	0.942	0.950	0.946	0.888	0.904	0.896
Tracking Status dan Data Tujuan Penerima	0.884	0.867	0.875	0.871	0.862	0.866	0.687	0.700	0.693
Macro Average	0.925	0.928	0.926	0.918	0.922	0.918	0.802	0.801	0.801

Lower F1 scores were found in the "Informasi Paket BAC & Kendala Pengirimannya" and "tracking Status" categories, ranging from 0.85 to 0.87 for both models. The decrease in precision in these two categories occurred due to overlapping terms such as "paket", "pengiriman", "status", and "tracking", which caused their contextual representations to overlap. The use of IndoBERT embeddings in this study successfully reduced this overlap by distinguishing the context between complaints that are very similar in meaning. These results align with previous research in [12], which reported that the IndoBERT model outperformed other models in all evaluation metrics for Indonesian text classification.

Table 10. Misclassification between similar categories

Category Pair	NN	SVM	NB
1 & 5 (Informasi Paket BAC & Tracking Status)	78	83	182
1 & 4 (Informasi Paket BAC &Isi Paket BAC)	26	31	66
2 & 3 (Akun BAD & Isi Aplikasi BAD)	10	8	53

Table 10 clearly shows the reduction in error achieved by the NN and SVM models compared to the NB model. The number of bidirectional misclassifications between Category 1 and Category 5 decreased by more than 57% when using NN with 78 errors and SVM with 83 errors, compared to NB which reached 182 errors. Meanwhile, confusion between Category 1 "Informasi Paket BAC" and Category 4 "Isi Paket BAC" decreased from 66 to 26–31 cases, and between Category 2 "Akun BAD" and Category 3 "Aplikasi BAD" decreased from 53 to below 10 cases.

Hasil evaluasi menunjukkan bahwa model NN dan SVM lebih baik dalam membedakan konteks antar kategori, hal ini sangat penting dalam pengelompokan teks [14]. Hasil evaluasi dengan Recall yang tinggi dan konsisten di atas 0.86 juga menunjukkan kinerja yang baik dalam memproses kategori minoritas, serta memperkuat kemampuan model IndoBERT dalam menangkap kekayaan bahasa dari data keluhan [9, 12].

4.7.2. Error Type and McNemar Test

In this study, an error type analysis was conducted to identify the main types of misclassification errors that occurred in each model. As explained in previous research [16], the process of classifying text involves managing numerous features representing all words that occur with varying frequencies. This creates unique challenges for algorithms processing text. These factors influence the differences in results obtained between models (see Table 11). The analysis results show that both NN and SVM can significantly reduce misclassification errors compared to naive Bayesian (NB). Furthermore, based on the results in Table 11, NN successfully reduced the two-way misclassification between classes (paket) and tracking from 182 cases to 78, while SVM also showed a similar improvement to 83 cases. This indicates that models using contextual embedding, such as IndoBERT, are more effective at handling frequency variations and contextual relationships than probabilistic approaches that assume feature independence. In addition, both NN and SVM can maintain recall values above 0.86 for each complaint category, indicating very consistent performance even in low-frequency

classes. The McNemar test was used to check whether the differences between the models were statistically significant [14].

This study used the McNemar test, and the results showed that both the NN and SVM models performed significantly better than the NB model ($p < 0.001$). However, the difference between the NN and SVM models was not significant ($p = 0.003$) (see Table 12). These results indicate that the performance improvements of the NN and SVM models are significant and reliable for identifying BA-related complaints.

Table 11. Error type found across models

Error Type	Description	NN	SVM	NB	Observation
Large Feature Space	Similar keywords (paket, pengiriman, tracking) cause confusion between Categories 1 and 5.	78 cases	83 cases	182 cases	NN and SVM reduced error by >55% vs NB.
Contextual ambiguity	Sentences referring to similar entities (isi paket and isi aplikasi) are misinterpreted semantically.	26 cases	31 cases	66 cases	Embedding-based models improved contextual separation.
Varying Word Frequencies (Data Imbalance)	Minority classes, such as *Tracking Status* underrepresented, are causing recall degradation.	0.867 recall	0.862 recall	0.700 recall	NN and SVM maintained higher recall stability.

Table 12. McNemar test results

Model Compare	A Correct, B Wrong	A Wrong, B Correct	p-value
NN vs NB	251	27	<0.001
SVM vs NB	237	26	<0.001
NN vs SVM	13	1	0.003

4.8. Model Contribution to Service Efficiency and Satisfaction

Implementation of the machine learning model in this study is expected to improve analytical accuracy and operational efficiency in complaint handling at XYZ University. Previously, the CS team manually read and categorized each complaint, resulting in delayed handling of complaints, incorrect decisions, and frequent misclassifications. The IndoBERT - based automated classification model eliminates manual steps for the CS team, thereby speeding up the handling process and ensuring more accurate category selection. Previous studies on higher education service management have shown that when student complaints are not handled quickly or fairly, it can result in decreased satisfaction and increased frustration levels, which in turn reduces the quality of institutional support for those students [19, 20]. By eliminating manual intervention in the CS team, the automated classification process contributes to shorter response times and more consistent complaint handling outcomes, improving perceptions of fairness for all students and improving efficiency in complaint handling.

Empirical findings from previous studies find the relation between procedural efficiency and satisfaction. According to [20], following fair procedures has a strong positive influence on satisfaction ($\beta = 0.936$, $p < 0.001$), and this satisfaction can strongly increase loyalty ($\beta = 0.999$, $p < 0.001$). In the study, the procedural justice referred to is related to timeliness, accessibility, and fairness in handling complaints. According to [21] in his research, the path coefficient is standardized between -1 and +1; values closer to +1 indicate a stronger positive relationship, while coefficient values closer to zero represent a weaker relationship. Based on this explanation, the data coefficient values reported in the study [20] indicate a strong positive relationship between satisfaction and loyalty. Likewise, in the study [19], it was found that service responsiveness is the strongest indicator of satisfaction ($\beta = 0.67$, $p < 0.001$), but in terms of administrative handling accuracy, it does not directly increase loyalty ($\beta = 0.55$, $p < 0.01$). These results confirm that speed and accuracy in complaint handling are key factors that can be addressed to increase satisfaction and loyalty in higher education.

Ineffective complaint handling can undoubtedly reduce satisfaction and undermine student trust, which directly impacts the institution's reputation [20]. By providing faster and more accurate complaint categorization, this proposed model mitigates the risks identified at XYZ University, thereby strengthening trust in the institution's responsiveness. Therefore, this approach should not simply be viewed as an optimization of complaint handling techniques, but as a comprehensive strategic improvement in service quality and the institution’s reputation.

5. Conclusion

Based on the final evaluation results of all models listed in Table 13, the NN model performed best compared to the NB and SVM models. The NN model with the best parameter combination achieved an accuracy evaluation score of 0.9196, precision of 0.9200, recall of 0.9196, and F1 score of 0.9196.

Table 13. Final evaluation of three algorithms

Model	Accuracy	Precision	Recall	F1 Score
NN	0.9196	0.9200	0.9196	0.9196
NB	0.7486	0.7601	0.7486	0.7445
SVM	0.9104	0.9120	0.9104	0.9108

The model uses four hidden layers with the number of neurons being 512, 256, 128, and 64, respectively. Each hidden layer uses a dropout method of 0.4 and batch normalization. The learning rate used is 0.0005, activation in all layers uses ReLU, the output layer uses softmax activation, the loss function used is CrossEntropyLoss, the optimization method used is Adam, and the epoch is 200. The SVM model also showed quite good results in second place, with evaluation values of accuracy 0.9104, precision 0.9120, recall 0.9104, and F1 score 0.9108. Meanwhile, the NB model shows the lowest results, with an accuracy evaluation result of 0.7486 and an F1 score of 0.7445, indicating that this model is less able to handle the complexity of language in complaint data. Because many students used the same words when writing complaints, there was overlap and ambiguity in each complaint. Furthermore, the use of similar words made it difficult for the model to distinguish complaints from different categories, which could also influence bias in the clustering process. However, based on the evaluation results, the Naïve Bayes model was not the right choice for classifying complaints related to BA.

This is because students at XYZ University used soft expressions and complex language patterns in conveying complaints. Error analysis was performed for each category, accompanied by statistical validation using the McNemar test. The results showed that the performance difference between the NN and SVM models was not statistically significant ($p > 0.05$), but both models were significantly better than the NB model ($p < 0.05$). These results indicate that IndoBERT embeddings are able to effectively capture the meaning of complaint texts and provide a strong foundation for classification tasks, especially on Indonesian data with complex contexts. Based on the evaluation results, using this model can help and make it easier for XYZ University to manage its BA complaints. Automatic categorization of complaints can speed up the handling process, reduce errors in categorizing complaints, and improve the institution's ability to respond more effectively to incoming complaints.

These improvements align with evidence that efficiency and accuracy in managing complaints contribute to increased student satisfaction and trust. Future studies may explore multi-label classification, real-time deployment within institutional systems to evaluate the impact of the efficiency of Service Level Agreement (SLA), and fine-tuning of IndoBERT for multilingual complaint datasets to strengthen model adaptability and application scope.

Data Availability

Data supporting this study are available at https://www.kaggle.com/datasets/hanson12321/categorizing-complaints. It is an open dataset and available to the public for the purpose of ensuring transparency and reproducibility. All complaint data were anonymized prior to analysis. Identifiers such as student names, student IDs, and contact information were removed in accordance with institutional ethical standards and applicable data protection regulations.

Credit Authorship Contribution Statement

Hanson Geraldi Pardede: Conceptualizing the idea, choosing the methodology, using software tools for detailed analysis, conducting research, getting necessary resources for this research, organizing the data, and writing the original draft.

Tuga Mauritsius: Writing review for this study& Editing, supervision for data analysis, and validation.

References

[1] Haris Ahmad Gozali, Mochamad Alfan Rosid, and Sumarno, "Classification of Student Complaints with the Naïve Bayes and Literature Methods," *Journal of Informatics, Network, and Computer Science*, vol. 3, no. 1, pp. 22-26, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[2] Muchamad Taufiq Anwar, Anggy Eka Pratiwi, and Khadijah Febriana Rukhmanti Udhayana, "Automatic Complaints Categorization using Random Forest and Gradient Boosting," *Advance Sustainable Science, Engineering and Technology (ASSET)*, vol. 3, no. 1, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] Nuzulul Khairu Nissa, and Evi Yulianti, "Multi-label Text Classification of Indonesian Customer Reviews using Bidirectional Encoder Representations from Transformers Language Model," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 5, pp. 5641-5652, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Yessy Asri et al., "Sentiment Analysis Based on Indonesian Language Lexicon and IndoBERT on User Reviews of PLN Mobile Application," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 38, no. 1, pp. 677-688, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[5] S.L. Ting, W.H. Ip., and Albert H.C. Tsang, "Is Naïve Bayes a Good Classifier for Document Classification," *International Journal of Software Engineering and Its Applications*, vol. 5, no 3, pp. 37-46, 2011. [Google Scholar] [Publisher Link]

[6] Isti Surjandari et al., "Application of Text Mining for Classification of Textual Reports: A Study of Indonesia's National Complaint Handling System," *Proceedings of the 2016 International Conference on Industrial Engineering and Operations Management*, Kuala Lumpur, Malaysia, pp 1147-1156, 2016. [Google Scholar] [Publisher Link]

[7] Manzhu Yu et al., "Deep Learning for Real-Time Social Media Text Classification for Situation Awareness using Hurricanes Sandy, Harvey, and Irma as Case Studies," *International Journal of Digital Earth*, vol. 12, no. 11, pp. 1230-1247, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[8] Bin Ning, Wu Junwei, and Hu Feng, "Spam Message Classification Based on the Naïve Bayes Classification Algorithm," *IAENG International Journal of Computer Science,* vol. 46, no. 1, pp. 46-53, 2019. [Google Scholar] [Publisher Link]

[9] Rio Wirawan, Erly Krisnanik, and Artika Arista, "Text Mining for News Forecasting on the Turnback Hoax Website," *International Journal on Informatics Visualization*, vol. 8, no. 1, pp. 96-106, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Aurangzeb Khan et al., "A Review of Machine Learning Algorithms for Text-Documents Classification," *Journal of Advances in Information Technology*, vol. 1, no. 1, pp. 4-20, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[11] Sindhuja Penchala et al., "Unveiling Text Mining Potential: A Comparative Analysis of Document Classification Algorithms," *EPiC Series in Computing: Proceedings of 39th International Conference on Computers and their Applications*, vol. 98, pp. 103-115, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Charan Singh Tejavath, and Tryambak Hirwarkar, "Analysis of Different Classification Algorithms for Text Data Mining," *Advances in Mathematics: Scientific Journal*, vol. 9, no. 6, pp. 3477-3485, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[13] Fajri Koto et al., "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-Trained Language Model for Indonesian NLP," *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain, pp. 757-770, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[14] Max Kuhn, and Kjell Johnson, *Applied Predictive Modeling*, 1nd ed., Springer, New York, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[15] Eric Bauer, and Ron Kohavi, "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants," *Machine Learning*, vol. 36, no. 1, pp. 105-139, 1999. [CrossRef] [Google Scholar] [Publisher Link]

[16] Charu C. Aggarwal, and Cheng Xiang Zhai, *An Introduction to Text Mining*, Mining Text Data, Springer, Boston, pp. 1-10, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[17] Elham Kariri et al., "Exploring the Advancements and Future Research Directions of Artificial Neural Networks: A Text Mining Approach," *Applied Sciences*, vol. 13, no. 5, pp. 1-18, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[18] Sayeda Muntaha Ferdous et al., "Sentiment Analysis in the Transformative Era of Machine Learning: A Comprehensive Review," *Statistics, Optimization and Information Computing*, vol. 13, no. 1, pp. 331-346, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Ling Lin et al., "Let's Make It Better: An Updated Model Interpreting International Student Satisfaction in China Based on a PLS-SEM Approach," *PLoS ONE*, vol. 15, no. 11, pp. 1-13, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[20] Idris Muh, Abidin Munirul, and Willya Evra, "Justice in Handling Complaints and Its Impact on Satisfaction and Loyalty in Higher Education," *Perspectives of Science and Education*, vol. 61, no. 1, pp. 24-39, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[21] Joseph F. Hair et al., *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*, SAGE Publications, 2021. [Google Scholar] [Publisher Link]

[22] Hermanto Hermanto, Ali Mustopa, and Antonius Yadi Kuntoro, "Algoritma Klasifikasi Naive Bayes Dan Support Vector Machine Dalam Layanan Komplain Mahasiswa," *JITK (Journal of Computer Science and Technology)*, vol. 5, no. 2, pp. 211-220, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[23] Levent Çallı, and Fatih Çallı, "Understanding Airline Passengers during COVID-19 Outbreak to Improve Service Quality: Topic Modeling Approach to Complaints with Latent Dirichlet Allocation Algorithm," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2677, no. 4, pp. 656-673, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[24] Jordana Bazzan et al., "An Information Management Model for Addressing Residents' Complaints through Artificial Intelligence Techniques," *Buildings*, vol. 13, no. 3, pp. 1-22, 2023. [CrossRef] [Google Scholar] [Publisher Link]