Vol. 7, No. 4, December 2025

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Integrating Aspect-Based Sentiment Analysis with CSI–IPA for Telecommunications App Development

Siti Hariza Marshella[1], Dedy Kurniawan[2*], Nabila Rizky Oktadini [3], Rudi Sanjaya[4], Putri Eka Sevtiyuni[5], Naretha Kawadha Pasemah Gumay[6]

[1,2,3,4,5,6] Information system department, Faculty of computer science, Universitas Sriwijaya, Palembang, Indonesia

Received: Nov 4, 2025
Revised: Nov 26, 2025
Accepted: Dec 21, 2025
Published: Dec 31, 2025

Corresponding Author:
Author Name*:
Dedy Kurniawan
Email*:
dedykurniawan@unsri.ac.id

DOI: 10.63158/journalisi.v7i4.1363

Abstract. This study integrates Aspect-Based Sentiment Analysis with Customer Satisfaction Index (CSI) and Importance Performance Analysis (IPA) to determine priority development features in digital service applications in the telecommunications sector. A total of 10,000 MyIM3 user reviews were analyzed using sentiment analysis with Pseudo-labeling and Fine-Tuned IndoBERT, then from the results of the analysis, negative sentiment was mapped into several topics using LDA. The topic is used to compile question indicators based on the five dimensions of SERVQUAL. After the questionnaire data is declared valid and reliable, CSI and IPA analysis is carried out. A CSI value of 79.22% indicates that user satisfaction is in the "borderline" category, but several aspects still need to be improved, especially system updates (RS3), application attention to user needs (E2), and feature personalization (E3) which are in quadrant I (Concentrate Here). This hybrid approach offers novelty by demonstrating how ABSA and LDA can be systematically integrated with CSI and IPA to provide more comprehensive and user-oriented insights. The limitations of this study include focusing on negative sentiment data for feature exploration, as it is most relevant for identifying problems and opportunities for improvement and development of telecommunication digital services.

Keywords: ABSA; CSI; IPA; Product Development Strategy; Telecommunications

1. INTRODUCTION

Responsive and reliable digital service quality is a crucial factor in enhancing customer satisfaction and loyalty, especially in the highly competitive telecommunications industry [1] To support these needs, Indosat IM3 provides the MyIM3 application as the main service channel. However, browsing on the Google Play Store shows many negative reviews related to login issues, disruptive promo pop-ups, inconsistent information, and slow service responses. Therefore, these public reviews serve as a potential data source for objective service quality evaluation [2]. Service quality evaluation is generally conducted using the Customer Satisfaction Index (CSI), which refers to five SERVQUAL dimensions [3], as well as a matrix that measures the performance and importance of service quality [4]. This matrix is the Importance-Performance Analysis (IPA), which maps service attributes based on their importance and performance levels to determine improvement priorities [5].

However, the two methods still rely on structured questionnaire data and have not yet utilized sentiment information sourced from user opinions in free text form. To optimize app evaluation, it is necessary to collect user review data from platforms like Google Play Store, which provide many user inputs. However, this data is unstructured, making it difficult to analyze without the support of Natural Language Processing (NLP) approaches. The development of language models such as IndoBERT and LDA enables automatic extraction of sentiment, topics, and user insights, allowing for a more comprehensive and experience-based evaluation of service quality.

Several previous studies have utilized automated text methods such as text mining and the CSI and IPA frameworks to evaluate service quality. Research by [6] integrated SERVPERF, CSI, and IPA to assess service performance in non-formal educational institutions and successfully identified priority attributes through the IPA quadrant. Another study on BUMN service applications [7] also combined CSI, IPA, and SERVQUAL, resulting in a CSI value of 75.85% and identifying seven attributes that need improvement. In addition to questionnaire-based evaluations, some studies have begun to combine machine learning methods with these frameworks. For example, Febrianto's research [8] merged sentiment analysis based on SVM on Google Play reviews with IPA to map important attributes, but it has not yet involved CSI as a satisfaction index.

Another study analyzed user reviews of the BRImo application using Naïve Bayes and CSI [9], but it has not integrated IPA, so it does not yet provide strategic priorities for service improvement. Meanwhile, research on e-commerce applications [10] combined SVM-based opinion mining with the CSI–IPA framework to produce strategic service mapping. However, all these studies are still limited to conventional text mining and have not achieved deep integration between sentiment analysis based on advanced NLP models like IndoBERT (ABSA) and topic modeling with LDA mapped to the SERVQUAL framework.

Although various previous studies have combined the CSI and IPA frameworks to evaluate digital service quality, and some studies have applied text mining to app reviews, the methods used are still simple, such as Naïve Bayes and SVM. To date, there has been no systematic research integrating in-depth analysis based on deep learning NLP specifically Aspect-Based Sentiment Analysis (ABSA) using IndoBERT with LDA topic modeling to automatically uncover service issues and connect them with the SERVQUAL, CSI, and IPA dimensions in telecommunication service applications. This gap highlights the need for an integrated approach capable of converting user opinions into service quality mapping and prioritizing digital application improvement strategies in a more comprehensive, user-experience-based.

Therefore, this research integrates Aspect-Based Sentiment Analysis with the CSI and IPA frameworks in a mobile customer service application within the telecommunications industry, using secondary data in the form of Google Play Store reviews. This approach utilizes NLP techniques to identify the most frequently discussed service aspects by users, then maps these aspects to several questions in the form of a questionnaire that can be validated. The results are then weighted within the CSI framework to assess user satisfaction and mapped onto the IPA matrix to determine priority indicators for product development, as well as conducting a gap analysis between expectations (aspect frequency) and perceptions (sentiment scores). The contribution and novelty (state-of-the-art) of this research lie in the utilization of public reviews as an alternative (secondary) data source for measuring service quality, as well as the integration of sentiment analysis (NLP) with the CSI and IPA frameworks in the context of digital applications, particularly in the telecommunications sector. It also involves the integration of automation and semi-automation processes and the development of a new analytical framework that combines modern NLP methods with service quality evaluation

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

instruments, thereby helping digital service providers better understand user needs and formulate more targeted development strategies.

2. METHODS

This research method consists of several main stages, starting from collecting user review data, text preprocessing, sentiment analysis using IndoBERT, LDA topic modeling, and integrating the results into the CSI–IPA framework. The overall flow of the research is presented in the following Figure 1.

.

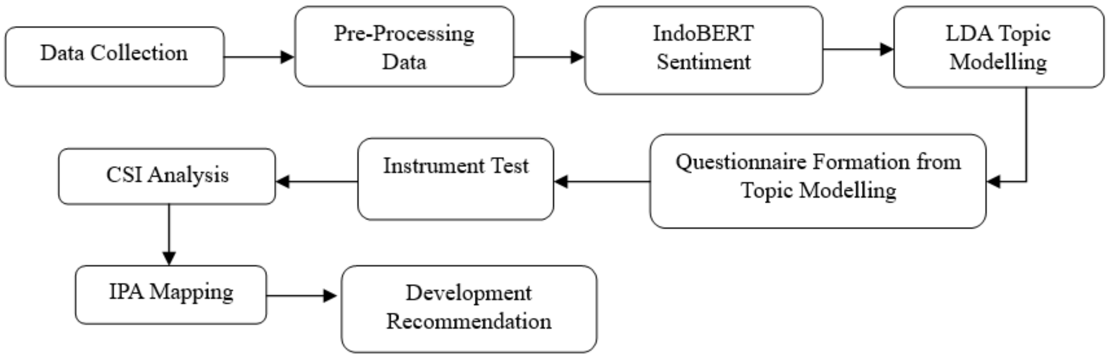

Figure 1. Flowchart Research Stages

This research uses a descriptive quantitative method by leverage both primary and secondary data. Primary data were acquired through the distribution of the questionnaires [11]. This phase is used to determine the Customer Satisfaction Index (CSI) framework and Importance Performance Analysis (IPA), while secondary data were sourced from user reviews of the MyIM3 application on Google Play Store. The quantitative method was chosen because it can objectively test the relationships between variables through statistical approaches, thereby minimizing bias and strengthening empirical evidence for the underlying theoretical assumptions [12]. The study population consists of active users of the MyIM3 application in Indonesia. The sampling technique used is purposive sampling, which draw in selecting samples based on specific criteria matched with the research objectives [13]. Based on previous research references, the sample includes 100 active MyIM3 user respondents for primary data and at least 1,000 user reviews from Google Play Store as secondary data. In general, the research process is carried out through two main stages,

1. **1.** Sentiment Analysis (Text Mining Stage), which involves collecting and processing secondary data in the form of user reviews of the MyIM3 application on the Google Play Store to identify the main topics of complaints and user perceptions.
2. **2.** User Satisfaction Analysis (CSI-IPA Stage), which involves developing a questionnaire based on sentiment analysis results, followed by CSI calculation and indicator mapping using IPA to determine the priority for application feature development.

2.1. Data collection (Data Scrapping)

User review data collection was carried out using the google-play-scraper library based on Python, which allows direct data extraction from the Google Play Store. This library was chosen because it is the official library , making it accurate for retrieving application metadata [14]. The data retrieval process involved calling the reviews() function on the MyIM3 app's package name (com.pure.indosat.care) with Indonesian language (lang='id') and Indonesia as the country (country='id'). A total of 10,000 recent reviews were downloaded from June to August 2025, including review text, ratings, and timestamps. All data were then stored in a table format using pandas to facilitate cleaning, preprocessing, and further analysis during NLP and topic modeling stages. These steps aim to clean the data from irrelevant characters and standardize the text format for effective analysis [14]. The scraping process was run with the following configuration:

1. Library = google-play-scraper
2. Parameter sort = Sort.NEWEST to ensure reviews are in the most recent order
3. Parameter count = 10000 to maximize data volume

This library works through Google's internal API, eliminating the need for web crawling or DOM parsing, making it more efficient and reducing the risk of blocking.

2.2. Data Pre-processing

The pre-processing stage is carried out to improve the quality of the corpus before further analysis. The process includes case folding, cleaning non-alphabet characters, normalization of non-standard words, stopword removal, tokenization, and stemming. Normalization is used to address language variations in user reviews, while stopword removal aims to reduce words that do not have significant semantic contribution to the

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

model[15] . The following configuration for the pre-processing and normalization stages can be seen in Table 1.

Table 1. Model Configuration Pre-processing

Pre-Processing Stage	Process Description	Tools / Library	Output Produced
Case Folding	Converts the entire text to lowercase.	Python (string lower)	Text with no capitalization differences.
Cleaning	Remove numbers, URLs, emojis, punctuation, and non-alphabetic characters.	Regex (re), emoji lib	Clean text is noise-free.
Tokenization	Breaking down sentences into word tokens.	IndoNLTK / NLTK	Word tokens per review.
Stopword Removal	Removing common words that carry no semantic meaning.	Sastrawi stopword list (custom extended)	Token means the main thing.
Stemming	Returns the word to its basic form.	Sastrawi Stemmer	Root word for analysis.
Normalization	Converting non-standard words to standard words (e.g., “gk” → “tidak”, “gabisa” → “tidak bisa”).	Python mapping dictionary (rule-based)	Consistent formal text.

The process of normalizing the text in this study was carried out to ensure that user reviews mixed between standard and non-standard languages can be processed consistently by the NLP model. Normalization was carried out using a slang dictionary based on mapping rules that were manually compiled based on the variety of informal languages of Indonesian users on the Google Play Store. This normalization dictionary contains non-standard word pairs and their equivalents in standard Indonesian, such as "gk" → "tidak", "bgt" → "sangat", and "gabisa" → "tidak bisa". The rule-based dictionary approach was chosen because it is more accurate to handle variations of Indonesian slang that are not covered in the general normalization module[16]. After the normalization of slang was carried out, this study used Literary stemming as a stage of word reduction. The selection of stemming is based on several scientific justifications. The Sastrawi stemmer is one of the most mature and widely used algorithm in Indonesian text mining research, and has been proven to be effective in improving model performance in various opinion analysis and text classification studies [17]. Also, stemming Sastrawi is better to handle variations of non-standard words and complex suffixes on large datasets [18]. Therefore, stemming is seen as more effective for NLP tasks such as sentiment analysis and topic modeling based on user reviews.

2.3. Pseudo-Labelling and Sentiment Analysis

Sentiment analysis is a technique of extracting opinions from texts to categorize opinions as positive, neutral, or negative [19] The modern approach utilizes a transformer language model that has been pretrained and then fine-tuned to specific domain data (IndoBERT for Indonesian). Fine-tuning allows the model to capture the local linguistic context (slang, abbreviations, spelling variations) thereby improving the classification accuracy of application review data [20]

In this research, the first process was carried out starting from pseudo-labeling to expand the labeled dataset using a semi-supervised approach. The initial model of Indonesian RoBERTa was used to predict sentiment labels on unlabeled data, then only predictions with a confidence score of ≥ 0.70 were maintained to ensure label quality and reduce the propagation of errors. This approach follows modern pseudo-labeling practices that emphasize strict filtration to maintain the quality of the dataset [21].. The dataset of pseudo-labeling results was then stratified into training and testing data. The following configuration for pseudo-labelling stages can be seen in Table 2.

Table 2. Model Configuration Pseudo-labelling

Component	Settings
Models	w11wo/indonesian-roberta-base-sentiment-classifier
Labelling Method	Pseudo-labeling semi-supervised
Confidence threshold	0.70
Batch size prediction	32

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Component	Settings
Output label	positive, neutral, negative, uncertain
Filtering	Only confidence ≥ 0.70 is used

After performing pseudo-labeling, the next step is to fine-tune IndoBERT using a high-confidence pseudo-labeled dataset. IndoBERT was chosen because it is a pre-trained Transformer-based model developed specifically for the Indonesian language and has been proven to produce better contextual representations compared to models based on traditional (word-level) architectures. The dataset was divided into two parts using stratified split, namely 80% for training and 20% for testing, to maintain balanced sentiment class distribution as recommended in supervised learning practices. Below is the IndoBERT model configuration presented in Table 3.

Table 3. Model Configuration IndoBERT

Component	Configuration
Model	indobenchmark/indobert-base-p1
Input max length	128 token
Batch size	16
Optimizer	AdamW
Learning rate	2e-5
Weight decay	0.01
Epoch	3
Evaluation strategy	steps (every 200 step)
Save strategy	steps
Load best model	Yes (based on eval_loss)
Loss Function	Cross Entropy
Device	CPU

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

The training process was carried out using the AdamW optimizer because it has better convergence stability on the Transformer model. The training parameters used include a learning rate of 2e-5, a batch size of 16, and a maximum sequence length of 128 tokens, which is suitable for application reviews that are generally short and do not require long context. Additionally, the settings eval_strategy=steps, save_strategy=steps, and load_best_model_at_end=True were used to ensure that the saved model is the one with the best validation loss. After training is complete, the fine-tuned IndoBERT model is used to perform sentiment classification on the entire review dataset, the addition of Confidence Threshold and Neutral Calibration before model evaluation resulting in more stable final predictions.

2.4. LDA Topic Modelling

Latent Dirichlet Allocation (LDA) is a probabilistic topic modelling method that views each document as a mixture of various latent themes (topics), and each topic is represented as a probability distribution of certain words. LDA allocates words to topics based on the frequency of co-occurrence and calculates the distribution of topics for each document using Bayesian inference [22]. LDA configuration model presented In Table 4.

Table 4. Model Configuration LDA

Component	Configuration
Library	Gensim
Dataset	Negative review of IndoBERT prediction results
Tokenization	Lowercase, cleaning, stemming (Sastrawi)
Stopwords	Stopwords Bahasa Indonesia + lexicon positive
Dictionary Filtering	no_below=5, no_above=0.5
Number of topics tested	5–20
Number of selected topics	Based on the highest coherence and the number of questionnaire indicators
Alpha	auto
Eta	auto
Passes	20

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Component	Configuration
Iterations	500
Coherence metric	Cv

In this study, to better understand the sources of issues in negative reviews, topic modeling using Latent Dirichlet Allocation (LDA) was conducted. Only reviews classified as negative by the IndoBERT model were analyzed using LDA to ensure that the generated topics focused on user pain points. Tokenization was performed by removing Indonesian stopwords and applying stemming using Sastrawi. The number of topics was tested within the range of 5–20 and was selected based on the highest Coherence Score (Cv) value and manual interpretability according to the needs of questionnaire indicator development. LDA parameters such as alpha and eta used auto settings so that the model could adaptively adjust the topic distribution according to the data structure.

2.5. Questionnaire Development and Test Instrument

The questionnaire instrument in this study was developed based on the adaptation of the SERVQUAL model for the digital service context. Considering that mobile applications have different characteristics from conventional services, the SERVQUAL indicators were adjusted to the context of usability, system reliability, and user experience as recommended by recent studies. The modified SERVQUAL can be effectively applied to digital applications while maintaining the five main dimensions: Tangibles, Reliability, Responsiveness, Assurance, and Empathy. [23] Here is an example of a SERVQUAL question that can be adapted, shown in Table 5 [24].

Table 5. Questionnaire Example

Dimension	Indicator / Questionnaire Question
Reliability	The application makes it easy for you to conduct transactions.
Responsiveness	Customer service handles complaints quickly and efficiently.
Assurance	The application ensures the security of users' personal data.
Empathy	The application frequently offers discounts and coupons.
Tangibles	The application provides notifications when there is new information.

Respondents will assess several indicators on a Likert scale (1-5) for importance and performance [25], [26]. Each indicator is selected to measure users' perceptions of the application service's performance and importance. The validity of the instrument was tested using Pearson Product-Moment correlation (r-calculated > r-table), while the reliability test was conducted using Cronbach's Alpha with a value ≥ 0.80, in accordance with the minimum standard for instrument feasibility in social research [26]

2.6. Data Analysis and Development Recommendations

The validated questionnaire data was analyzed using the CSI method to calculate user satisfaction levels based on importance weights and performance [27]. Additionally, a Gap Analysis was conducted to identify the differences between user perceptions and expectations. Furthermore, the results of the calculations were mapped using IPA to determine the indicators that should be prioritized for service improvement and development. The final stage of the research resulted in recommendations for the development of the MyIM3 application based on the CSI and IPA analysis results. The recommendations focus on improving features or service aspects with high importance levels but still low performance, serving as a basis for system development strategies and enhancing user experience.

3. RESULTS AND DISCUSSION

This stage details the results of sentiment analysis of customer reviews regarding the My IM3 application, following research steps: data collection, pre-processing, classification, and questionnaire development. Data is then gathered from respondents, the research instrument is tested using validity and reliability tests, and the questionnaires are weighted based on CSI weighting and mapped using IPA. This process leads to conclusions about which features are most urgent for development based on their performance and importance.

3.1. Data Collection

The first stage of this research is scraping 10,000 customer review data taken from the latest update on the Google Play Store. The graph below shows the number of reviews based on the ratings of the My IM3 application.

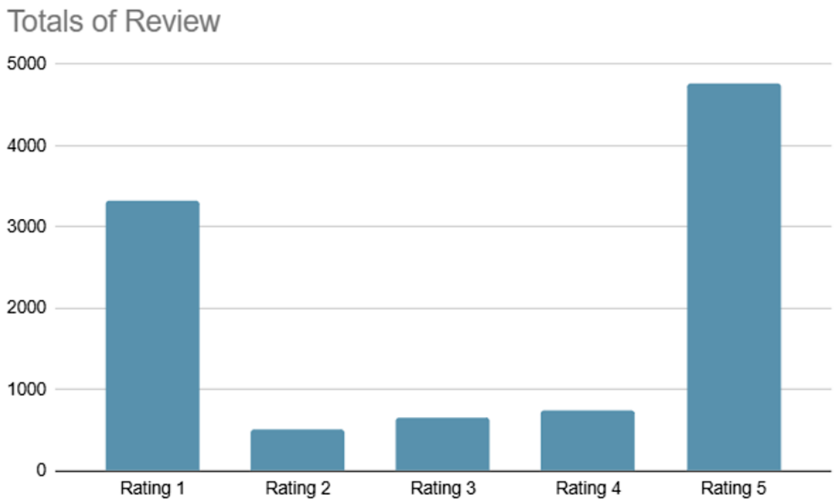

Figure 2. Rating Classification

Based on Figure 2, the user who gave a 5-star review numbered 4,762, which is the highest among the total 10,000 reviews. However, the 1-star reviews also make up a significant proportion, totaling 3,315 reviews. Additionally, there are 755 4-star reviews, followed by 648 3-star reviews, and the last is 520 2-star reviews. After the data is collected or scrapped from the Google Play Store, the next phase is data preprocessing before later the data will go through the analysis stage.

3.2. Pre-processing Data

In the pre-processing phase, the data is cleaned by changing the data format to lowercase text, also known as case folding. Then, the data is cleansed by removing irrelevant attributes such as URLs, punctuation, and numbers. Next, tokenization is performed, which involves breaking sentences into words. After that, text normalization is applied to convert non-standard words into standard words. Once the words are standardized, stopwords such as conjunctions or words that are less necessary for analysis are removed. Finally, affixed words are reduced to their root form, such as changing the word “membeli” to the base word 'beli”.

Table 6 shows an example of the results of *the preprocessing* process on the review text data. The *clean_text* column is the output of the text cleanup stage which includes lowercase normalization, URL removal, symbol removal, and space fireplace. The data from this *preprocessing* result is then used in the sentiment analysis stage.

Table 6. Preprocessing

No	Content (Indonesia)	Clean Text (Indonesia)
1	*Indosat modal dong sinyal jelek amat, kaya Tel...*	*indosat modal dong sinyal jelek amat kaya tel*
2	*aplikasi arek maem mbok spam notifikasi*	*aplikasi arek maem mbok spam notifikasi*
3	*pembelian paket internet gagal, tapi saldo udah...*	*pembelian paket internet gagal tapi saldo udah*
4	*mntap*	*mntap*
5	*oke*	*oke*

3.3. Sentiment Analysis

Sentiment analysis in this study was performed through two main stages, namely pseudo-labeling using the RoBERTa model and fine-tuning the IndoBERT model using the resulting semipseudo labels. This two-stage approach was chosen to improve the quality of labels in the review data so that the final model performs better in the data domain used.

1) Pseudo-Labeling Using Indonesian RoBERTa

The initial stage of sentiment analysis is pseudo-labeling, which is the process of automatically generating sentiment labels using a pre-trained model. The model used is: w11wo/indonesian-roberta-base-sentiment-classifier. Then the model and tokenizer are loaded and run in evaluation mode to ensure the stability of predictions. Prediction is carried out using a batching mechanism measuring 32 texts per process. This batch approach is used to avoid empty input errors, prevent out-of-memory, improve processing speed. Each batch is tokenized and processed to generate a logit score, then retrieved the argmax value as a sentiment class (0 = negative, 1 = neutral, 2 = positive). These pseudo labels are stored in the pseudo_label and pseudo_text columns. The pseudo-label distribution was analyzed to see the composition of the initial sentiment before training. Only data that has a confidence threshold of > 0.75 meets the training and testing data.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

2) Split Data for Training and Evaluation

The dataset is then splitted into data train (80%) and data test (20%) using *train_test_split*. The label used is a pseudo label from RoBERTa. This step make sure that the IndoBERT model can be properly evaluated on data that was not used during training.

3) Tokenization Using IndoBERT Tokenizer and Fine-Tuning Model IndoBERT

The model used in the fine-tuning stage is: indobenchmark/indobert-base-p1 The tokenization process is carried out using the IndoBERT tokenizer in the clean_text column, and the pseudo label is moved into the labels column. The IndoBERT model was then derived (fine-tune) using a pseudo-labeled dataset. This fine-tuning allows IndoBERT to learn sentiment patterns that are more in line with the review domain used than using a direct pre-trained model.

4) Sentiment Prediction Using the Trained IndoBERT Model

After the training process is completed, the IndoBERT model is used again to make predictions on the test dataset (test_token). Predictions are made through trainer.predict(test_token). The output in the form of a logit value is converted into a sentiment class through the argmax function. The numerical labels are re-mapped into positive, neutral, and negative text labels. Then all the results are combined into the Table 7.

Table 7. Analysis Sentiment

No	Content (Indonesia)	Clean Text (Indonesia)	Sentiment
1	*di mana-mana lelet, katanya makin stabil...*	*di mana mana lelet katanya makin stabil mana ada*	negative
2	*pulsa selalu ilang gk jelas.!*	*pulsa selalu ilang gk jelas*	negative
3	*jaringan tolong di perbaiki lah...*	*jaringan tolong di perbaiki lah lagi kerja tiba tiba hilang*	negative
4	*good*	*good*	positive
5	*uang anda kebanyakan? ga tau mau buat apa?*	*uang anda kebanyakan ga tau mau buat apa coba*	negative

Overall, from the results of the sentiment analysis with data onfidence threshold >0.75 shown in Table 7 dan Table 8, conclude that many of users convey negative sentiment,

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

which is demonstrated by the percentage of negative reviews as high as the positive with the proportion of positive reviews as many as 948, and neutral reviews as many as 87 and negative reviews as many as 890. This stage generates a final table of sentiment which is then used as the basis for further analysis, including sentiment mapping for LDA Topic Modelling.

Table 8. Analysis Sentiment in Total

Sentiment	Total
Negative	890
Positive	948
Neutral	87
Total	1925

5) Model Evaluation

The model evaluation stage was carried out to measure how well the IndoBERT model that has gone through *the pseudo-labeling* process is able to predict sentiment on the test data. This evaluation uses two main components, namely the Classification Report and the Confusion Matrix. The pseudo-labeled IndoBERT model performs very well, especially for two main classes: positive and negative. The 93% accuracy indicates a highly reliable model for app review sentiment analysis.

Table 9. Classification Report

Sentiment	Precision	Recall	F1-Score	Support
Negative	0.92	0.96	0.94	890
Neutral	0.70	0.46	0.56	87
Positive	0.95	0.95	0.95	948
Accuracy			0.93	1925
Macro Avg	0.86	0.82	0.81	1925
Weighted Avg	0.93	0.93	0.93	1925

Based on Table 9, the model is excellent for both positive and negative classes, as evidenced by the F1-score ≈ 0.95 meanwhile, the neutral class has a lower performance

(F1-score = 0.56), which is common because the amount of neutral data is much less (only 87 samples). However, overall the accuracy of the model is very high, at 93% and also the weighted average is close to 0.93, indicating that the model's performance is stable.

Table 10. Confussion Matrix Result

Actual _Predicted	Negative	Neutral	Positive
Negative	850	9	31
Neutral	30	40	17
Positive	40	8	900

Meanwhile from the Table 10, for the results of the confusion matrix, the model is almost always correct in predicting positive and negative sentiment. The biggest error was in the neutral class, which many were mispredicted as negative (30 cases) and positive (17 cases).

To overcome neutral-class linguistic ambiguity in reviews that do not show clear emotions, this study applies threshold-based neutral calibration after probabilistic predictions are calculated with softmax. If the IndoBERT model predicts a positive or negative class but the confidence value is below the 0.55 threshold, the prediction is corrected to neutral. This mechanism prevents misclassification of ambiguous sentences such as informative comments or minor complaints that are often erroneously mapped to extreme sentiment due to the similarity of word distributions. Overall, the model has shown good results so that it can be used for sentiment analysis on the review dataset effectively.

3.4. LDA Topic Modelling

At this stage, topic modeling is performed using Latent Dirichlet Allocation (LDA) to identify the main topics that frequently appear in user reviews with negative sentiment. A total of 890 negative review data previously obtained from NLP sentiment analysis modeling using ROBERTa and IndoBert were reprocessed to identify the dominant complaint pattern. The LDA process is applied to texts that have been cleaned, tokenized, and assembled into a corpus.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Based on the processing of negative sentiment reviews, the LDA model succeeded in forming several of topics that represented groups of user complaints. Each topic is represented based on the *keywords* that appear most frequently and are associated with the contextual meaning of the existing review. The measurement results using a coherence score with 15 topics are 0.4204, and taking 15 topics is also in line with the minimum number of constructs for the questionnaire, with 5 indicators measured: a minimum of 2-3 questions or a minimum of 15 questions [28]. From the LDA topics, 15 topic were initially grouped into high-level thematic categories to facilitate interpretation of the analysis results. The results are summarized in Table 11.

Table 11. LDA Topic Modelling

LDA Topic (Keywords)	Topic Complaints	Total
iklan, ganggu, pulsa	*Masukan Diabaikan*	19
buruk,kualitas,lemah	*UX buruk/Eror aplikasi*	22
gua, email, gak	*Keterbatasan Bantuan Interaktif*	23
pulsa, berkurang, hilang	*Akurasi Data Diragukan*	27
notif, banget, masuk	*Notifikasi Berlebihan/Mengganggu*	29
sinyal, udah, jelek	*Keandalan Jaringan Rendah*	32
ngelag, min, aplikasi	*Navigasi Aplikasi Buruk*	33
sinyal, gak, mengecewakan	*Pemberitahuan Gangguan*	35
beli, kuota, gak	*Kegagalan Transaksi/Beli*	35
tolong, indosat,beli	*Resolusi Masalah Lambat*	37
ngeleg, lemot, jarigannya	*Keterlambatan/Lag Sistem*	44
parah, im3,main	*Kurang Transparansi Status*	45
beli, buruk, update	*Pembaruan Tidak Efektif*	49
lemot, data, hilang	Kualitas Dukungan Buruk	52
mahal,jelek,banget	Harga/promo Tidak Sesuai	138

Overall, there are some important findings regarding customer complaint patterns:

1. The most discussed topic was "The price/promo is not as needed" (138 comments) related to user preferences, which were dominated by keywords such as *expensive*, *ugly*, and *really*. This topic shows that most customers feel that the price of the service is not worth the quality received. This complaint is

an indication of fundamental dissatisfaction with the value proposition of the service from the application.

2. The next topics that came up a lot were "Poor Support Quality" (52 comments) and "Ineffective updates" (49 comments). Many reviews mention issues such as *slow, data loss, poor, and updating,* which indicate that users are experiencing technical bottlenecks and find the service provider's response unhelpful. This reflects the gap in the operational and customer service sides.
3. Complaints about system lag and lag are also quite significant, shown by the *topic System Lag* (44 comments). Keywords such as *slow*, *slow*, and *network* indicate an unstable application and network experience.
4. Some other topics have to do with transaction failures, slow resolution of issues, and excessive notifications. These topics point out various aspects of the user experience that interfere with and decrease satisfaction, both on the technical and non-technical sides.
5. On some topics such as *Doubtful Data Accuracy* or *Limitations of Interactive Help*, users highlight data reliability issues, loss of credit, or lack of clarity in interactive help. This indicates that there is a gap between the expectations and the digital experience of users.

In general, the results of LDA show that user complaints not only focus on network issues, but also include factors such as price, service, data accuracy, and the effectiveness of system updates. This information can be the basis for companies to make more targeted improvements, especially on the aspects of service that trigger the most dissatisfaction. Thus, LDA analysis provides a comprehensive picture of the key issues felt by customers, allowing for more targeted service development strategies and user experience improvements. The analysis of LDA also plays matters role in the development of the questionnaire, because the identified complaint topics can be used as a reference to formulate more detailed and relevant question items. The information is then used as a basis for secondary data collection through CSI and IPA, So that the instrument compiled is able to capture the source of user dissatisfaction more comprehensively.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

3.5. Questionnaire Formation

From 190 respondents, they were given 30 questions divided into 2 indicator assessments, namely importance and performance which have 5 dimensions of SERVQUAL, namely Responsiveness, Empathy, Reliability, Tangibles, Assurance. Each of these dimensions has 3 questions formed from the modelling topic, the related question items are presented in the Table 12.

Table 12. SERVQUAL Question

Dimension	Attribute Dimension Questions	Attribute Code	LDA Topics
Tangibles	The MyIM3 application interface design is attractive, modern, and consistent.	T1	*buruk,kualitas,lemah*
	The notification and message displays in the application do not disturb the user experience.	T2	*notif, banget, masuk*
	The menu layout and icons in the MyIM3 application are easy to understand and not confusing.	T3	*ngelag, min, aplikasi*
Reliability	The MyIM3 app rarely experiences errors, lag, or disruptions during use.	R1	*sinyal, udah, jelek*
	Quota, credit, and transaction information are always accurate and updated in real-time.	R2	*pulsa, berkurang, hilang*
	The application system rarely experiences delays when displaying usage data.	R3	*ngeleg, lemot, jarigannya*
Responsiveness	MyIM3 customer service team responds quickly and resolves user complaints.	RS1	*tolong, indosat,beli*
	The MyIM3 application provides clear notifications when there are system disruptions or maintenance.	RS2	*sinyal, gak,mengecewakan*

Dimension	Attribute Dimension Questions	Attribute Code	LDA Topics
	The MyIM3 app regularly updates its system to address user issues.	RS3	*beli, buruk, update*
Assurance	Purchasing internet packages and credit top-ups through the MyIM3 app is safe and trustworthy.	A1	*buka, beli, kuota, gak*
	The MyIM3 application keeps users' personal data confidential well.	A2	*lemot, data, hilang*
	The MyIM3 application always provides clear information about the transaction status (successful, failed, or pending).	A3	*parah, im3,udah*
Empathy	The MyIM3 application provides interactive guides and assistance when users encounter problems.	E1	*gua, email, gak*
	The application pays attention to user input from feedback or sentiments in the Google Play Store reviews and improves the service based on that input.	E2	*iklan, ganggu, pulsa*
	Promotions and new features are tailored to the needs and preferences of users identified from user feedback.	E3	*mahal,jelek,banget*

3.6. Instrument Test

The questionnaire in this study is divided into 2 parts. The first part concerns user demographics, including questions about gender, age, occupation, residence, and duration of app usage. The second part pertains to service quality (SERVQUAL), which contains of tangibles, reliability, responsiveness, assurance, and empathy. These are measured using a Likert scale (1-5) to assess the importance and performance of each feature considered for development. After 30 questions were distributed to 190 respondents, it was found that all questions were valid because the calculated r-value (performance and importance) was greater than the table r (0.1417, n=190), as shown in Table 13.

Table 13. Validity Test Result

Atributte	r-table	r-value Performance	r-value Importance	Description Performance	Description Importance
T1	0,1417	0,711	0,700	Valid	Valid
T2	0,1417	0,696	0,650	Valid	Valid
T3	0,1417	0,766	0,724	Valid	Valid
R1	0,1417	0,656	0,825	Valid	Valid
R2	0,1417	0,696	0,833	Valid	Valid
R3	0,1417	0,741	0,839	Valid	Valid
RS1	0,1417	0,835	0,833	Valid	Valid
RS2	0,1417	0,805	0,803	Valid	Valid
RS3	0,1417	0,773	0,784	Valid	Valid
A1	0,1417	0,788	0,803	Valid	Valid
A2	0,1417	0,729	0,835	Valid	Valid
A3	0,1417	0,708	0,782	Valid	Valid
E1	0,1417	0,807	0,757	Valid	Valid
E2	0,1417	0,817	0,758	Valid	Valid
E3	0,1417	0,814	0,760	Valid	Valid

Then, for the reliability test Table 14 shows that the expectations and perceptions of users are already in the very reliable category because the Cronbach's Alpha (α) value is greater than 0.80 for each SERVQUAL dimension.

Table 14. Reability Test

Dimension	Cronbach_Alpha_ Performance	Cronbach_Alpha_ Importance	Description
Tangibles	0,851	0,830	Reliable
Reliability	0,831	0,917	Reliable
Responsiveness	0,900	0,902	Reliable
Assurance	0,864	0,902	Reliable

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Dimension	Cronbach_Alpha_ Performance	Cronbach_Alpha_ Importance	Description
Empathy	0,905	0,872	Reliable

3.7. User Satisfaction Measurement

After conducting validity and reliability tests on the questionnaire instrument and obtaining valid and reliable results, the next stage will be to evaluate user satisfaction with the My IM3 app by comparing and measuring the gap between user expectations (MI) and user perceptions (MP). Table 15 shows the results of the gap analysis in the service quality of My IM3 application.

Table 15. Gap Analysis

Dimension	Item	MI	MP	Gap	Sort
Tangibles	T1	4,084	4,147	0,063	15
Tangibles	T2	4,205	3,768	-0,437	2
Tangibles	T3	4,200	3,932	-0,268	9
Reliability	R1	4,211	3,768	-0,442	1
Reliability	R2	4,279	4,163	-0,116	13
Reliability	R3	4,242	3,989	-0,253	10
Responsiveness	RS1	4,232	3,837	-0,395	4
Responsiveness	RS2	4,211	3,847	-0,363	6
Responsiveness	RS3	4,347	3,958	-0,389	5
Assurance	A1	4,305	4,105	-0,200	11
Assurance	A2	4,284	4,221	-0,063	14
Assurance	A3	4,242	4,074	-0,168	12
Empathy	E1	4,211	3,868	-0,342	7
Empathy	E2	4,247	3,926	-0,321	8
Empathy	E3	4,237	3,837	-0,400	3
	Total	4,2358	3,962666667	-0,27293333	

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

After obtaining the results of the gap analysis per dimension of the question, the next stage can be measured overall user satisfaction using CSI (Customer Satisfaction Index) analysis, the results of the CSI analysis are presented in the Table 16.

Table 16. Customer Satisfaction Index Result

Dimension	MI	MP	WF	WS	CSI%
Responsiveness	4.263	3.881	0.201	0.780	15.60
Empathy	4.232	3.877	0.199	0.773	15.46
Reliability	4.244	3.973	0.200	0.794	15.89
Tangibles	4.163	3.949	0.197	0.777	15.54
Assurance	4.277	4.133	0.202	0.836	16.73
			Total	3,96	79,22%

Based on the results of the Customer Satisfaction Index (CSI) calculation for the five SERVQUAL dimensions, a CSI value of 79.22% was obtained. This value indicates that users' satisfaction level with the mobile customer service application falls into the 'somewhat satisfied' category, with a 'borderline' position. This means users feel that the application service has sufficiently met their expectations, but there are still some aspects that need improvement to achieve a higher level of satisfaction. The mapping of which aspects need to be improved to reach greater satisfaction can be seen in Table 17.

Table 17. CSI Aspects Result

Dimension	Item	MI	MP	WF	WS	CSI_%	Sort
Tangibles	T1	4,084	4,147	0,064	0,267	5,331	10
Tangibles	T2	4,205	3,768	0,066	0,249	4,987	1
Tangibles	T3	4,200	3,932	0,066	0,260	5,198	7
Reliability	R1	4,211	3,768	0,066	0,250	4,995	2
Reliability	R2	4,279	4,163	0,067	0,280	5,607	14
Reliability	R3	4,242	3,989	0,067	0,266	5,326	9
Responsiveness	RS1	4,232	3,837	0,067	0,256	5,111	4
Responsiveness	RS2	4,211	3,847	0,066	0,255	5,099	3
Responsiveness	RS3	4,347	3,958	0,068	0,271	5,416	11

Dimension	Item	MI	MP	WF	WS	CSI_%	Sort
Assurance	A1	4,305	4,105	0,068	0,278	5,563	13
Assurance	A2	4,284	4,221	0,067	0,285	5,692	15
Assurance	A3	4,242	4,074	0,067	0,272	5,440	12
Empathy	E1	4,211	3,868	0,066	0,256	5,127	6
Empathy	E2	4,247	3,926	0,067	0,262	5,249	8
Empathy	E3	4,237	3,837	0,067	0,256	5,117	5
				Total	0,264	79,22	

Based on the assessment conducted by comparing user expectations and perceptions using CSI weighting, the lowest scores are found in item T2, which relates to notifications and messages in the application, followed by item R1, which pertains to errors and lag in the application, and then item RS2, which is related to system disruptions and maintenance. This indicates that many users complain that the performance of these items does not meet their expectations. Other calculation results show that the Assurance dimension obtained the highest score compared to other SERVQUAL dimensions. These findings indicate that MyIM3 application users have a strong level of trust in the security, reliability of information, and the competence of the services provided by the application.

To determine the urgency of feature updates in the application, it is not only necessary to perform CSI analysis comparing user expectations and perceptions but also to map it using IPA (Importance-Performance Analysis) to identify whether the underperforming features are truly urgent or highly needed for immediate development. In this way, the analysis can ultimately provide input regarding application development, focusing on features that are not only frequently complained about because they do not meet expectations and are not yet optimal but also are urgent or important to develop quickly. If development is not carried out promptly, it could disrupt the application's functionality. Below are the mapping results from the IPA of the MyIM3 application.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

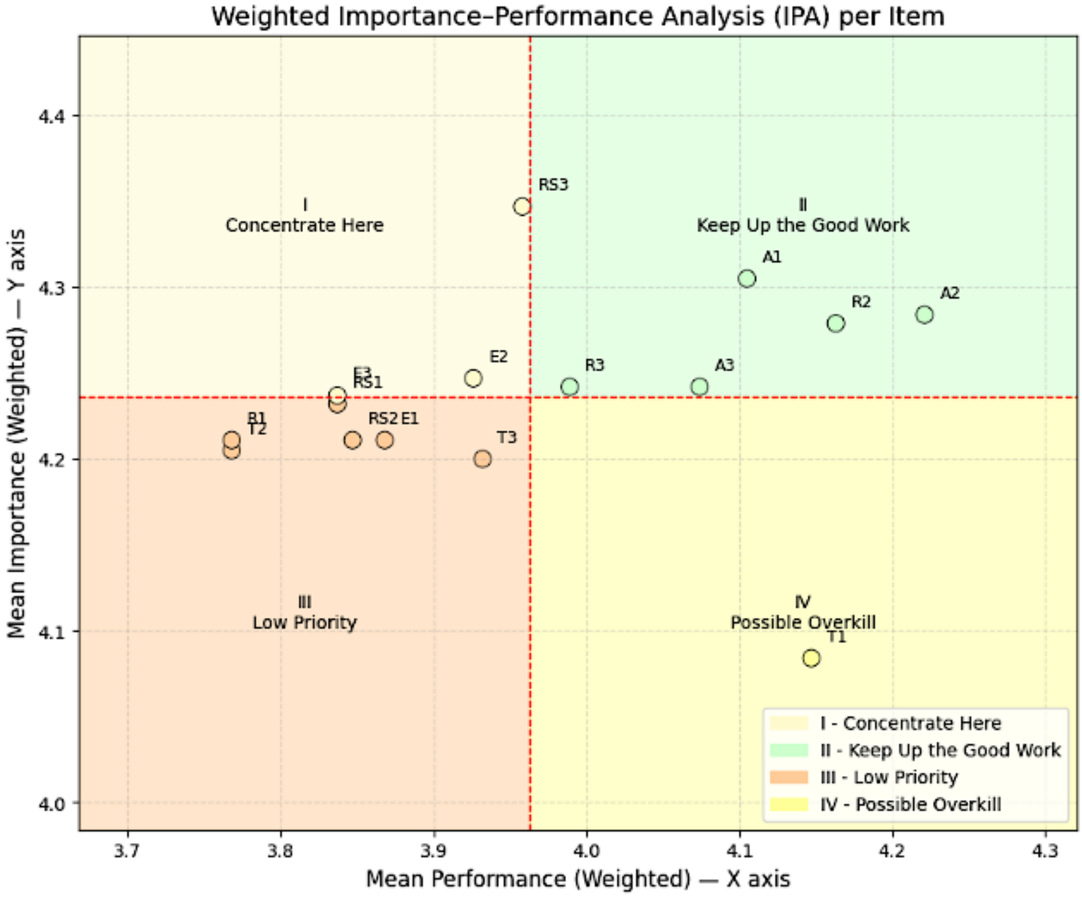

Figure 3. Importance-Performance Analysis

From the results of the science mapping conducted on several question items with matriks IPA, the following results were obtained:

Table 18. Quadrant 1 (Concentrate Here)

Item	Questions
E2	The application pays attention to user input from feedback or sentiments in Google Play Store reviews and improves the service based on that input.
E3	Promotions and new features are tailored to the needs and preferences of users identified from user feedback.
RS3	The MyIM3 app regularly updates its system to address user issues.The MyIM3 app regularly updates its system to address user issues.

Based on Table 18, items included in this quadrant are RS3 (System Update), E2 (Application attention to personal user needs), and E3 (User personalization in terms of features and pricing). These three items have a high level of importance but their performance is still below average. This indicates that users consider responsiveness and empathy to be very important, but these aspects have not been well realized in the application. The recommended strategy is to strengthen user engagement through reliable, quick response systems, the development of AI-based intelligent chatbots, and personalized service features based on user behavior and preferences.

Table 19. Quadrant II (Keep Up The Good Work)

Item	Questions
A1	Purchasing internet packages and credit top-ups through the MyIM3 app is safe and trustworthy.
A2	The MyIM3 application keeps users' personal data confidential well.
A3	The MyIM3 application always provides clear information regarding the transaction status (successful, failed, or pending).
R2	Quota, credit, and transaction information are always accurate and updated in real-time.
R3	The application system rarely experiences delays when displaying usage data.

Table 19 shows that items included in this quadrant are R2 (Application service accuracy), R3 (Application data display speed), A1 (User data security), A2 (Confidentiality of user information), and A3 (Service professionalism). These attributes have equally high levels of importance and performance, indicating that users are sufficiently satisfied with aspects of reliability and security assurance. The company needs to maintain this high performance through improvements in system reliability, data protection, and service consistency to keep user trust intact.

Items included in this quadrant shows in Table 20 are T2, T3, R1, RS1, RS2, and E1. These attributes have both low importance and performance values, so they are not a top priority for development at this time. However, gradual improvements are still recommended because users are not yet satisfied with the service aspects, such as improving the application's navigation display (user experience), speeding up initial response to complaints, and addressing bugs or application disruptions to support a professional image of the app.

Table 20. Quadrant III (Low Priority)

Item	Questions
T2	The notification and message display in the application do not disturb the user experience.
T3	The menu layout and icons in the MyIM3 application are easy to understand and not confusing.
R1	The MyIM3 application rarely experiences errors, lag, or disruptions when used.
RS1	MyIM3 customer service team responds quickly and resolves user complaints.
RS2	The MyIM3 application provides clear notifications when there are system disruptions or maintenance.
E1	The MyIM3 application provides interactive guidance and assistance when users encounter problems.

Table 21. Quadrant IV (Possible Overkill)

Item	Questions
T1	The MyIM3 application interface design is attractive, modern, and consistent.

In Table 21, Item T1 (Application display and feature completeness) falls into this quadrant, where its performance is already high but its level of importance is relatively low. This means, based on the calculation of importance and performance presented in the IPA quadrant, users are already satisfied with the application's appearance aspects, and it is not an urgent area for improvement based on the user perspective, so the company does not need to allocate too many resources for enhancements in this area. The main focus should be shifted to attributes in Quadrants I and III, which have a greater impact on user satisfaction. From the results of mapping the IPA quadrant, it can be seen that each quadrant has its list of items, here is the conclusion of the IPA quadrant which can be seen in table 22

Table 22. Summary IPA

Quadrant	Interpretation	Item
Quadrant I (Consentrate Here)	Low performance, high importance	E2,E3,RS3

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Quadrant	Interpretation	Item
Quadrant II (Keep Up The Good Work)	High performance, high importance	A1, A2, A3, R2, R3
Quadrant III (Low Priority)	Low performance, low importance	T2, T3, R1, RS1, RS2, E1
Quadrant IV (Possible Overkill)	High performance, low importance	T1

3.8. Development Recommendations

Based on the CSI analysis results and mapping conducted on several application feature items by IPA, there are 3 items with high urgency but still relatively low performance. Therefore, the researcher provides several development recommendations to improve the application features so they can be better and address some user complaints. Here are some recommended development features provided in Table 23.

Table 23. Development Recommendations

No	Recommend Features	Main Objectives	Development Strategy	Impact
1	Automatic Service Personalization	Enhancing the application's empathy towards user needs (E2, E3).	Develop a machine learning-based recommendation system to customize promotions, prices, and service suggestions according to user preferences.	Users feel noticed and served personally.
2	Smart Virtual Assistant (AI Customer Assistant)	Improve response speed and service suitability (RS3).	Upgrading the chatbot based on NLP deep learning with integration to a fast and interactive customer database and an option to hand over to a human customer service representative.	Faster responses, more accurate answers, more satisfied users.

No	Recommend Features	Main Objectives	Development Strategy	Impact
3	Quick Feedback & Adaptive Service	Bridging the gap between high interest and low performance.	Provide micro-surveys after service interactions and implement *auto-learning feedback loops* for feature improvements.	Improved services based on direct user feedback.

Also, to strengthen application development in a more strategic and structured direction, the researcher added prioritizing matrix as a suggestion for future application development, prioritizing matrix can be seen in table 24.

Table 24. Prioritizing Matrix

Priority	Recommendation	Reason
Short Term (0–3 Months)	Price and promo adjustments according to user behavior and the use of historical data	The highest complaints in negative reviews & in Quadrant I
Medium Term (3–6 Months)	Responsiveness improvements, reduced pop-ups, bug-fixing applications and transactions	Regarding user experience and significant topics in LDA
Long Term (6–12 Months)	Personalize the service, improve server stability, perform system updates based on user feedback	Strategic attributes that increase business value and user retention

Apart from some of the recommendations given, the risks of the implementation of changes also need to be considered, for that the researcher also provides some of the risks of implementation and its mitigation which can be seen in table 25.

Table 25. Risk Mitigation

Risk	Category	Risk	Strategy
Privacy Risks in Personalization Tools	Data Privacy	High	Use privacy-by-design, implement anonymization,

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Risk	Category	Risk	Strategy
			ensure compliance with the PDP Law 2022
Decreased App Performance after the addition of new features	Technology / Performance	High	Perform profiling and code optimization, implement modular architecture, use A/B testing before full launch
NLP Classification Errors on chatbots/automated help	NLP Reliability	Medium	Expand the training dataset, implement human-in-the-loop for ambiguous cases, regularly fine-tune the model
Service Interruptions During Application Updates	Operational	Medium	Schedule maintenance at night, provide a rollback system, and clear disturbance notifications.

3.9. Discussion

The results of this study underscore the effectiveness of integrating IndoBERT-based Aspect-Based Sentiment Analysis (ABSA), LDA Topic Modeling, and the CSI–IPA framework as a holistic methodology for understanding user experience and identifying actionable improvement areas in digital service applications—far surpassing the diagnostic capability of traditional questionnaire-based methods alone.

The sentiment analysis using IndoBERT, enhanced with pseudo-labeling techniques and confidence threshold calibration, successfully extracted nuanced user sentiment from over 10,000 real-world reviews of the MyIM3 application. This stage revealed widespread negative sentiment clusters concentrated around price fairness, customer service responsiveness, and network instability, capturing spontaneous user frustrations in natural usage contexts—data that are typically lost in structured surveys.

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Subsequently, LDA Topic Modeling distilled these sentiments into interpretable, high-frequency complaint themes, such as ineffective promotions, system lag, failed transactions, and intrusive notifications. These findings illuminate deeper structural issues not captured by sentiment polarity alone. For instance, while the review might register as negative, LDA shows *why*—whether it's a confusing UI, unresponsive assistance, or a mismatch between price and perceived value. The granularity provided by LDA allows for the development of targeted indicators that accurately reflect lived user experiences.

When this unstructured insight is mapped onto structured evaluation tools like CSI and IPA, the analysis gains strategic relevance. CSI quantifies user satisfaction across five SERVQUAL dimensions—Tangibles, Reliability, Responsiveness, Assurance, and Empathy—while IPA identifies which service aspects are perceived as *important* but underperforming. Notably, Assurance emerged as the strongest-performing dimension, suggesting that users generally trust the platform's security, privacy, and transaction integrity—a positive reflection of MyIM3's foundational service design. This aligns with expectations for mature digital platforms in regulated sectors such as telecommunications.

However, significant gaps were identified in Responsiveness and Empathy, particularly regarding the timeliness and relevance of customer support, system updates, and the platform's ability to adapt to user feedback. These gaps were further validated by their placement in Quadrant I of the IPA matrix (Concentrate Here)—highlighting them as both critically important and currently underperforming.

This triangulated methodology demonstrates how each analytical tool contributes distinct yet synergistic insights:

1. IndoBERT-based sentiment analysis captures the emotional tone and volume of user feedback at scale.
2. LDA modeling decodes the structural and thematic patterns behind user dissatisfaction.
3. CSI–IPA mapping quantifies these insights into measurable gaps and strategic priorities.

Together, this integrated framework supports a data-driven, user-centric development strategy—one that moves beyond intuition or anecdotal evidence. It enables digital service providers to prioritize development efforts on attributes that most significantly influence satisfaction, loyalty, and long-term user retention. In particular, the mapping of features such as RS3 (system updates), E2 (user input responsiveness), and E3 (personalized promotions) into the "Concentrate Here" quadrant indicates an urgent need to shift from a one-size-fits-all model to a hyper-personalized, agile service delivery model. This can be achieved through machine learning-based personalization engines, intelligent chatbots, and real-time feedback loops—all of which were proposed in the development recommendations section.

Furthermore, this study contributes methodologically by offering a novel integration of NLP with established service quality models. Unlike earlier research relying on traditional classifiers like Naïve Bayes or SVM, this approach leverages transformer-based language models fine-tuned on contextual user data, resulting in higher prediction accuracy and more relevant service insights. It also bridges the gap between quantitative performance metrics and qualitative user feedback, providing a richer, more actionable understanding of digital service quality. This research confirms that the synergy between advanced NLP techniques and service quality frameworks offers a powerful tool for adaptive service evaluation and improvement. In a dynamic and competitive digital environment, such an approach equips organizations with the insight needed to respond to evolving user expectations and deliver consistently high-quality experiences.

4. CONCLUSION

This study shows that integrating NLP techniques, specifically IndoBERT fine-tuned sentiment and LDA topic modelling with the CSI–IPA framework, provides a systematic way to identify user-needed service attributes for digital telecommunication applications. Instead of relying solely on theoretical SERVQUAL dimensions, the research extracts negative sentiment reviews, identifies the dominant topic through LDA, and uses these topics as empirical foundations for constructing more relevant SERVQUAL questionnaire indicators. The resulting CSI score of 79.22% places user satisfaction in the “borderline” category. At the same time, IPA results reveal that most indicators fall into Quadrants II and III, suggesting adequate but not yet optimized performance. Meanwhile, attributes

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

related to system updates, personal needs, and personalization lie in Quadrant I and require immediate improvement. Overall, the study emphasizes the importance of enhancing responsiveness and empathy while maintaining reliability and assurance to create a more adaptive and user-centred digital service experience.

Theoretically, this research contributes to the digital service quality literature by proposing an integrated pipeline for NLP negative sentiment extraction, topic modelling, and SERVQUAL indicator construction. This hybrid pipeline offers a novel, data-driven mechanism for deriving service attributes directly from user-generated content, reducing reliance on purely conceptual SERVQUAL assumptions. The mapping of data-driven topics to SERVQUAL dimensions illustrates a methodological advancement that future researchers can use to refine or localize service quality measurements for digital platforms. This approach also demonstrates how textual feedback can be systematically transformed into measurable constructs for satisfaction evaluation.

Practically, the findings equip telecommunication service providers with a clearer understanding of what users genuinely prioritize based on real-world complaints. The identification of urgent topics system responsiveness, system updates, personalization, and attention to individual user-needs provides clear directions for improving the MyIM3 application's service quality. The proposed pipeline can be adopted as a continuous monitoring tool to detect emerging user issues, inform product development decisions, and support strategic planning. Developers are advised to enhance responsiveness through real-time support features, improve personalization mechanisms, maintain stability and assurance in core services, and regularly evaluate changes using CSI–IPA to ensure alignment with user expectations.

Future studies may enhance topic extraction by exploring models such as BERTopic or LLM-driven clustering to reduce topic overlap and increase interpretability. Researchers can also integrate multi-platform datasets such as Twitter or Instagram complaints, in-app feedback, and customer service records to build a more holistic understanding of user needs. Longitudinal studies are also recommended to evaluate how CSI–IPA scores and user topic patterns evolve following system improvements. Finally, applying this pipeline across different telecom applications or digital platforms may strengthen the generalizability of the framework and highlight industry-wide service quality trends.

REFERENCES

[1] F. Saputra, Y. Lake, E. T. Taena, [2]universitas Timor, and [3]universitas Timor, "Pengaruh Kualitas Pelayanan Terhadap Kepuasan Pelanggan Telekomunikasi PT. Telkom Indonesia," *Jurnal Manajemen Bisnis Dan Organisasi (JMBO)*, vol. Vol 3 No 2, 2024, doi: 10.58290/jmbo.v3i2.327.

[2] Y. Yudiana, A. Y. Agustina, and N. Khofifah, "Prediksi Customer Churn Menggunakan Metode CRISP-DM Pada Industri Telekomunikasi Sebagai Implementasi Mempertahankan Pelanggan," *Indonesian Journal of Islamic Economics and Business*, vol. 8, no. 1, pp. 1–20, 2023, doi: 10.30631/ijoieb.v8i1.1710.

[3] O. S. Ighomereho, T. S. Afolabi, and A. O. Oluwakoya, "Impact of E-service quality on customer satisfaction: a study of internet banking for general and maritime services in Nigeria," *Journal of Financial Services Marketing*, vol. 28, no. 3, pp. 488–501, Sep. 2023, doi: 10.1057/s41264-022-00164-x.

[4] S. Wahono and H. Ali, "Peranan Data Warehouse, Software Dan Brainware Terhadap Pengambilan Keputusan (Literature Review Executive Support Sistem for Business)," *Jurnal Ekonomi Manajemen Sistem Informasi*, vol. 3, no. 2, pp. 225–239, 2021, doi: 10.31933/jemsi.v3i1.

[5] A. Syaifullah, "Analisis Perbandingan Pengalaman Pengguna Dan Kualitas Layanan Pada Aplikasi Dana Dan Ovo Menggunakan Metode Ueq Dan Ipa," *Jurnal Teknologi Sistem Informasi dan Aplikasi*, 2024, vol. 7, no. 3, pp.1128-1139, 2024, doi: 10.32493/jtsi.v7i3.40512.

[6] T. L. Windyaningrum and C. D. Indrawati, "Integrasi Metode Service Performance (SERVPERF), Importance Performance Analysis (IPA) Dan Customer Satisfaction Index (CSI) Dalam Penentuan Prioritas Pengembangan Lembaga Pendidikan Nonformal," *JUSTER: Jurnal Sains Dan Terapan*, vol. 2, no. 1, pp. 112–121, 2023, doi : 10.57218/juster.v2i1.522.

[7] A. R. Maulani, T. Wahyudi, and R. Rahmawati, "Pengukuran Tingkat Kepuasan Konsumen Aplikasi Mypertamina Dengan Metode Servqual, CSI, Dan IPA," *INTEGRATE: Industrial Engineering and Management System*, vol. 7, no. 3, pp. 19–25, 2023.

[8] A. N. Febrianto, "Analysis of Google Play Store Reviews Using Natural Language Processing (NLP) and Importance Performance Analysis (IPA) for the Development of Mobile Banking Application Businesses Wondr BNI," *Enrichment: Journal of Multidisciplinary Research and Development*, vol. 3, no. 6, pp. 1198–1212, 2025, doi : 10.55324/enrichment.v3i6.501.

[9] R. Marbun, "Analisis Sentimen Ulasan Pengguna Terhadap Aplikasi BRImo BRI pada Situs Google Play Store Menggunakan Metode Naive Bayes dan Customer Satisfaction Index (CSI)", Diss, Universitas Lancang Kuning, 2024.

[10] R. Rosmaladewi, "Pengukuran Kepuasan Pelanggan Lazada Berdasarkan Opinion Mining Menggunakan Algoritma Support Vector Machine, Customer Satisfaction Index dan Importance Performance Analysis," Tesis, Universitas Hasanuddin, Makassar, Indonesia, 2023.

[11] A. R. Fadilla and P. A. Wulandari, "Literature review analisis data kualitatif: tahap pengumpulan data," *Mitita Jurnal Penelitian*, vol. 1, no. 3, pp. 34–46, 2023.

[12] A. Kusumastuti, A. M. Khoiron, and T. A. Achmad, *Metode penelitian kuantitatif*. Deepublish, 2021.

[13] E. P. Setiani, "The impact of ESG scores on corporate financial performance: Moderating role of gender diversity," *Nominal: Barometer Riset Akuntansi dan Manajemen*, vol. 12, no. 1, pp. 128–139, 2023, doi: 10.21831/nominal.v12i1.59778

[14] H. S. : J. S. I. Halomoan Limbong, "Optimasi Analisis Sentimen Ulasan Aplikasi Amikom One Menggunakan SMOTE pada Algoritma Artificial Neural Network Optimization of Sentiment Analysis for Amikom One Application Reviews Using SMOTE with Artificial Neural Network Algorithm," *Sistemasi : Jurnal Sistem Informasi*, vol. 13, no. 5, pp. 2540–9719, 2024.

[15] K. C. Astuti, A. Firmansyah, and A. Riyadi, "Implementasi text mining untuk analisis sentimen masyarakat terhadap ulasan aplikasi digital Korlantas Polri pada Google Play Store," *Remik: Riset dan E-Jurnal Manajemen Informatika Komputer*, vol. 8, no. 1, pp. 383–394, 2024, doi: 10.33395/remik.v8i1.13421.

[16] A. S. Aribowo, H. Basiron, N. F. A. Yusof, and S. Khomsah, "Cross-domain sentiment analysis model on indonesian youtube comment," *International Journal of Advances in Intelligent Informatics*, vol. 7, no. 1, pp. 12–25, 2021, doi: 10.26555/ijain.v7i1.554.

[17] Z. Abidin, A. Junaidi, and Wamiliana, "Text Stemming and Lemmatization of Regional Languages in Indonesia: A Systematic Literature Review," *Journal of Information Systems Engineering and Business Intelligence*, vol. 10, no. 2, pp. 217–231, Jun. 2024, doi: 10.20473/jisebi.10.2.217-231.

[18] Rianto, A. B. Mutiara, E. P. Wibowo, and P. I. Santosa, "Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation," *J Big Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-021-00413-1.

[19] E. B. Satriawan, S. H. Wijoyo, and D. E. Ratnawati, "Analisis Sentimen Terhadap Pendapat Masyarakat Mengenai Pilkada 2024 Menggunakan Metode Support Vector Machine," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 12, no. 4, pp. 809–818, 2025, doi: 10.25126/jtiik.124.

[20] F. R. Andhika, W. Witanti, and P. N. Sabrina, "Analisis Sentimen Menggunakan Metode IndoBERT Pada Ulasan Aplikasi Zoom Menggunakan Fitur Ekstrasi GloVe," *METIK JURNAL (AKREDITASI SINTA 3)*, vol. 9, no. 2, pp. 439–448, 2025, doi: 10.47002/g38fxb78.

[21] E. Arazo, D. Ortego, P. Albert, N. E. O'connor, and K. Mcguinness, "Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning," *nternational joint conference on neural networks (IJCNN)*, pp. 1–8, 2020.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[22] N. Cahyono, N. Nurcahyo, A. Fauzan, and R. Agung, "Analisa Perbandingan Latent Semantic Indexing (LSI) dan Latent Dirichlet Allocation (LDA) untuk Topic Modelling Aplikasi Identitas Kependudukan Digital (IKD)," *Building of Informatics, Technology and Science*, vol. 6, no. 3, pp. 1638–1647, 2024, doi: 10.47065/bits.v6i3.5970..

[23] F. R. Wilujeng, G. D. Rembulan, D. Andreas, and H. Tannady, "Meningkatkan kepuasan pelanggan pada dua bisnis e-commerce terbesar di Indonesia dengan menggunakan analisis Servqual dan IPA," *Prosiding Seminar Nasional Sains dan Teknologi (Semnastek)*, vol. 16, pp. 1–9, Nov. 2019.

[24] B. S. Silalahi and F. J. Kaunang, "Analisis Service Quality pada Aplikasi DANA berdasarkan Sudut Pandang Pelanggan di Daerah Bandung Barat Menggunakan Metode Servqual dan Model Kano," *TelKa: Jurnal Teknologi Informasi dan Komunikasi*, vol. 12, no. 02, pp. 121–133, Oct. 2022, doi: 10.36342/teika.v12i02.2957.

[25] M. Agustina, V. Sahfitri, and T. Astuti, "Analysis of Online Transportation User Satisfaction Using the Customer Satisfaction Index (CSI) and Important Performance Analysis (IPA) Methods in Palembang City," *Journal of Information Systems and Informatics*, vol. 4, no. 4, pp. 1063–1075, Dec. 2022, doi: 10.51519/journalisi.v4i4.396.

[26] M. M. Rahman, "Sample size determination for survey research and non-probability sampling techniques: A review and set of recommendations," *Journal of Entrepreneurship, Business and Economics*, vol. 11, no. 1, pp. 42–62, 2023.

[27] D. Rahmadani, M. L. Hamzah, and E. Saputra, "The Influence of E-Customer Service, Mobile Application Design, Brand Image, and Customer Satisfaction on Interest in Mobile Banking Adoption," in *2024 4th International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, IEEE, 2024, pp. 1–6, doi: 10.1109/esmarta62850.2024.10639012.

[28] J. F. Hair, G. T. M. Hult, C. M. Ringle, and Marko. Sarstedt, *A primer on partial least squares structural equation modeling (PLS-SEM)*. Sage, 2017.