Combining Feature Embedding Based on the BERT Architecture and Random Forest Algorithm to Identify the Handling Section and Risk Priority Level in an ISP Company in Indonesia

Rhemzy Putra Maulana[1] and Yulianto[2*]
[1,2]Computer Science Department, School of Computer Science, Bina Nusantara University
Jakarta, Indonesia 11480
Email: [1]rhemzy.maulana@binus.ac.id, [2]yulianto003@binus.ac.id

***Abstract*—Revision Presently, Internet usage in society has grown considerably. It presents significant opportunities for companies to develop business related to internet service providers (ISPs). In Indonesia, both domestic and foreign ISPs are competing. To survive and remain popular with Indonesian customers, ISPs need to improve their products and services quickly. Hence, companies need to collect user feedback through customer loyalty applications and reviews to improve their service or technology. However, manually auditing thousands of textual reviews is labor-intensive and subject to human error, making it highly inefficient for agile, real-time corporate decision-making. The research aims to address this operational bottleneck by improving the automation process and providing decision-making information to ISP management. More specifically, the research contributes to implementing and comparing two transfer learning models: Bidirectional Encoder Representations from Transformer (BERT) and IndoBERT for tokenization and embedding. The embedding results are then fed into a random forest for classification. The dataset is collected via a scraping process from the Play Store application, focusing on user feedback and ratings. A total of 1,192 records of user feedback reviews from ISP loyalty apps on the Play Store in March 2024 were gathered. These are manually labeled by a general manager from the company into binary and multi-class categories. Splitting the dataset into 70% for training and 30% for testing achieves a classification accuracy of 70% for the problem domain with IndoBERT + Random Forest. However, multi-class classification for categorizing user feedback into a priority score model achieves an accuracy of 58%.**

***Index Terms*—Sentiment Analysis, Bidirectional Encoder Representations from Transformers (BERT), Indonesian Bidirectional Representations from Transformers (IndoBERT), Random Forest, Natural Language Processing**

Received: Feb. 06, 2025; received in revised form: Aug. 21, 2025; accepted: Sep. 22, 2025; available online: July 27, 2026.
*Corresponding Author

I. INTRODUCTION

IN the current global era, all human activities require the Internet to connect with one another. It happens not only between humans but also with machines, especially in Indonesia [1]. This phenomenon is also seen as a major opportunity for Internet Service Provider (ISP) companies in Indonesia to grow continuously and to compete with others [2, 3]. According to a survey conducted by the Indonesia Internet Service Provider Association, many Internet service providers compete to offer services at good prices and high quality in Indonesia [3, 4]. As a result, maintaining and increasing customer loyalty is of great importance to ISP business owners [5].

ISPs often rely on loyalty programs to retain customers and build long-term relationships, one way of which is through customer loyalty applications [6]. Currently, several telecommunications operators that also serve as ISPs in Indonesia still maintain a competitive presence. They include Telkomsel, XL Axiata, Indosat Ooredoo, and others. Examples of their loyalty applications are MyTelkomsel, MyXL, Bima+, MySmartfren, and myIM3 [2]. With the rapid development of social media and technology, users of Internet services can easily share and write about their experiences, both positive and negative. They usually write their reviews on TikTok, Instagram, Twitter, Google Play Store, and other platforms. These customer loyalty applications are generally available on the Google Play Store and can be used by users to write reviews and by companies to gather user evaluations [7].

User reviews are usually in the form of opinions about the application's user interface, mobile network speed, network signal stability, network plan prices, and other review factors related to the provision of input for or evaluation of the customer loyalty application [8]. In the customer loyalty application, users can rate reviews from 1 to 5, with 1 as the lowest and 5 as the highest [9]. Play Store users can also submit opinions or reviews about the ISP's services [7]. Hence, ISPs need to maintain the quality of their network and customer loyalty applications to increase positive perceptions and maintain or increase user retention [10]. As user reviews are an important factor in supporting their revenues, ISPs must follow up on user feedback, especially negative reviews. They can be used as materials for evaluating recurring incidents or similar problems [11].

Risk assessment is particularly important in operational activities [12]. Potential risks that may occur or have occurred in the past require further investigation or review to mitigate them if they occur and to prevent them from occurring in the future [12]. For example, users have left negative reviews on the Google Play Store due to being unable to register for the customer loyalty application. If the problem is left unfixed, it will create a negative perception and impact the company's branding and revenue. Companies should collect user reviews to improve user satisfaction with products or services [13]. However, for ISPs with a wide range of products, conducting manual risk assessments requires a great deal of time and human resources. As a result, the potential for human error is very high if the risk assessment method still relies on manual, individual evaluation.

The use of machine learning can be a solution for products with large user bases. Existing reviews, user complaints, and experiences can be summarized. It can result in data that can be grouped and classified based on predetermined parameters. This preventative measure avoids disruptions or complaints that can overwhelm the application. The use of machine learning for risk classification is necessary for companies with products that grow very dynamically. The potential risk category from the summary of existing data can assist in taking preventive action at an early stage. When the risk actually occurs, the action or solution is taken. Thus, operations are not disrupted because the risk has been predicted earlier using machine learning [14].

Several natural language processing tasks based on Bidirectional Encoder Representations from Transformers (BERT), such as DistilBERT, mBERT-aux-NLIB, mBERT-aux-NLIM, IndoBERTW-CNN, and Word2Vec-CNN-XGBoost, are tested for sentiment classification to identify whether user reviews are negative, neutral, or positive. The first research has shown that the BERT architecture is outstanding in classification tasks, such as for classifying the Dbpedia14 dataset with an accuracy score of 0.9888 [15]. The second research is in the prevention of false alarms of disasters caused by the spread of fake news on the Internet, especially on Twitter. It has achieved 83% accuracy in classifying hoax news using Convolution Neural Network (CNN) Indonesia's crawled data [16].

The third research has used customer reviews of Indonesian hotels. The results show that the IndoBERT-CNN architecture achieves the highest accuracy, at 96.36% [13]. The fourth research has also proposed a combination of the BERT-Bidirectional Long Short-Term Memory - Convolution Neural Network (BiLSTM-CNN) architecture for text classification. An open dataset released in 2021 from Xinyu City, Jiangxi Province, containing data about the Innovation Application Competition, is used. The classification results show an accuracy of 0.9536 [17].

The fifth research has shown the IndoBERT architecture for text opinion classification. It trains the model using three optimizers: Root Mean Square Propagation (RMSProp), Nadam, and Adam. The emotions dataset, as the benchmark dataset, is collected from Twitter, specifically for the Indonesian language. The results show that the IndoBERT architecture trained with the Adam optimizer achieves the highest classification accuracy, at 90.21% [18]. The sixth study proposed the IndoBERT architecture to classify social media posts as pornography, hate speech, radicalism, or defamation [19]. It is developed from the problem of hate speech, defamation, and pornography on social media platforms, which can have negative impacts on society. The data are collected using a robotic automation process that automatically captures user comments from Instagram, Twitter, and Kaskus. As a result, the proposed model can classify social reviews with 90% accuracy. The seventh study has also implemented the IndoBERT architecture to analyze user feedback ratings on the Google Play Store [20]. It aims to improve online medical care services through mobile applications. The dataset is scraped from the Google Play Store. The sentiment analysis of user feedback has a score of 96%.

The research aims to develop a classification model for customer feedback from ISP subscribers who have used the mobile loyalty application. The idea is to digitize the manual process. Hence, ISP companies can respond to user feedback and improve their services based on reviews of loyalty applications on the Google Play Store. The classification process involves categorizing user feedback and reviews related to service operations or technical issues. This process is then

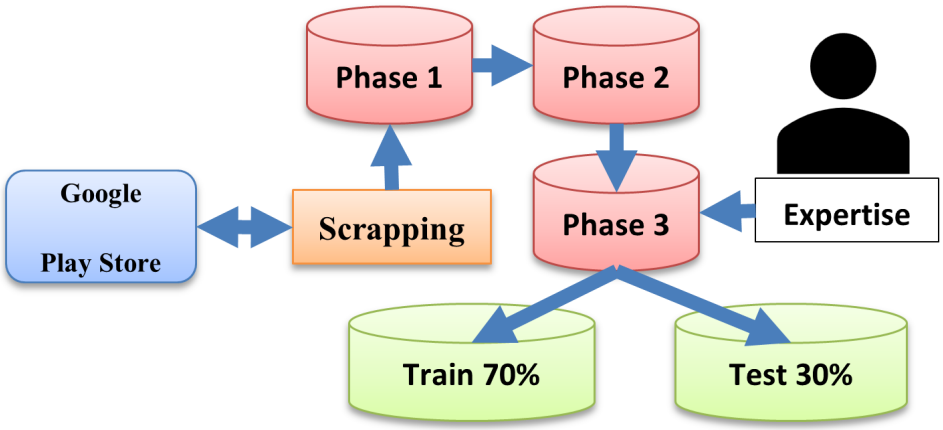

Fig. 1. The process of collecting and pre-processing the dataset.

followed by identifying the feedback category as low priority (1), medium priority (2), or high priority (3). The model uses a hybrid of BERT with a double Random Forest approach. The goal of this model is to facilitate more efficient decision-making in the management of loyalty applications.

BERT is chosen as the pre-trained model because it is the most recent method developed by Google that can improve classification performance by converting letters into vectors [21]. Additionally, the usage of BERT makes preprocessing optional because BERT can use existing letter inputs to determine the correct context for classification [22]. For further comparison, the IndoBERT transfer learning architecture is proposed, as the dataset consists primarily of Indonesian expressions. The outputs of BERT and IndoBERT are called embeddings, and the total number of outputs is 768 feature vectors. These feature vectors are then fed into a double Random Forest architecture. Random Forest is used to classify text risk labels obtained from the dataset with BERT as the pre-trained model. The Random Forest method is chosen because it can handle complex features and classify multiple labels. Random Forest has 70% data classification accuracy in previous research [23]. The two proposed Random Forests have different tasks. The first Random Forest identifies areas for improvement, whether in operational services or technical affairs. The second Random Forest architecture identifies priority levels, categorizing the user feedback review as level 1, 2, or 3.

II. Research Method

As mentioned previously, the research aims to implement the BERT + double Random Forest and IndoBERT + double Random Forest hybrid models to categorize user feedback and reviews with ratings between 1 and 3 as input containing information that needs improvement, whether for service or technical matters. The proposed model also identifies whether the user feedback reviews are categorized as priority 1 (high risk), priority 2 (medium risk), or priority 3 (low risk). The dataset comprising user feedback on Telkomsel, an ISP in Indonesia, is collected from the Google Play Store. The feedback data collected from Google Play Store generally does not have appropriate labels as required by Telkomsel. Hence, manual labeling needed to be performed by adding additional labels. The added labels are the original ratings, in relation to the risk priority level, collected from the scraping process, along with an additional label related to the risk control recommendation. The google_play_scraper Python library is used to scrape user feedback and ratings from the Play Store [24].

The destination scraper link is https://play.google.com/store/apps/developer?id=Telkomsel&hl=id. The data that were scrapped from March 4, 2024 to March 15, 2024. As shown in Fig. 1, in phase 1, user feedback reviews with ratings of higher than 3 are excluded. User reviews with ratings of 1, 2, or 3 are considered bad reviews. User reviews with a rating of 1 are considered high risk. Meanwhile, those with a rating of 2 and 3 are considered medium and low risk, respectively. However, these ratings must be verified by the relevant General Manager. As a result, the General Manager of Telkomsel is needed to participate in validating the results of the manual labeling category.

As shown in Fig. 1, the original ratings given by users are excluded. In Phase 3, the General Manager replaces the user rating score with two categorizations: priority score and problem domain. The priority score

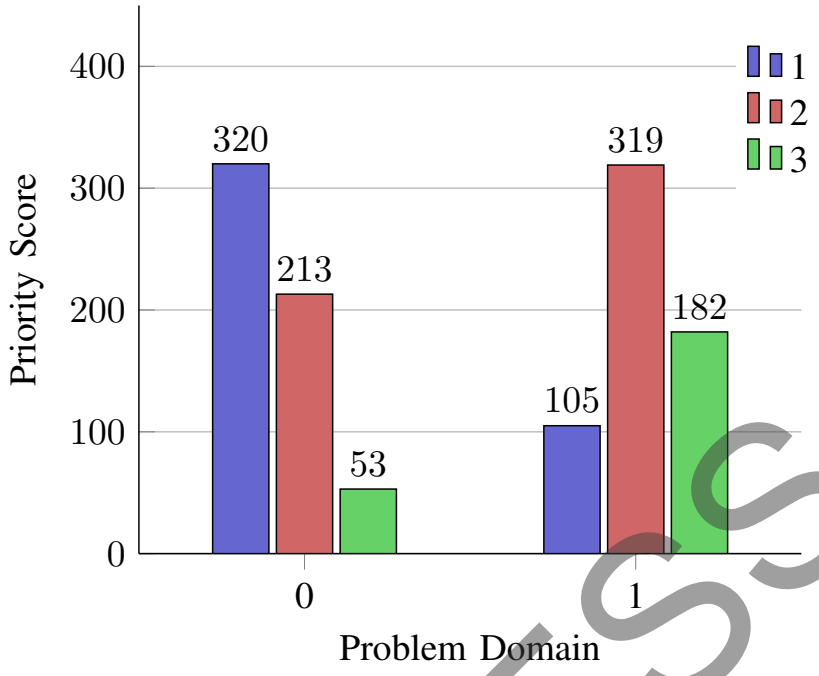

Fig. 2. Chart of the number of reviews with certain priority scores in each problem domain.

TABLE I
NUMBER OF REVIEWS IN EACH CATEGORY IN PHASE 3.

Priority Score			Problem Domain	
Score 1	Score 2	Score 3	Category 0	Category 1
425	532	235	586	606

comprises three categories. Level 1 is low priority, level 2 is medium priority, and level 3 is high priority. The General Manager assigns the levels after reading each review individually. The second categorization added by the General Manager is the problem domain, which comprises two categories. The label of 0 represents necessary service improvements, and the label of 1 represents necessary technical improvements. In Phase 3, the total number of reviews representing user feedback collected through the scraping process is 1,192.

As shown in Table I, the reviews in the dataset are all assigned to the two categories: problem domain and priority score. Figure 2 shows a visual comparison. Around 425 user reviews are categorized as a priority score of 2. Meanwhile, 235 user reviews have a priority score of 3.

It can be seen in Fig. 2 that the problem domain labeled as 0 consists of 320 reviews with a score of 1,213 reviews with a score of 2, and 53 reviews with a score of 3. Meanwhile, the problem domain labeled as 1 comprises 1,319 reviews with a score of 2 and 182 with a score of 3. Therefore, user feedback with a problem domain of 0, indicating the need for service improvement, comprises 586 reviews. Meanwhile, user feedback with a problem domain of 1, which indicates necessary technical improvement, comprises 606 reviews. The dataset in Phase 3 is then split into two parts: 70% for training the classifier and 30% for testing. The portion of the dataset used to train the classifier is fed into the Random Forest, which is the approach used in this research. This process is illustrated in Fig. 3.

Table II shows six sample records of user feedback reviews collected from loyalty apps on the Play Store with ratings of three stars or less. These reviews represent raw data about complaints that need to be handled by ISP management. The raw data contain slang, typos, and informal language typical of app reviews. There are three columns. The first column is the ‘priority score’, assigned by management. A priority score of 3 means it is important to be solved immediately, a priority score of 2 is medium, and a priority score of 1 is not too important. The third column is ‘problem domain’, which represents which division is responsible for responding to user complaints. A digit of 0 means the improvement needs to be addressed by the service department, and a digit of 1 means the technical division should address the improvement.

Cite this article as: R. P. Maulana and Yulianto, "Combining Feature Embedding Based on the BERT Architecture and Random Forest Algorithm to Identify the Handling Section and Risk Priority Level in an ISP Company in Indonesia", CommIT Journal 20(2), 207–218, 2026.

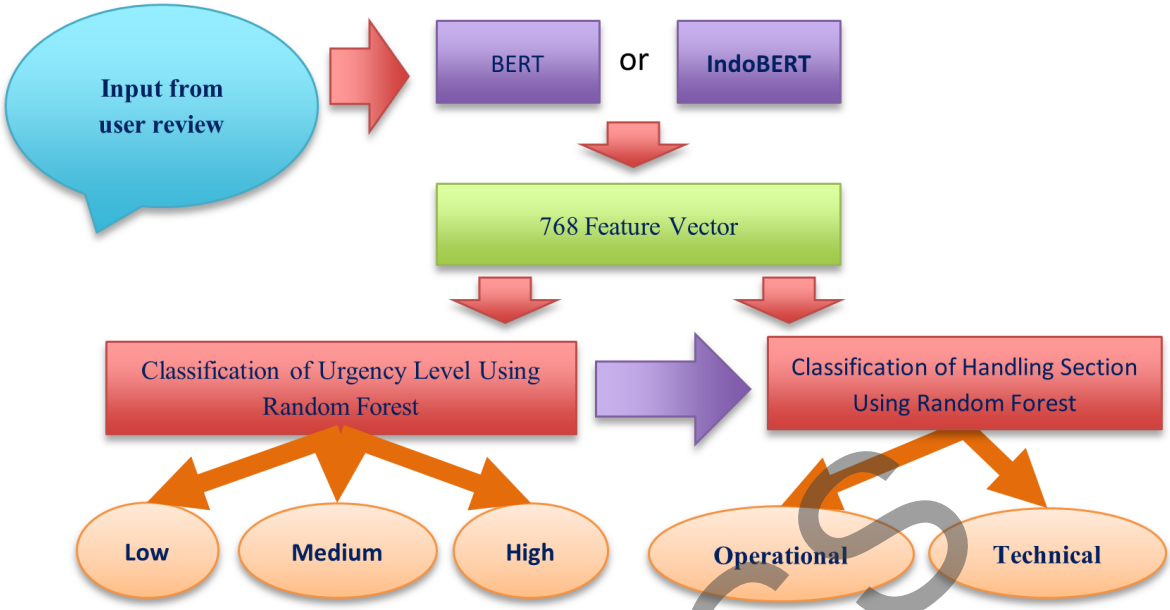

Fig. 3. Flowchart of user input processing. Note: Bidirectional Encoder Representations from Transformers (BERT) and Indonesian Bidirectional Representations from Transformers (IndoBERT).

TABLE II
EXAMPLES OF USER FEEDBACK IN INDONESIAN WITH PRIORITY SCORE AND PROBLEM DOMAIN.

User Review	Priority Score	Problem Domain
Jujur starling paling berkualitas. Apapun buatan Indonesia bersifat pemeras korupsi. Dan Belanda jajah Konoha yang korupsi Bupati Konoha selama 300 tahun. Dan kalaupun ketahuan warga konoha. Belanda nya pun di fitnah. Korupsi sudah mendarah daging. Miris	1	1
Kak, kenapa setelah update malah sering error,setiap mau bukak akun harus masukin no terus.	3	1
Update terbaru malah kacau, masa aktif paket tak bisa di cek, Kuota 17GB baru 5hri tau tau ada pemeritahuan kuota tinggal 500mb padahal jarang pakai karna sering tersambung wifi	3	0
Saya ada kendala dimana update an yang sekarang saat keluar aplikasi itu otomatis nge logout sendiri, entah memang fitur baru atau bug , jadinya saya harus ulang lagi memasukan no hp untuk membuatkan link untuk masuk ke aplikasi, dimana itu sangat merepotkan, saya harap bisa dipertimbangkan lagi metode login yg skrng, mungkin bisa dikembalikan metode login nya ke awal atau dibuatkan fitur yg lebih efisien, jika bisa buat agar supaya bs login menggunakan username/userid/email jg spya lebih mudah.	3	1
Tolong dong bro gua udah beli paket disini lumayan ini perbulan nya bisa buat bayar wifi unlimited sebulan, JARINGAN NYA JANGAN LEMOT!!! Bikin malu aje kuota doang mahal jaringan abal hadeuh	3	0
Tolong lah untuk sinyal di perbaiki yg bener, saya ini pelanggan lama kalian. 10 tahun bukan waktu yg singkat coy. Kuota ga pernah ada yg murah, sinyal makin kesini makin jelek, niat apa enggak ini kalian menyenangkan pelanggan?	2	0

Table III is the English translation of Table II. The reviews in Table III are faithfully translated, keeping their core ideas. The research uses the original Indonesia reviews for analysis.

To assist the company's managerial staff in enhancing their services by aggregating feedback from the loyalty application, the researchers present a machine learning architecture comprising a feature extractor and a classification algorithm, as illustrated in Fig. 3. The objective of this method is to provide ISPs with recommendations to improve their services, whether in service or technical matters. The implementation feeds the proposed input model with user feedback reviews, which can be integrated with the customer loyalty application's Application Programming Interface (API). In this research, 30% of the user feedback reviews are set aside, as shown in Fig. 1.

The next step involves processing the user review data with the feature extractor. Here, the two word-embedding models, BERT [25] and IndoBERT [26], are used for feature extraction and benchmarking. After carrying out the word embedding process to obtain the feature vectors, classification is then performed twice to obtain the different labels, as the classification

TABLE III
EXAMPLES OF USER FEEDBACK IN ENGLISH WITH PRIORITY SCORE AND PROBLEM DOMAIN.

User Review	Priority Score	Problem Domain
Honestly, starling [sic, Starlink] has the best quality. Anything made in Indonesia is for extortion and corruption. And the Dutch colonized Konoha, corrupting the Konoha Regent for 300 years. And even if Konoha residents found out, the Dutch were slandered. Corruption is ingrained in our blood. it's sad.	1	1
Why do I often get errors after the update? Every time I want to log on, I have to keep entering the [phone] number.	3	1
The latest update is a mess, the active period of a package cannot be checked, a 17GB quota has only been active for 5 days, and suddenly there is a notification that only 500MB is left in the quota even though I rarely use it because I am often connected to WiFi.	3	0
I have a problem in the current update where if I exit the application it logs out by itself. I don't know if it is a new feature or a bug, so I have to reenter to create a link to log into the application, which is very inconvenient. I hope the current login method can be reconsidered, maybe the previous login method can be brought back or a more efficient feature can be created, if possible, make it so that I can log in using a username/userid/email so it is easier.	3	1
Please, bro, I bought a plan here, the monthly price is pretty high, it can pay for unlimited Wi-Fi for a month, THE NETWORK SHOULDN'T BE SLOW!!! it is embarrassing, the quota is expensive, the network is terrible, Oh dear	3	0
Please fix the signal properly. I'm a long-time customer. 10 years isn't a short time, man. The quota is never cheap, and the signal is getting worse. Do you intend to please your customers or not?	2	0

TABLE IV
A CONFUSION MATRIX WITH THREE LABELS.

		Predicted		
		Label 1	Label 2	Label 3
Actual	Label 1	TP1	F12	F13
	Label 2	F21	TP2	F23
	Label 3	F31	F32	TP3

for the priority score and problem domain. Random Forest is the proposed classification algorithm in the research [27]. The first classification process involves determining the importance of necessary improvements based on the reviews. User feedback reviews with a low importance categorization may not be a concern to ISP management, but those with medium or high importance should be considered for improvement. After completing this process, the next step is to determine the area for improvement, which also involves classification by the Random Forest algorithm to determine whether the necessary improvement concerns operational or technical matters.

A crucial step in the research process is to evaluate the performance of a classification model [28]. It helps to determine how well the model can accurately predict the correct category for a new, unseen dataset [28]. Several performance metrics are used to assess the model's accuracy: confusion matrix, precision, recall, F1-score, and accuracy [29]. The confusion matrix table shown in Table IV has three label categories. The columns represent the predicted label, and the rows show the actual label. A confusion matrix is a visualization tool used in machine learning to evaluate the performance of a classification model [30].

In the confusion matrix, three attributes are mapped diagonally from the top-left to the bottom-right: TP1, TP2, and TP3. These represent True Positives, meaning that items with the attribute are correctly classified according to the true label. F12 indicates that an actual classification of Label 1 is wrongly classified as Label 2, while F13 indicates that an actual classification of Label 1 is wrongly classified as Label 3. Similarly, F21 and F23 represent actual Label 2 classifications that are incorrectly classified as Label 1 and Label 3, respectively. Meanwhile, F31 represents the misclassification of Label 3 as Label 1, and F32 represents the misclassification of Label 3 as Label 2.

Precision in Eq. (1) is used to evaluate positive classifications that are truly positive for each category [31]. True Positives, as indicated by the confusion matrix in Table IV, exist for each label, where TP1 is True Positive for Label 1, TP2 is True Positive for Label 2, and TP3 is True Positive for Label 3. Calculating the precision for Label 1, for example, is done using Eq. (2). The number of True Positive for Label 1 (TP1) is divided by the sum of TP1, the number of F21 False Positive (FP), and the number of F31 FP.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Precision Label 1} = \frac{TP1}{TP1 + F21 + F31} \tag{2}$$

The recall rate (Eq. (3)), also known as the True Positive rate, can be calculated for each class [32]. It measures the proportion of all True Positives that are correctly classified as positive. The recall rate is also evaluated using the confusion matrix data. Calculating the recall for Label 1, for example, is done using Eq. (4). The number of True Positive for Label 1 (TP1) is divided by the sum of TP1, the number of F12 FP,

and the number of F13 FP.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Recall Label 1} = \frac{TP1}{TP1 + F12 + F13} \tag{4}$$

The third metric that can be evaluated using the data mapping in the confusion matrix is the F1-score, given by Eq. (5) [33]. The F1-score is used to measure the harmonic mean for each category [33]. The example is shown in Eq. (6). Prec. shows precision, while Rec. shows recall. Finally, the accuracy of the model can be calculated using Eq. (7). This value is calculated by dividing the sum of all correctly classified data by the sum of all data from the confusion matrix.

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

$$\text{F1-Score Label 1} = \frac{2 \times \text{Prec. Label 1} \times \text{Rec. Label 1}}{\text{Prec. Label 1} + \text{Rec. Label 1}} \tag{6}$$

$$\text{Accuracy} = \frac{\text{Correct Classification}}{\text{Total Classification}} \tag{7}$$

III. Results and Discussion

This section presents the results and discussion of classification performance using the word embedding process and two different transfer learning architectures, with different output labels. Case 1 involves classification based on the combination architecture of BERT without fine-tuning, using the two Random Forests. In contrast, Case 2 involves classification based on the combination architecture of IndoBERT without fine-tuning, using the two Random Forests. The first Random Forest is used to classify the priority score. The second Random Forest is used to classify the problem domain. In Case 1 and Case 2, 30% of user feedback reviews are used for the classification process. Both of the Random Forest architectures are trained using the portion of data (70%) that is separated for the training purpose as explained in Fig. 1.

The dataset, as presented in Table I, is divided into two portions: the training dataset with 70% (see Table V) and the testing dataset with 30% (see Table VI). Table V shows the training dataset, which consists of 313 reviews with a score of 1, 360 with a score of 2, and 161 with a score of 3, according to their priority score. Meanwhile, the training dataset comprises 424 reviews in category 0 and 410 in category 1, based on their problem domain.

The remaining portion of the data (30%) forms the testing dataset, as shown in Table VI. The testing dataset likewise includes reviews categorized by priority score and problem domain. In the testing dataset, there are 112 reviews with a score of 1, 172 reviews with a score of 2, and 74 reviews with a score of 3. According to the problem domain, the testing dataset comprises 162 reviews in category 0 and 196 reviews in category 1.

TABLE V
Number of Reviews for Training Dataset (70%).

Priority Score			Problem Domain	
Score 1	Score 2	Score 3	Category 0	Category 1
313	360	161	424	410

TABLE VI
Number of Reviews for Testing Dataset (30%).

Priority Score			Problem Domain	
Score 1	Score 2	Score 3	Category 0	Category 1
112	172	74	162	196

TABLE VII
A Confusion Matrix of Priority Score Classification for Case 1.

		Predicted			Total
		Category 1	Category 2	Category 3	
Actual	Category 1	72	40	0	112
	Category 2	35	134	3	172
	Category 3	14	58	2	74

In Case 1, the transfer learning architecture of BERT, without fine-tuning the word embeddings, is used to extract feature vectors from the test data. The classifications of priority score and problem domain are involved in this case. The results of classification for priority score are presented in the confusion matrix in Table VII. As shown in Table VII, for reviews classified as score 1, 72 reviews are correctly classified, while 40 reviews are incorrectly classified as score 2. For reviews classified as score 2, 134 reviews are correctly classified. Meanwhile, 35 reviews are incorrectly classified as score 1, and 3 reviews are incorrectly classified as score 3. For reviews classified as score 3, only 2 reviews are correctly classified. Most of the reviews are incorrectly classified as score 1 (58 reviews) and score 2 (14 reviews).

The confusion matrix in Table VII shows the performance of the classification model, which can be further evaluated through several metrics. The evaluation results are shown in Table VIII. For the precision metric, the highest value is score 1 at 60%, and the lowest is score 3 at 40%. For the recall metric, the highest value is score 2 at 78%, and the lowest is score 3 at 3%. For the F1-score metric, the highest value is 66% for score 2.

The next step was the classification process to

Cite this article as: R. P. Maulana and Yulianto, "Combining Feature Embedding Based on the BERT Architecture and Random Forest Algorithm to Identify the Handling Section and Risk Priority Level in an ISP Company in Indonesia", CommIT Journal 20(2), 207–218, 2026.

TABLE VIII
PERFORMANCE METRICS OF PRIORITY SCORE CLASSIFICATION FOR CASE 1 BY USING BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT) AS THE WORD EMBEDDING OR FEATURE EXTRACTOR.

Label	Precision	Recall	F1-Score
1	0.60	0.64	0.62
2	0.58	0.78	0.66
3	0.40	0.03	0.05

TABLE IX
THE RESULT OF MAPPING IN CONFUSION MATRIX USING BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT) AS THE WORD EMBEDDING AND THE PROBLEM DOMAIN AS THE LABEL CATEGORY FOR CASE 1.

		Predicted		Total
		Category 0	Category 1	
Actual	Category 0	100	62	162
	Category 1	69	127	196

TABLE X
THE PROBLEM DOMAIN CATEGORY CLASSIFICATION RESULTS USING BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT) AS THE WORD EMBEDDING OR FEATURE EXTRACTOR FOR CASE 1.

Label	Precision	Recall	F1-Score
0	0.59	0.62	0.60
1	0.67	0.65	0.66

identify the problem domain as the area of necessary improvement. It is to see whether improvement is needed only for service matters (category 0) or for technical or engineering matters (category 1). Table IX shows the mapping of classification results for the second Random Forest architecture, which is trained to classify the two categories above as the output. The input to the Random Forest is the embedding produced by the transfer-learning BERT architecture without fine-tuning.

According to the confusion matrix in Table IX, for category 0, 100 reviews are correctly classified, and 62 reviews are incorrectly classified as category 1. Meanwhile, for category 1, 127 reviews are correctly classified, and 69 reviews are incorrectly classified as category 0. From Table IX, the metrics of each category can be calculated. In Table X, the precision values for the classification performance of categories 0 and 1 are 59% and 67%, respectively. The recall values for the classification performance are 62% for category 0 and 65% for category 1.

In Case 2, the IndoBERT transfer learning architecture, with word-embedding-based feature extraction, is used to extract feature vectors from the testing data. Like Case 1, two random forest algorithms are trained

TABLE XI
THE CONFUSION MATRIX OF THE CLASSIFICATION RESULT USING INDOBERT AS THE WORD EMBEDDING AND THE PRIORITY SCORE AS THE LABEL CATEGORY FOR CASE 2.

		Predicted			Total
		Category 1	Category 2	Category 3	
Actual	Category 1	71	41	0	112
	Category 2	32	135	5	172
	Category 3	10	62	2	74

TABLE XII
THE PRIORITY SCORE CLASSIFICATION RESULTS USING INDOBERT AS THE WORD EMBEDDING OR FEATURE EXTRACTOR FOR CASE 2.

Label	Precision	Recall	F1-Score
1	0.63	0.63	0.63
2	0.57	0.78	0.66
3	0.29	0.03	0.05

and tested. This case uses the extraction results of IndoBERT. The first random forest algorithm is used to classify priority scores. Priority score classification has three categories that indicate how important a problem is, with score 1 being the lowest and score 3 the highest. A score of 3 indicates a problem that needs to be resolved as soon as possible. The results for priority score classification are mapped to a confusion matrix in Table XI.

Table XI shows the priority score classification results for Case 2 (as the combination of IndoBERT with random forest). From 112 reviews classified as score 1, the model classifies 71 reviews, while 41 are incorrectly classified as score 2. In the 172 reviews with score 2, after being put through the model, 135 are correctly classified, while 32 are incorrectly classified as score 1. Moreover, 5 reviews are incorrectly classified as score 3. From the 74 reviews of score 3, using the model, only 2 are correctly classified. Around 10 reviews are incorrectly classified as score 1, and 62 reviews are incorrectly classified as score 2.

Table XII shows the performance metrics for each priority score. The proposed model is found to have moderate performance, especially for recognizing scores 1 and 2, with the F1-score between 0.63 and 0.66. However, the performance for score 3 is significantly lower, with an F1-score of only 0.05. The performance of the proposed model for score 1 is balanced, with the F1-score of only 0.05. In score 1, it also has the same precision and recall values of 0.63. The results indicate a reasonable trade-off between identifying positive instances and avoiding false positives.

The second random forest algorithm is used for problem domain classification. Problem domain clas-

TABLE XIII
THE RESULT OF MAPPING IN CONFUSION MATRIX USING INDOBERT AS THE WORD EMBEDDING AND THE PROBLEM DOMAIN AS THE LABEL CATEGORY FOR CASE 2.

		Predicted		Total
		Category 0	Category 1	
Actual	Category 0	117	45	162
	Category 1	63	133	196

TABLE XIV
THE PROBLEM DOMAIN CATEGORY CLASSIFICATION RESULTS USING INDOBERT AS THE WORD EMBEDDING OR FEATURE EXTRACTOR FOR CASE 2.

Label	Precision	Recall	F1-Score
0	0.65	0.72	0.68
1	0.75	0.68	0.71

sification has two categories that represent areas to be improved by the ISP management. Category 0 represents service matters, and the category of 1 represents technical matters. The results for problem domain classification are mapped in Table XIII. The classification results show that, for the 162 reviews in category 0, 117 are correctly classified, and 45 are incorrectly classified as category 1. Then, for the 196 reviews in category 1, 133 reviews are correctly classified, and 63 reviews are incorrectly classified as category 0.

The classification results are used to compute the precision, recall, and F1-score metrics, as shown in Table XIV. For category 0, the combination approach of the model has a higher value for recall (0.72) than for precision (0.65). This result suggests that the proposed model, combining IndoBERT and the Random Forest algorithm, is effective at identifying positive instances in category 0. However, it tends to classify some positive instances of category 1 incorrectly as category 0. For category 1, the proposed model has a higher value of precision (0.75) than recall (0.68). It indicates that the model often predicts category 1 instances correctly, but it may miss other positive instances.

Table XV presents an overall comparison of the classification using the testing dataset for priority score classification into three categories. The results show no difference between using BERT with Random Forest and using IndoBERT with Random Forest. The results are the same for precision, recall, F1-score, and accuracy. In terms of precision, both models have the same value, 0.53, indicating that their predictions are relatively accurate when correct. For the recall metric, both models also have the same recall value (0.58). They can identify a moderate proportion of positive cases. For the F1-score, both models have a value of 0.52, indicating a good balance between precision and recall.

TABLE XV
COMPARISON OF THE PRIORITY SCORE CLASSIFICATION RESULTS.

Model Architecture	Weighted Average			Accuracy
	Precision	Recall	F1-Score	
BERT + Random Forest	0.53	0.58	0.52	0.58
IndoBERT + Random Forest	0.53	0.58	0.52	0.58

Note: BERT: Bidirectional Encoder Representations from Transformers (BERT).

TABLE XVI
COMPARISON OF THE PROBLEM DOMAIN CLASSIFICATION RESULTS.

Model Architecture	Weighted Average			Accuracy
	Precision	Recall	F1-Score	
BERT + Random Forest	0.66	0.66	0.66	0.66
IndoBERT + Random Forest	0.70	0.70	0.70	0.70

Note: BERT: Bidirectional Encoder Representations from Transformers (BERT).

Table XVI shows the comparison between the two model architectures in the test scenario for problem domain classification using the testing dataset. It presents the evaluation metrics for the two different model architectures of BERT with Random Forest and IndoBERT with Random Forest. The evaluated metrics are Precision, Recall, F1-Score, and Accuracy. Both models show identical values across all metrics, with 0.66 for BERT with Random Forest and 0.70 for IndoBERT with Random Forest. The results indicate that both models perform similarly in terms of their ability to classify instances correctly, the recall of true positive instances, and the balance between precision, recall, and overall accuracy.

The data in Table XV represent the performance of both models in classifying whether user feedback reviews need to be addressed by the customer service or technical support department of an ISP. However, the accuracy rate is still only 58%. Moreover, in Table XVI, the classification accuracy achieved is 70% for the hybrid method of IndoBERT with Random Forest. It is superior to BERT with Random Forest. Nevertheless, the accuracy is relatively low, indicating that the model's predictive performance requires improvement. A higher accuracy rate is necessary for more reliable and precise risk classification, particularly in decision-making contexts involving customer feedback and operational evaluations. Future

enhancements can include optimizing hyperparameters, experimenting with deeper neural architectures, expanding and balancing the dataset, or integrating additional contextual features. These changes will improve the model's ability to generalize across diverse input data. Increasing the accuracy will also strengthen the model's practical applicability and enhance stakeholder trust in the system's predictions.

IV. Conclusion

In the research, a hybrid model is proposed to help ISPs to manage the issue of manually identifying user feedback reviews from loyalty applications on the Google Play Store. It is important to determine the problem domain (whether improvements are needed in service or technical matters within ISP management) and to identify the priority risk score (the immediacy of improving or resolving the matters) on a scale of 1 to 3. Two different hybrid architectures are compared for the proposed model: BERT + double Random Forest and IndoBERT + double Random Forest. Then, the dataset with two additional subcategories is used to train the proposed model, particularly the Random Forest classifier. Meanwhile, the BERT and IndoBERT models are frozen because they are only used as feature embeddings.

The results show that combining IndoBERT with Random Forest yields better performance than using BERT, achieving 70% accuracy for problem domain classification. However, for priority score classification, both architectures in the proposed methods show similar unsatisfactory results, with an accuracy of only 58%. For reference, the training dataset is only used for the Random Forest, while BERT and IndoBERT are only used as word embeddings, which may affect testing results. There is still an opportunity to improve classification accuracy. Because the results from the proposed models across both architectures show that performance can still be improved, a real human evaluation is needed to double-check the classification results.

The research has several limitations that need to be acknowledged. First, the dataset used to train the model is limited to user feedback ratings from Google Play Store with a single internet service provider in Indonesia (Telkomsel). It may limit the generalizability of the results to other ISPs or different platforms. Second, manual labeling of the dataset introduces potential for subjectivity and inconsistency, even when internal ISP managers perform it.

Future research can address the limitations and further explore the potential of sentiment analysis to improve ISP services. Expanding the dataset to include feedback from multiple IPSs and platforms will improve the generalizability of the findings. Investigating methods to reduce subjectivity in manual labeling, such as using multiple annotators or automated labeling techniques, can improve model accuracy. Additionally, exploring different deep learning architectures for feature extraction, particularly to enhance the accuracy of binary classification for problem domain identification, can improve classification performance. Next, a hybrid deep learning architecture (such as a text graph convolutional neural network with a graph neural network) can be used because the grammatical structure of user feedback reviews is often informal. Finally, it is possible to research the impact of implementing the sentiment analysis results on customer satisfaction and ISP performance.

Acknowledgement

The authors would like to acknowledge the financial support provided by the Department of Computer Science at Bina Nusantara University's School of Computer Science during the research period (2024–2025), The authors are grateful for the provided funding, which made the successful execution of this project, as well as for access to the university's research facilities and technical infrastructure.

Author Contribution

Conducted the literature review and determined the model architecture, R. P. M.; Contributed to the development of the application used to crawl the data in the Play Store and performed the data pre-processing, R. P. M.; Contributed to providing the new labelling data to meet the company's needs, R. P. M.; Developed the machine learning programme to assist with processing the data, R. P. M.; Supervised and provided recommendations during the literature review process, Y.; Contributed to the development of a machine learning testing model, Y.; and Contributed to the writing of the paper, Y.

Data Availability

The data that support the findings of the research are openly available in Zenodo at https://doi.org/10.5281/zenodo.16759141.

References

[1] W. Riani and S. Haryadi, "Effect of uneven distribution of broadband Internet services in developing countries during economic recession," *Journal of Distribution Science*, vol. 20, no. 10, pp. 1–9, 2022.

[2] B. H. Hayadi *et al.*, "An extensive exploration into the multifaceted sentiments expressed by users of the myIM3 mobile application, unveiling complex emotional landscapes and insights," *Journal of Applied Data Sciences*, vol. 5, no. 2, pp. 357–366, 2024.

[3] Y. Tirana and Sfenrianto, "Factors on mobile application user satisfaction in the largest Indonesian Internet Service Provider (ISP)," *CommIT (Communication and Information Technology) Journal*, vol. 17, no. 2, pp. 199–208, 2023.

[4] R. Yati, "Survey APJII: Pengguna Internet di Indonesia tembus 215 juta orang," 2023. [Online]. Available: https://bit.ly/4gRLh8n

[5] S. Scherrer, S. Tabaeiaghdaei, and A. Perrig, "Quality competition among internet service providers," *Performance Evaluation*, vol. 162, pp. 1–36, 2023.

[6] W. S. Ismail, M. M. Ghareeb, and H. Youssry, "Enhancing customer experience through sentiment analysis and natural language processing in e-commerce," *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA)*, vol. 15, no. 3, pp. 60–72, 2024.

[7] R. I. Kurnia and A. S. Girsang, "Classification of user comment using Word2Vec and deep learning," *Advances in Science, Technology and Engineering Systems Journal*, vol. 6, no. 2, pp. 566–572, 2021.

[8] I. J. Tae, A. Broillet-Schlesinger, and B. Y. Kim, "Effect of motivational factors on the use of integrated mobility applications: Behavioral intentions and customer loyalty," *Information*, vol. 15, no. 9, pp. 1–19, 2024.

[9] T. H. Rohwiyati, A. I. Setiawan, L. Wahyudi, E. Dwi, and M. Amperawati, "E-trust and e-service quality on e-loyalty: Role of e-satisfaction and customer privacy," *Journal of Ecohumanism*, vol. 3, no. 4, pp. 3130–3143, 2024.

[10] C. Y. Li *et al.*, "Mobile social service user identification framework based on action-characteristic data retention," *IEEE Access*, vol. 8, pp. 127 748–127 767, 2020.

[11] M. Ariyanti, S. Widiyanesti, and W. H. Aprillia, "Service quality analysis of Telkomsel case study based on online customer reviews in Google Play Store," in *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, vol. 5. Greater Noida, India: IEEE, Feb. 9–10, 2024, pp. 1669–1674.

[12] M. Dhotay, M. Dharrao, S. Deokate, A. Bongale, and D. Dharrao, "Multimodal sentiment analysis: Emerging innovations, core challenges, and future directions," *Discover Artificial Intelligence*, vol. 6, pp. 1–28, 2026.

[13] E. Yulianti and N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT: Single-sentence and sentence-pair classification approaches," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 5, pp. 3579–3589, 2024.

[14] N. Paltrinieri, L. Comfort, and G. Reniers, "Learning about risk: Machine learning for risk assessment," *Safety Science*, vol. 118, pp. 475–486, 2019.

[15] W. Liao *et al.*, "Mask-guided BERT for few-shot text classification," *Neurocomputing*, vol. 610, 2024.

[16] D. Dharrao *et al.*, "An efficient method for disaster tweets classification using gradient-based optimized convolutional neural networks with BERT embeddings," *MethodsX*, vol. 13, pp. 1–10, 2024.

[17] Y. Xiong, G. Chen, and J. Cao, "Research on public service request text classification based on BERT-BiLSTM-CNN feature fusion," *Applied Sciences*, vol. 14, no. 14, pp. 1–11, 2024.

[18] J. I. T. Krisna, A. Luthfiarta, L. D. Cahya, S. Winarno, and A. Nugraha, "Comparing optimizer strategies for enhancing emotion classification in IndoBERT models," *Advance Sustainable Science, Engineering and Technology (ASSET)*, vol. 6, no. 2, pp. 1–8, 2024.

[19] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," *International Journal of Electrical & Computer Engineering (IJECE)*, vol. 14, no. 1, pp. 1071–1078, 2024.

[20] H. Imaduddin, F. Y. A'la, and Y. S. Nugroho, "Sentiment analysis in Indonesian healthcare applications using IndoBERT approach," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, pp. 113–117, 2023.

[21] O. Galal, A. H. Abdel-Gawad, and M. Farouk, "Rethinking of BERT sentence embedding for text classification," *Neural Computing and Applications*, vol. 36, no. 32, pp. 20 245–20 258, 2024.

[22] N. Darraz, I. Karabila, A. El-Ansari, N. Alami, and M. El Mallahi, "Integrated sentiment analysis with BERT for enhanced hybrid recommendation systems," *Expert Systems with Applications*, vol. 261, 2025.

[23] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, "Sentiment analysis of social media Twitter with

case of anti-LGBT campaign in Indonesia using naïve bayes, decision tree, and random forest algorithm," *Procedia Computer Science*, vol. 161, pp. 765–772, 2019.

[24] Handrizal, T. H. F. Harumy, Herriyance, and M. A. Ilmi, "Sentiment analysis based on 7P marketing mix aspects of the Indriver application service using the BERT algorithm, based on user reviews on the Google Play Store," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 19, pp. 6136–6144, 2023.

[25] K. Aziz *et al.*, "Unifying aspect-based sentiment analysis BERT and multi-layered graph convolutional networks for comprehensive sentiment dissection," *Scientific Reports*, vol. 14, pp. 1–22, 2024.

[26] D. Y. Yefferson, V. Lawijaya, and A. S. Girsang, "Hybrid model: IndoBERT and long short-term memory for detecting Indonesian hoax news," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 2, pp. 1913–1924, 2024.

[27] T. Wang, "Improved random forest classification model combined with C5. 0 algorithm for vegetation feature analysis in non-agricultural environments," *Scientific Reports*, vol. 14, pp. 1–13, 2024.

[28] S. Redjeki, S. Abadi, D. Kurniati, S. R. C. Nursari, A. Damayanti, and E. Iskandar, "Clustering on sentiment analysis: Effect of Twitter dataset," *Journal of advanced research*, vol. 51, no. 1, pp. 39–51, 2024.

[29] Q. Xi and P. Jiang, "Design of news sentiment classification and recommendation system based on multi-model fusion and text similarity," *International Journal of Cognitive Computing in Engineering*, vol. 6, pp. 44–54, 2025.

[30] I. Kanwal, F. Wahid, S. Ali, A.-U. Rehman, A. Alkhayyat, and A. Al-Radaei, "Sentiment analysis using hybrid model of stacked auto-encoder-based feature extraction and long short term memory-based classification approach," *IEEE Access*, vol. 11, pp. 124 181–124 197, 2023.

[31] K. Barik and S. Misra, "Analysis of customer reviews with an improved VADER lexicon classifier," *Journal of Big Data*, vol. 11, pp. 1–29, 2024.

[32] L. A. Gaafar, A. Z. Ghalwash, A. A. Youssif, and H. A. Ghalwash, "Machine and deep learning models for multiclass sentiment classification," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 22, pp. 8325–8339, 2024.

[33] A. Maiti, A. Abarda, and M. Hanini, "The impact of feature extraction techniques on the performance of text data classification models," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 35, no. 2, pp. 1041–1052, 2024.