

Improving Application Services Using Root Cause Analysis and User Sentiment Analysis

Lintang Iqhtiar Dwi Mawarni^{1*}, I Kadek Dwi Nuryana¹ , Ika Qutsiati Utami² , Nisa Dwi Septiyanti³

^{1*} Universitas Negeri Surabaya, Surabaya, East Java, 60231, Indonesia
lintang.21031@mhs.unesa.ac.id, dwinuryana@unesa.ac.id

² Universitas Airlangga, Surabaya, East Java, 60115, Indonesia
ika.q.utami@ftmm.unair.ac.id

³ National Central University, Zhongli, Taoyuan, 320, Taiwan
nisaseptiyanti@g.ncu.edu.tw

Abstract. Amid the wave of digital transformation, the Digital Korlantas Polri application emerged as a solution and breakthrough in the digital-based SIM (driver's license) issuance process. However, concerns about the application's performance quality remain, as evidenced by user comments on the Google Play Store. A disparity was found between the app's rating and the content of user reviews, raising questions about the application's actual quality. This study aims to identify the root causes of user issues with the Digital Korlantas Polri application and generate improvement recommendations based on the identified problems. The research utilizes the pre-trained IndoBERT model for sentiment classification, followed by semi-supervised topic modeling using Guided LDA to uncover hidden patterns in the review data. The sentiment classification model was adopted from a previously validated IndoBERT model with reported benchmark performance; the model uses a labeled dataset for each sentiment. Furthermore, the pre-trained GPT-2 model is employed as a text generator to produce application improvement recommendations based on the identified issues. Evaluation results show that the Guided LDA model reached a coherence score of 0.51, and the GPT-2 model yielded a perplexity value of 1.43. These findings demonstrate that the proposed integrated framework is effective in identifying critical service issues and transforming analytical insights into actionable recommendations, providing a scalable approach for enhancing user experience in digital public service applications.

Keywords: *IndoBERT; Guided LDA; GPT-2; Topic Modelling; Root Cause Analysis*

1. Introduction

Digital transformation has significantly reshaped various sectors, including public services such as healthcare, education, and government. Digital technologies are now widely adopted to improve the efficiency and quality of public services, including in the law enforcement sector. One notable implementation is the Digital Korlantas POLRI application, developed by the Indonesian National Police to facilitate online access to traffic-related services, such as the issuance and renewal of driving licenses. However, like many other public service applications, user responses toward this application

are mixed. Despite receiving relatively high ratings, many user reviews on the Google Play Store express dissatisfaction. This mismatch between rating scores and review content highlights the need for a more in-depth analysis of user perceptions and experiences. To address this gap, various computational approaches have been employed to analyze user feedback in public service applications. Sentiment analysis has been widely applied to classify user opinions into positive, negative, and neutral categories, enabling organizations to gauge overall user satisfaction ([Darwis et al., 2021](#); [Abrori, 2024](#)). However, these studies generally stop at polarity classification without identifying the underlying causes of dissatisfaction, limiting their utility for targeted service improvement.

Beyond sentiment classification, topic modeling techniques such as Latent Dirichlet Allocation (LDA) have been employed to summarize user feedback and extract recurring themes from large-scale review data ([Hyun et al., 2024](#); [Lee & Oh, 2024](#)). While LDA provides a useful overview of discussion topics, unsupervised LDA often produces less interpretable and less coherent topics due to its reliance on statistical co-occurrence alone, particularly when applied to noisy, informal user-generated text. This limitation has prompted the development of semi-supervised approaches such as Guided LDA, which incorporate seed words derived from domain knowledge to guide topic discovery toward more meaningful and actionable themes. Recent advances in large language models (LLMs) have introduced new possibilities for generating contextual recommendations based on analytical insights. Models such as GPT-2 and its successors have demonstrated promising capabilities in producing human-like text for various applications, including automated response generation and recommendation systems ([Naveed et al., 2025](#); [Zhao et al., 2024](#)). However, LLM-based recommendation systems have rarely been integrated with Root Cause Analysis (RCA) for digital public services. Most existing implementations treat sentiment and topic analysis as isolated components, rather than as part of an integrated framework that systematically diagnoses problems and generates solutions.

Previous studies have applied Root Cause Analysis using fishbone diagrams to map contributing factors to service failures in various domains, including healthcare and manufacturing ([Graber et al., 2024](#); [Park, 2021](#)). However, these applications typically rely on manual identification of causes, which can be time-consuming and subject to human bias. While some studies have explored sentiment-based RCA ([Park, 2021](#)), the integration of automated sentiment analysis, topic modeling, RCA, and LLM-based recommendation generation remains underexplored in the context of digital public services. Therefore, this study integrates IndoBERT for sentiment classification, Guided LDA for topic modeling, Root Cause Analysis using fishbone diagrams, and GPT-2 for recommendation generation into a unified analytical framework. The analysis is guided by the End User Computing Satisfaction (EUCS) framework, which provides five structured dimensions for evaluating service quality: content, accuracy, form, ease of use, and timeliness ([Halim, 2020](#)). By combining these approaches, this study aims to provide a comprehensive methodology for diagnosing user dissatisfaction and generating actionable service improvements.

1.1. State of the Art

To position this study within the existing literature, Table 1 summarizes prior research on sentiment analysis, topic modeling, Root Cause Analysis, and LLM-based recommendation systems in the context of digital service evaluation.

Table 1 Comparison of Previous Studies and the Proposed Approach

Study	Sentiment Analysis	Topic Modeling	RCA	LLM- based Recommendation	EUCS Framework	Aplication Domain
Darwis et al. (2021)	✓ (Naïve Bayes)	✗	✗	✗	✗	Social Media
Arum Abrori (2024)	✓ (SVM)	✓ (LDA)	✗	✗	✗	Restaurant
Hyun et al. (2024)	✗	✓ (LDA)	✗	✗	✗	Architecture
Halim (2020)	✗	✗	✗	✗	✓ (EUCS Analysis)	Mobile App
Park (2021)	✓	✗	✓ (Sentiment-based)	✗	✗	General
Graber et al. (2024)	✗	✗	✓ (Manual)	✗	✗	Healthcare Recommender
Zhao et al. (2024)	✗	✗	✗	✓ (LLM-based)	✗	Systems

As shown in Table 1, while previous studies have addressed individual components of the analytical pipeline, none have integrated all four approaches sentiment analysis, topic modeling, Root Cause Analysis, and LLM-based recommendation generation within a unified framework. Furthermore, applying this integrated approach to digital public service evaluation, particularly using the EUCS framework to structure topic discovery, represents a novel contribution to the literature.

1.2. Contributions of This Study

Based on the identified research gaps, this study makes the following contributions:

- Integration of IndoBERT and Guided LDA for Public Service Evaluation:** This study combines a pre-trained IndoBERT model for sentiment classification with semi-supervised Guided LDA for topic modeling. While previous studies have applied sentiment analysis or topic modeling separately, integrating both approaches enables a more comprehensive understanding of user dissatisfaction, moving beyond polarity classification to identify specific problem areas.
- EUCS-Guided Topic Discovery:** Unlike conventional unsupervised LDA, which relies solely on statistical co-occurrence, this study incorporates seed words derived from the End User Computing Satisfaction (EUCS) framework to guide topic modeling. This approach ensures that the extracted topics are aligned with established service quality dimensions, improving topic interpretability and practical relevance.
- Integration of Root Cause Analysis with Topic Modeling:** This study maps Guided LDA topics to RCA using fishbone diagrams, creating a structured diagnosis of service failures. By visualizing the relationships between contributing factors across multiple dimensions, this approach provides a systematic method for identifying root causes of user dissatisfaction.
- LLM-Based Recommendation Generation for Public Services:** This study fine-tunes a GPT-2 model on official response data to generate contextual service improvement recommendations. This represents a novel application of generative language models in the

digital public service domain, enabling automated and data-driven suggestion generation based on identified issues.

- e. **Prototype Implementation for Practical Deployment:** This study develops an integrated prototype that combines analytical dashboards with automated recommendation generation. This prototype bridges the gap between data-driven insights and practical service improvement, offering a scalable solution for enhancing user experience in digital government platforms.
- f. **Empirical Application to Digital Public Service:** While previous studies have examined user feedback in various domains, the application of this integrated framework to a digital public service application ([POLRI](#)) provides empirical evidence of its effectiveness in a high-stakes government service context, where user satisfaction is critical for public trust.

By addressing the limitations of prior research and integrating multiple analytical approaches, this study contributes a comprehensive framework for diagnosing and addressing user dissatisfaction in digital public services.

2. Methods

This research adopts a combination of CRISP-DM (Cross-Industry Standard Process for Data Mining) and LLM-OPS (Large Language Model--Operational Pipeline for Suggestions) methodologies. The integration of both approaches is designed to achieve two main objectives: (1) identifying the root causes of user dissatisfaction based on sentiment and topic analysis, and (2) generating improvement recommendations using a generative language model.

2.1. CRISP-DM (*Cross Industry Standard Process for Data Mining*)

The standard methodology used in this research is a data mining framework designed to guide the data analysis process from start to finish. Developed in the late 1990s, CRISP-DM (Cross Industry Standard Process for Data Mining) provides a flexible structure that helps data teams understand and manage each phase of a project in a more organized manner ([Schröer, 2021](#)).

CRISP-DM was selected as the primary analytical framework because it provides a structured, reproducible workflow that ensures methodological transparency, with each phase clearly defined from business understanding through deployment. As the de facto industry standard for data mining projects, CRISP-DM has been validated across numerous domains, and its iterative nature allows for refinement based on evaluation results, which is essential for optimizing topic modeling parameters.

This methodology consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The step-by-step of the method can be seen in Figure 1.

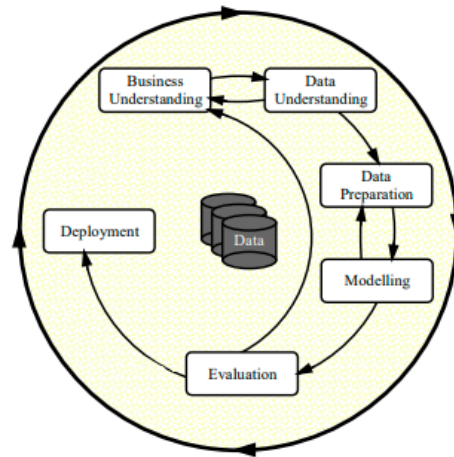


Figure 1 CRIPS-DM Method (Schröder, 2021)

Each phase in the CRISP-DM methodology is carried out as follows:

- a. **Business Understanding:** This is the first step to understand the objectives and project requirements from a business perspective. In this study, the goal is to identify user dissatisfaction issues within the Digital Korlantas POLRI application and to generate actionable recommendations for service improvement.
- b. **Data Understanding:** This phase involves collecting and exploring the dataset. The focus is on understanding the structure, patterns, and key characteristics of the data relevant to the problem, which in this case includes user reviews extracted from the Google Play Store using Google Play Scraper between May 2022 and March 2025. A total of 33,264 reviews were collected. The dataset was then sentiment-labeled using a pre-trained IndoBert model from Hugging Face, and the model was employed in a zero-shot inference setting without additional fine-tuning.
- c. **Data Preparation:** This step includes cleaning and preprocessing the data to ensure it is ready for modeling. Processes such as removing duplicates, handling missing values, case folding, stopword removal, and tokenization are applied to standardize and structure the text data.
- d. **Modeling:** In this phase, topic modeling is performed to uncover underlying topics from user reviews. The study compares two models, LDA (Latent Dirichlet Allocation) and Guided LDA, to determine the most effective model for identifying root causes. The best-performing model is selected based on coherence score evaluation.

Guided LDA was implemented using seven predefined topic groups derived from the EUCS framework and issues identified in the initial LDA results. The number of topics was selected by evaluating candidate models containing 2–10 topics using C_v coherence, UMass coherence, and log-likelihood. The final model was trained for 2,000 iterations with alpha and eta values of 0.01. A seed confidence value of 0.15 was used to control the influence of the predefined words on topic assignment.

Table 2 Guided LDA Configuration Parameters

Parameter	Value	Description
Number of topics (n_topics)	7	Number of latent topics generated by the final Guided LDA model
Number of iterations (n_iter)	2	Number of sampling iterations performed during model training
Document–topic prior (alpha)	0.01	Controls the sparsity of the topic distribution within each document
Topic–word prior (eta)	0.01	Controls the sparsity of the word distribution within each topic
Seed confidence	0.15	Determines the strength of the predefined seed words in guiding topic formation
Random state	42	Ensures reproducibility of the modeling results
Refresh interval	20	Frequency for displaying model-training progress
Candidate number of topics	2–10	Topic-number range evaluated during parameter selection
Topic evaluation metrics	(C_v) coherence, UMass coherence, and log-likelihood	Metrics used to assess topic quality and determine the final number of topics
Final topic selection	7 topics	Selected based on the comparison of topic coherence results

Table 3 EUCS-Based Seed Words Used in Guided LDA

Topic group	Related EUCS dimension	Seed words
Application features	Content	fitur, layanan, perpanjangan, pengajuan, pengembangan, sim, unduh
Payment and administrative process	Accuracy	biaya, proses, prosedur, pembayaran, tolak, ditolak, persyaratan, satpas, hasil
Application interface	Format	tampilan, desain, interface, bug, warna, layout, tema, grafis
Ease of application use	Ease of Use	akses, tombol, navigasi, bingung, sulit, kesulitan, mudah, user_friendly
Login process	Ease of Use	login, masuk, akun, username, password
Registration and verification	Ease of Use	verifikasi, daftar, registrasi, akun, otp, sms, email, masuk, nik, ktp
Processing time and system responsiveness	Timeliness	loading, lambat, cepat, timeout, hang, lelet, stuck, lag, responsif, crash

The seed words were derived from the five dimensions of End-User Computing Satisfaction: content, accuracy, format, ease of use, and timeliness. The ease-of-use dimension was divided into three more specific topic groups: general usability, login, and registration because the initial LDA results indicated that these issues appeared as distinct complaint patterns. Before being assigned to the Guided LDA model, seed words that were not present in the corpus vocabulary were removed.

- e. Evaluation: The evaluation process involves comparing the topic coherence values produced by LDA and Guided LDA. The model with the higher coherence score and more relevant topic distribution is selected for the next step.

- f. **Deployment:** The final phase involves saving the best model and analysis results for future use.

The adoption of CRISP-DM in this study provides a structured and reproducible workflow for handling large-scale unstructured text data. By following the framework's sequential phases, the research ensures that analytical decisions align with the study objectives and are supported by empirical evidence. This is particularly important when dealing with user-generated content, which often contains noise, informal language, and varying levels of detail.

Moreover, the iterative nature of CRISP-DM allows flexibility in refining models based on evaluation results. For instance, the topic modeling phase was iteratively adjusted by varying the number of topics and incorporating guided seed words derived from the EUCS framework. This iterative process enhances the interpretability and relevance of the extracted topics, which is essential for conducting meaningful root cause analysis. By embedding CRISP-DM within the broader analytical pipeline, this study ensures methodological transparency and supports the replicability of the proposed framework for similar public service applications.

2.2 LLM-OPS (Large Language Model Operational Pipeline for Suggestions)

To complement the CRISP-DM methodology, this study integrates the LLM-OPS approach to generate improvement recommendations based on the analysis results. LLM-OPS is a pipeline that operationalizes the use of Large Language Models (LLMs) such as GPT or T5 for generating human-like suggestions or summaries from structured analytical insights.

LLM-OPS was integrated to address the recommendation generation objective because it is specifically designed for operationalizing large language models in production-like environments, covering the entire lifecycle from data preparation through deployment with comprehensive evaluation stages. The integration of these methodologies ensures both analytical rigor in diagnosing user dissatisfaction and practical utility in generating automated, data-driven recommendations, bridging the gap between analytical insights and actionable service improvements. The step-by-step of the method can be seen in Figure 2.

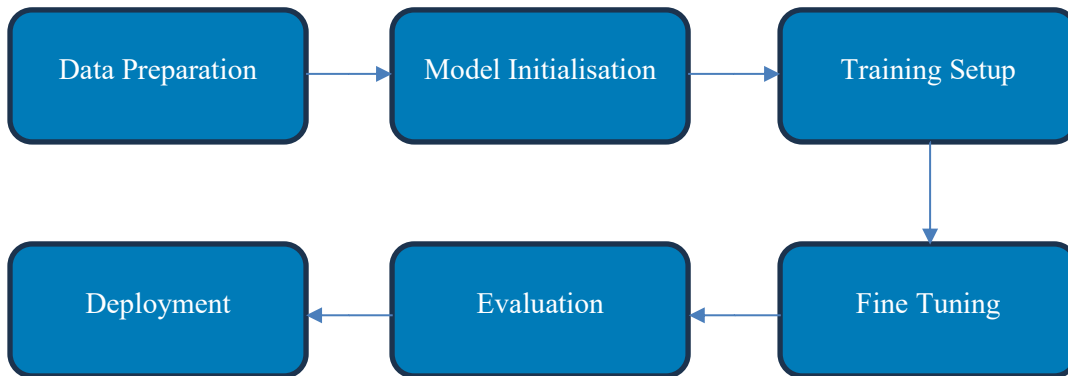


Figure 2 LLM OPS Method (Liu et al., 2023)

The LLM-OPS (Large Language Model Operational Pipeline for Suggestions) methodology is implemented in the following stages:

- Data Preparation:** The training dataset consists of user comments labeled with negative sentiment, specifically those that received replies from the official Digital Korlantas POLRI account. These comment-reply pairs were selected as training samples to teach the model how to generate appropriate service recommendations in response to negative feedback.
- Model Initialization:** The pretrained model is initialized using a GPT-2 architecture available on Hugging Face, which has been previously fine-tuned on Bahasa Indonesia text data. This model serves as the base for further fine-tuning in the context of service improvement responses.
- Training Setup:** Training arguments are defined, including batch size, learning rate, number of epochs, logging steps, and evaluation strategy. These hyperparameters are configured to ensure efficient and stable training performance.
- Fine-Tuning:** The fine-tuning process is conducted on the prepared dataset, using different data configurations to observe the model's response to various input types. The training is closely monitored through loss values and performance trends across epochs.

The recommendation model was developed by fine-tuning the Indonesian GPT-2 model `cahya/gpt2-small-indonesian-522M`. Each training instance combined a user complaint and the corresponding official response using the format “Keluhan: [review] Reply: [reply].” The text was tokenized to a maximum length of 128 tokens, while the model was trained for ten epochs using a batch size of four. The learning rate and optimizer were not manually specified; therefore, the default Hugging Face Trainer settings were used.

Table 4 GPT-2 Fine-Tuning Configuration Parameters

Parameter	Value	Description
Base model	<code>cahya/gpt2-small-indonesian-522M</code>	Pretrained Indonesian GPT-2 model obtained from Hugging Face
Model architecture	GPT-2 LM Head Model	Causal language model used for recommendation generation
Tokenizer	GPT-2 Tokenizer	Tokenizer associated with the pretrained model
Training input format	Keluhan: [Review] Reply: [Reply]	Text format combining the user complaint and official response
Maximum sequence length	128 tokens	Maximum token length for each training instance
Padding strategy	Maximum-length padding	Pads every input sequence to the specified maximum length
Truncation	Enabled	Truncates input sequences exceeding 128 tokens
Training batch size	4	Number of training samples processed per device in each step
Evaluation batch size	4	Number of evaluation samples processed per device in each step
Number of epochs	10	Number of complete passes through the training dataset
Learning rate	(5×10^{-5})	Default learning rate used by the Hugging Face TrainingArguments because no custom value was explicitly specified
Optimizer	AdamW	Default optimizer applied by the Hugging Face Trainer

Evaluation strategy	Every epoch	Evaluates the model after each training epoch
Model-saving strategy	Every epoch	Saves a model checkpoint after each epoch
Logging interval	Every 10 steps	Records training information every ten optimization steps
Language-modeling objective	Causal language modeling	Predicts subsequent tokens based on preceding context
Masked language modeling (mlm)	FALSE	Disabled because GPT-2 is an autoregressive causal language model
Mixed-precision training	Enabled when CUDA is available	Uses FP16 computation when supported by the GPU
Reporting integration	Disabled	No external experiment-tracking platform was used
Hugging Face Hub upload	Disabled	The fine-tuned model was not automatically uploaded to the Hub
Dataset split configurations evaluated	50:50, 70:30, 80:20, and 90:10	Training and test proportions compared during the experiment
Final model configuration	80:20:00	Selected based on the balance between automatic metrics and human evaluation

During inference, recommendation text was generated using stochastic decoding with top-k and nucleus sampling. The temperature was set to 0.7 to maintain relatively focused outputs, while a repetition penalty of 1.5 and a no-repeat bigram constraint were applied to reduce repetitive phrases. The model generated one recommendation with a maximum sequence length of 100 tokens for each user complaint.

Table 5 Generation Configuration During Inference

Parameter	Value	Description
Maximum generated sequence length	100 tokens	Maximum total length of the generated output sequence
Number of returned sequences	1	Produces one recommendation for each input complaint
Sampling method	Enabled	Generates text probabilistically rather than through deterministic greedy decoding
Top-(p) sampling	0.92	Selects tokens from the smallest cumulative probability distribution reaching 0.92
Top-(k) sampling	50	Limits token selection to the 50 most probable candidates
Temperature	0.7	Controls output randomness; values below 1 produce more focused responses
Repetition penalty	1.5	Penalizes tokens that have already appeared in the generated sequence
No-repeat (n)-gram size	2	Prevents repeated bigrams from appearing in the output
Number of generated recommendations	1	Returns a single generated recommendation for each prompt
Early stopping	Enabled	Stops generation when the model reaches an ending condition
Special-token removal	Enabled	Removes model-specific special tokens during output decoding
Computing device	GPU when available; otherwise CPU	Automatically places the model on an available computing device
Parameter	Value	Description
Maximum generated sequence length	100 tokens	Maximum total length of the generated output sequence
Number of returned sequences	1	Produces one recommendation for each input complaint

- e. Evaluation: Model evaluation is performed by monitoring training loss, validation loss, and perplexity throughout the fine-tuning process. Additionally, BERTScore and Cosine Similarity are used to assess the semantic quality and relevance of the generated text in relation to ground-truth responses.
- f. Deployment: The final fine-tuned model is deployed by integrating it into a prototype system. This prototype allows users or stakeholders to input a user complaint or issue and receive AI-generated recommendation text directly from the model.

3. Results and Discussion

3.1. Sentiment Analysis Result

The sentiment analysis process classified a total of 33,264 user reviews from the Digital Korlantas POLRI application into three categories: positive, neutral, and negative. The distribution of sentiment is as follows in Figure 3

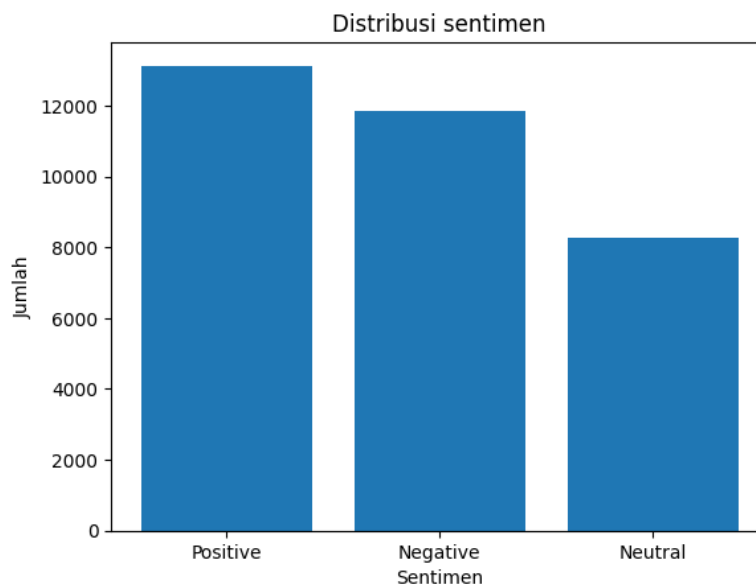


Figure 3 Sentiment Distribution

The IndoBERT model successfully classified the dataset into 13,124 positive reviews, 11,856 negative reviews, and 8,283 neutral reviews. The results indicate that the number of negative reviews is slightly lower than the number of positive ones. However, this still reflects user dissatisfaction in several key areas. This proportion of negative feedback is consistent with findings from previous studies on e-government applications, where a gap between service delivery and user expectations is frequently reported. However, this research uses a review that labels negative reviews to analyze causes. After data preparation, the dataset contains 11,382 negative reviews.

3.2. Root Cause Analysis

To uncover the primary sources of user dissatisfaction, negative reviews were analyzed using Latent Dirichlet Allocation (LDA) and Guided LDA. Both models were compared based on their coherence score, which measures topic interpretability. The results are summarized in Table 1.

Table 6 Evaluation of each model

Experiment	LDA Coherence	Guided LDA Coherence
2	0.39066	0.44024
3	0.39236	0.50635
4	0.43436	0.49825
5	0.39860	0.50007
6	0.41174	0.47982
7	0.44173	0.51667

The Guided LDA model achieved a coherence score of 0.51, outperforming the standard LDA model, which reached a maximum coherence score of 0.44 with 7 topics. According to Rahimi et al. (2024), coherence values above 0.5 generally indicate interpretable topics, suggesting that the Guided LDA model successfully generated semantically coherent topic clusters. This score is also higher than those reported in previous studies applying topic modeling to user feedback analysis, which typically report coherence values ranging from 0.42 to 0.48 (Hyun et al., 2024; Lee & Oh, 2024). The improvement can be attributed to the incorporation of EUCS-derived seed words, which guided the model toward more semantically meaningful and domain-relevant topic groupings compared to the purely statistical approach of unsupervised LDA.

A further comparison was conducted to analyze the extent to which the keywords from each model contributed to the dataset. This was done by examining the presence of keywords generated by both the LDA and Guided LDA models within each user review. The results are presented in Table 2.

Table 7 Contribution of Each Model Based on Keyword Presence in Reviews

Contribution Category	Total Reviews
LDA	864
Guided LDA	423
Both model	9,251
Not both Model	843

From the table above, both models perform well in identifying hidden topics within the dataset. The Guided LDA model demonstrated superior keyword specificity. Analysis of keyword presence within reviews revealed that 9,251 reviews (81.3% of negative reviews) contained keywords from both models, while 864 reviews contained only LDA keywords and 423 reviews contained only Guided LDA keywords. The lower number of reviews matched exclusively by Guided LDA (423 compared to 864 for LDA) indicates that Guided LDA's keywords are more specific and tightly bound to particular aspects, likely due to the influence of the predefined seed words. This suggests that Guided LDA offers higher topic precision, producing more focused and interpretable topic clusters even though it may miss some broader discussions captured by LDA.

To contextualize the findings of this study, the performance and analytical depth of the proposed approach were benchmarked against prior research in related domains. In terms of topic modeling quality, the Guided LDA model in this study achieved a coherence score of 0.51 within the public service context, outperforming standard LDA implementations in earlier works—such as Abrori (2024) in restaurant review analysis (0.42), Hyun et al. (2024) in architecture spatial themes (0.45), and Lee & Oh (2024) in space syntax pattern analysis (0.48). This performance gain stems directly from the semi-supervised mechanism, which leverages domain-specific seed words derived from the EUCS framework to guide topic formation.

Furthermore, while single-dimensional methods have been widely applied, such as Naïve Bayes for sentiment classification on social media (Darwis et al., 2021) or standalone topic modelling this study successfully integrates fine-tuned IndoBERT with Guided LDA to capture both emotional polarity and underlying thematic issues simultaneously, identifying seven distinct EUCS-aligned operational problems. Finally, while Park (2021) introduced a conceptual sentiment-based Root Cause Analysis (RCA) framework in a general context, this study advances that work by operationalizing a sentiment-informed RCA integrated directly with topic modeling, providing empirical validation of its practical effectiveness in resolving public service delivery challenges.

Based on these findings, the Root Cause Analysis (RCA) is conducted using the topics generated by the Guided LDA model, as it demonstrated better coherence and more specific keyword contribution. The names of the aspects are derived from the EUCS (End-User Computing Satisfaction) dimensions, based on the results of the LDA model, which classified each topic into a corresponding dimension. These aspect names represent a combination of the topic modeling results and the EUCS framework, and were subsequently used as seed words for the Guided LDA model.

The sub-aspects represent the specific keywords or themes extracted from each topic generated by the Guided LDA model. The results of the root cause mapping are illustrated in Figure 4.

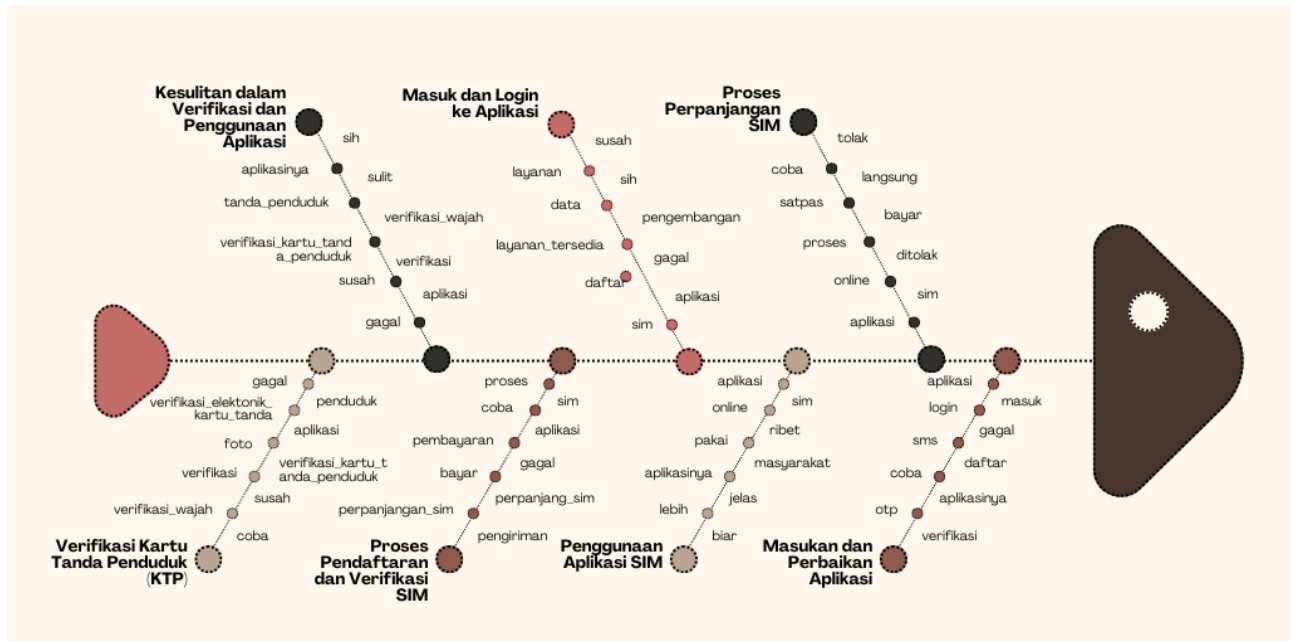


Figure 4 Fishbone Diagram using Guided LDA model Result

The findings from the Guided LDA-based root cause analysis provide important insights into user behavior and expectations toward digital public services. The dominance of issues related to system reliability, service responsiveness, and ease of use indicates that technical stability remains a critical concern among users. This finding aligns with previous studies on e-government systems, which emphasize that system errors and slow response times significantly influence negative user perceptions.

The seven topics identified by Guided LDA were mapped to EUCS dimensions based on the dominant keywords within each topic cluster. Each topic was then interpreted to identify specific root causes of user dissatisfaction, guided by the fishbone diagram analysis. The mapping between topics, EUCS dimensions, keywords, and identified root causes is presented in Table 8. This structured

interpretation enables a systematic understanding of user complaints and provides a foundation for generating targeted service improvement recommendations.

Table 8 Topic Interpretation and Root Cause Mapping Based on Guided LDA Results

Aspect	Keywords	Root Cause
App Feedback and Fixes (Content)	min, pembaruan, sim, perbaikan, masukan_nomor, gagal, login, otp	• Lack of responsiveness to user feedback • Application update process does not resolve existing issues • Repeated login or OTP input failures
Usage of SIM Application (Ease of Use)	ribet, layanan_tersedia, pengembangan, tidak jelas, masyarakat	• Non-intuitive application navigation • Service information is not clearly explained • The application is difficult for the general public to use
SIM Registration and Verification Process (Accuracy)	tes_kesehatan, unggah, gagal, pembayaran, proses, sim	• Frequent failure in document upload process • Lack of step-by-step clarity, particularly during the health test • Disruptions during payment and verification
National ID (KTP) Verification (Accuracy)	verifikasi_elektronik, susah_banget, nomor_telepon, foto, gagal	• e-KTP verification system frequently fails to read documents • Inaccurate facial recognition feature • Requested photo is difficult to align with system standards
SIM Renewal Process (Timeliness)	perpanjang_sim, ditolak, prosesnya, online, pembayaran	• Many processes are not immediately approved • Data mismatch leads to automatic rejection • Processing time is inconsistent with user expectations
Application Login (Ease of Use)	login, daftar, nomor_handphone, email, layanan_tersedia	• Difficulties during initial registration process • SMS OTP not received by users • Issues during phone number or email verification
Verification and Application Usage Difficulties (Accuracy)	verifikasi_wajah, tanda_penduduk, sulit, aplikasi, masuk	• Unresponsive facial verification • Frequent app errors during the identity verification process • Absence of technical guidance when errors occur

The topic interpretation reveals several critical patterns in user dissatisfaction. First, Accuracy-related issues dominate the topics, with three separate clusters (SIM registration verification, e-KTP verification, and facial verification) all pointing to systemic problems in the application's verification processes. Users consistently reported failures in document reading, facial recognition, and data matching, suggesting fundamental issues in the application's integration with government databases and biometric systems.

Second, Ease of Use emerges as a significant concern, with users describing the application as "ribet" (complicated) and expressing confusion about step-by-step procedures, particularly during health tests and document uploads. This indicates that the application's user interface and user

experience design may not adequately accommodate users with varying levels of digital literacy, which is a critical consideration for public services that serve diverse populations.

Third, Content-related issues, including lack of responsiveness to feedback and non-intuitive navigation, suggest that the application's information architecture and feature organization require improvement. Users expressed frustration with repeated login failures and OTP issues, indicating potential technical debt in the authentication system.

Interestingly, Timeliness issues, while present, were less frequently mentioned compared to accuracy and ease of use. This suggests that while users expect efficient processing, the primary drivers of dissatisfaction are related to the application's ability to correctly process user data and provide clear guidance throughout the user journey.

These findings align with previous studies on e-government applications, which have consistently identified system accuracy and usability as critical success factors ([Halim, 2020](#); [Park, 2021](#)). The dominance of accuracy-related complaints suggests that technical integration with backend systems, particularly with the e-KTP database and biometric verification, represents the most urgent area for service improvement. Furthermore, mapping these root causes to specific EUCS dimensions provides a structured foundation for generating targeted recommendations.

These insights justify using Root Cause Analysis (RCA) as a complementary qualitative approach. By visualizing the relationships between contributing factors through a fishbone diagram, stakeholders can more easily identify priority areas for intervention. Ultimately, this structured diagnosis serves as a critical input for the recommendation generation phase, ensuring that the outputs produced by the language model are grounded in empirical findings rather than generic assumptions.

3.3. Fine tuned model

The purpose of the LLM-OPS component is to generate service improvement suggestions in response to user issues identified during the sentiment and topic modeling phases. The model was fine-tuned on comment–response pairs from the Digital Korlantas POLRI dataset. Fine-tuning was conducted using multiple dataset split configurations: 50:50, 70:30, 80:20, and 90:10 (train: test ratio). The evaluation results of each configuration are shown in Table 3.

Table 9 LLM Evaluation Metric by Dataset Split

Metric	50:50	70:30	80:20	90:10
Training Loss	0.101	0.117	0.1	0.103
Validation Loss	0.490	0.489	0.408	0.354
Perplexity	1.63	1.63	1.50	1.43
BERT Score (F1)	0.786	0.776	0.779	0.781
Cosine Similarity	0.824	0.869	0.752	0.847

From the table, it can be seen that the 90:10 split produced the best overall results, achieving the lowest validation loss (0.354) and perplexity (1.43), as well as high semantic similarity as measured by BERTScore (0.781) and Cosine Similarity (0.847). These scores indicate that the model was able to generate coherent, relevant, and semantically accurate recommendations.

In addition to the quantitative evaluation, a human assessment was conducted to validate the quality and usefulness of the model's output. This evaluation compared the generated outputs from

each dataset split configuration to assess which version produced the most relevant and actionable recommendations from a human perspective. The evaluation was conducted on 10 respondents by comparing four outputs from different split datasets.

The assessment focused on three criteria:

- a. Relevance: how well the recommendation addressed the user’s problem.
- b. Clarity: how clear and understandable the recommendation was.
- c. Usefulness: how practical or actionable the recommendation was.

Each criterion was rated on a 1–5 Likert scale. The average scores are summarized in Table 4.

Table 10 Average Human Evaluation by Dataset Split

Criteria	50:50	70:30	80:20	90:10
Relevance	3.7	3.9	4.4	4.4
Clarity	4.3	4	4.3	4
Usefulness	4.2	3.8	4.4	4.4

The 80:20 dataset split configuration was selected as the final model based on a trade-off between automatic metrics and human evaluation. Although the 90:10 split achieved slightly superior quantitative scores (validation loss: 0.354, perplexity: 1.43, BERTScore: 0.781), human evaluation revealed a tendency toward overfitting, where responses replicated exact training phrases rather than generating adaptable recommendations. Conversely, while the 80:20 split yielded slightly lower automatic metrics (validation loss: 0.408, perplexity: 1.50, BERTScore: 0.779), it scored consistently high in human evaluation across relevance (4.4), clarity (4.3), and usefulness (4.4) by offering a superior balance between specificity and generalization. Lower training splits (70:30 and 50:50) failed to capture sufficient response patterns, resulting in weaker utility. This selection underscores that automatic metrics do not always align with human judgment in LLM quality assessment ([Liu et al., 2023](#)), prioritizing practical generalization for real-world deployment.

The results from the LDA and Guided LDA models are integrated into the prototype. The prototype displays the identified root causes and allows users to interact with the fine-tuned GPT-2 model to generate improvement recommendations. The interface and functionality of the prototype are illustrated in Figure 5 and Figure 6.

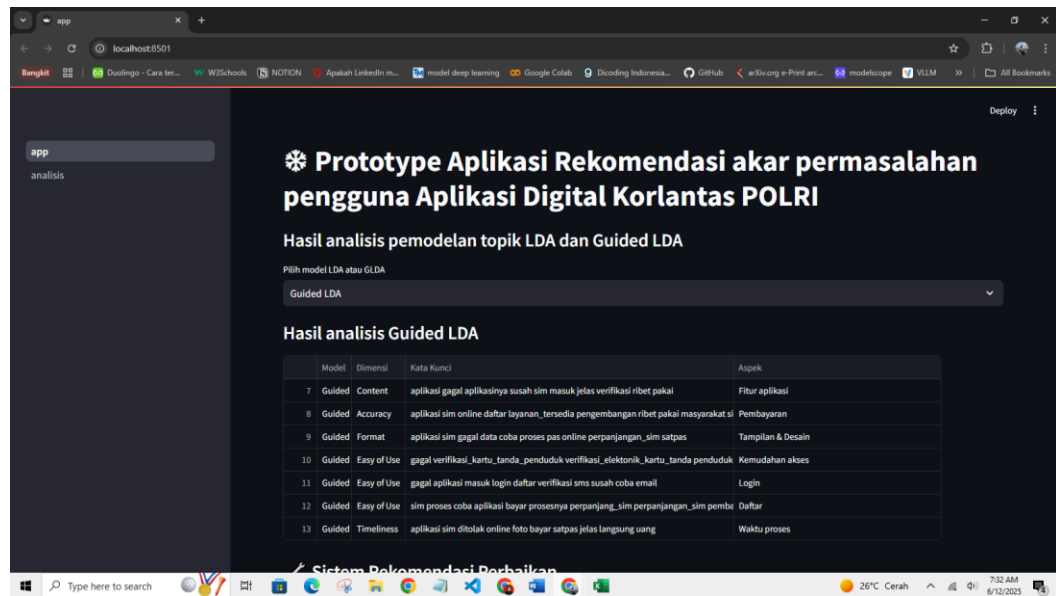


Figure 5 Dashboard displaying the results of LDA and Guided LDA topic modeling.

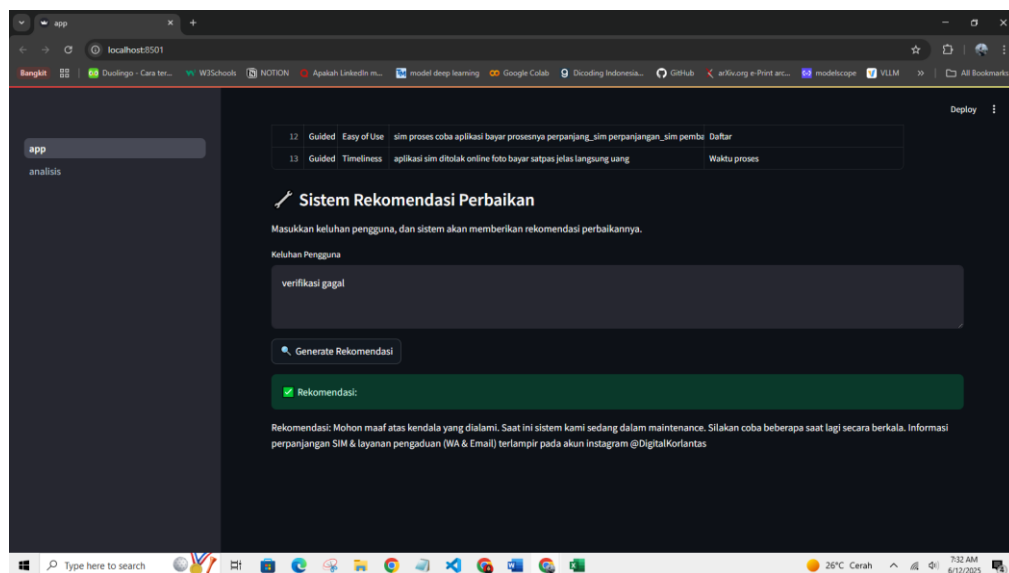


Figure 6 Recommendation section of the prototype powered by the fine-tuned GPT-2 model

4. Discussion

The sentiment analysis results revealed that 35.6% of user reviews expressed negative sentiment. This substantial proportion indicates meaningful dissatisfaction with the Digital Korlantas POLRI application despite its relatively high published rating. This finding confirms the initial observation that prompted this study: rating scores alone are insufficient indicators of service quality, as users may leave high ratings while simultaneously expressing significant frustrations in their review text. The discrepancy between aggregate ratings and textual feedback suggests a positivity bias in user ratings, where users may rate based on the application's overall utility while still noting specific pain points, or alternatively, rate based on a single positive experience while ignoring persistent technical issues (Darwis et al., 2021; Halim, 2020). This pattern is consistent with findings from previous studies on e-government applications, where user ratings often exhibit a positivity bias while textual feedback reveals more nuanced dissatisfaction. The practical implication of this finding is that public service organizations should not rely solely on aggregate rating scores as performance metrics;

instead, a more comprehensive evaluation approach that incorporates sentiment analysis of textual feedback is essential for identifying service quality issues that may be masked by high average ratings. For the Digital Korlantas POLRI application, the sentiment distribution suggests that while the application serves its core function, significant improvements are needed in specific areas to reduce the proportion of dissatisfied users, particularly given that the application serves as a critical public service interface for millions of Indonesian citizens ([Utami et al., 2023](#)).

The Guided LDA model successfully identified seven interpretable topics aligned with the EUCS framework, achieving a coherence score of 0.51 that exceeds the typical range of 0.42-0.48 reported in previous LDA-based studies on user feedback ([Hyun et al., 2024](#); [Lee & Oh, 2024](#); [Abrori, 2024](#)). According to [Rahimi et al. \(2024\)](#), coherence values above 0.5 generally indicate interpretable topics, suggesting that the Guided LDA model successfully generated semantically coherent topic clusters. The improvement over standard LDA (0.44) can be attributed to the semi-supervised nature of Guided LDA, which incorporates domain-specific seed words derived from the EUCS framework to guide the topic discovery process toward semantically meaningful clusters ([Zhou et al., 2023](#)). This approach addresses a key limitation of unsupervised LDA, which relies solely on statistical co-occurrence and often produces topics that are fragmented or misaligned with domain-specific concerns, particularly when applied to noisy, informal user-generated text ([Wu et al., 2024](#)). The dominance of Accuracy-related topics (three out of seven topics) and Ease of Use (two out of seven topics) in the results suggests that technical verification processes and usability represent the most critical pain points for users. This finding is consistent with previous research on e-government services, which has consistently identified system reliability and user experience as primary determinants of user satisfaction ([Halim, 2020](#); [Park, 2021](#)). However, this study provides a more granular understanding by identifying specific sub-issues within each dimension, such as e-KTP verification failures, facial recognition inaccuracies, and unclear step-by-step guidance during health tests, which is essential for developing targeted interventions. The integration of topic modeling with Root Cause Analysis using fishbone diagrams represents an advancement over previous work. At the same time [Park \(2021\)](#) explored the relationship between sentiment analysis and RCA conceptually; this study operationalizes sentiment-informed RCA integrated directly with topic modeling, providing empirical validation of its practical effectiveness in resolving public service delivery challenges.

The fine-tuned GPT-2 model demonstrated strong performance in generating service improvement recommendations, with the 80:20 configuration achieving the optimal balance between automatic metrics and human-perceived quality. The model achieved a perplexity of 1.50 and BERTScore of 0.779, with human evaluation scores of 4.4 for relevance, 4.3 for clarity, and 4.4 for usefulness on a 5-point scale. These results indicate that the model successfully learned the semantic patterns of official responses and can generate coherent, relevant recommendations that align with user complaints ([Zhang & Shafiq, 2024](#); [Sayeed et al., 2023](#)). The performance is particularly noteworthy given the relatively small training dataset (approximately 960 comment-response pairs), suggesting that GPT-2 can effectively learn from limited data when fine-tuned on domain-specific text, consistent with findings from previous studies on LLM fine-tuning with limited resources ([Naveed et al., 2025](#); [Liu et al., 2023](#)). The selection of the 80:20 configuration over the 90:10 configuration, despite the latter achieving better automatic metrics (perplexity: 1.43, validation loss: 0.354), highlights an important consideration in LLM evaluation: automatic metrics do not always correlate with human-perceived quality ([Liu et al., 2023](#)). The 90:10 model, while producing numerically superior results, generated overly specific responses that replicated training patterns rather than producing generalizable recommendations, indicating potential overfitting. This phenomenon has been observed in previous studies on text generation, where models trained on limited data tend to memorize rather than generalize ([Naveed et al., 2025](#)). The 80:20 configuration, in contrast,

demonstrated better generalization capability, generating responses that were balanced between specificity and adaptability, which is essential for practical deployment in a recommendation system. The integration of GPT-2 with the analytical findings (sentiment analysis and topic modeling) represents a novel contribution to the literature on digital public service evaluation. While previous studies have applied LLMs for recommendation generation in various domains ([Zhao et al., 2024](#)), this study demonstrates the value of grounding LLM outputs in empirical analytical findings rather than relying solely on the model's pre-trained knowledge. By incorporating root cause analysis into the recommendation generation pipeline, the model can produce more targeted and contextually appropriate recommendations, offering a scalable methodology for public service organizations to transform user feedback into actionable service improvements.

The theoretical implications of this study extend to several areas of digital service evaluation research. First, the study demonstrates the value of integrating multiple analytical approaches—sentiment analysis, topic modeling, Root Cause Analysis, and LLM-based recommendation generation—within a unified framework. While previous studies have applied these techniques individually ([Darwis et al., 2021](#); [Hyun et al., 2024](#); [Graber et al., 2024](#)), this study provides empirical evidence that their integration yields more comprehensive and actionable insights than any single approach alone. The framework bridges the gap between descriptive analysis (what users are complaining about) and prescriptive analysis (what should be done about it), addressing a key limitation in current service evaluation research ([Sayeed et.al., 2023](#)). Second, the study contributes to the methodological advancement of topic modeling in service evaluation contexts through the use of EUCS-derived seed words to guide Guided LDA. This method addresses the common criticism that unsupervised LDA produces topics that are difficult to interpret or misaligned with domain-specific concerns, validating the effectiveness of semi-supervised approaches for domain-specific user feedback analysis ([Zhou et.al., 2023](#); [Wu et al., 2024](#)). Third, the study advances understanding of the relationship between user ratings and textual feedback, finding that high average ratings can coexist with substantial negative feedback, which highlights the limitations of single-metric service quality assessment and underscores the need for multi-dimensional evaluation approaches. This finding has implications for service quality theory, suggesting that user satisfaction should be conceptualized as a multi-faceted construct that is not fully captured by aggregate rating scores ([Halim, 2020](#)). The practical implications of this study extend to several stakeholders involved in digital public service delivery. For application developers and service managers, the structured root cause analysis provides clear priorities for improvement interventions, with the identification of Accuracy as the most critical issue suggesting that immediate attention should be directed toward enhancing the application's verification capabilities, particularly in terms of e-KTP integration, facial recognition accuracy, and data validation processes ([Utami et al., 2023](#)). The mapping of specific root causes to EUCS dimensions enables resource allocation based on impact and urgency, facilitating more efficient service improvement planning. For policymakers and government agencies, the study demonstrates the value of systematic user feedback analysis in digital transformation initiatives, offering a methodology that can be replicated across other public service applications to enable evidence-based decision-making in service improvement ([POLRI, 2021](#)). The integration of LLM-based recommendation generation offers the potential to automate aspects of user feedback response, reducing the human resources required for complaint handling while maintaining response quality and consistency. For the broader academic community, the study provides a replicable methodology for evaluating digital public services that can be adapted to different contexts and application domains, with the combination of sentiment analysis, topic modeling, RCA, and LLM-based recommendation generation offering a flexible framework for user experience research in various digital service contexts.

Limitations and Future Work

Despite the promising results, this study has several limitations that should be acknowledged. First, the sentiment classification model (IndoBERT) was used in a zero-shot inference setting without additional fine-tuning on the specific dataset, relying on its previously validated benchmark performance on Indonesian sentiment analysis tasks ([Sayeed et.al., 2023](#)). While manual validation confirmed acceptable performance, a fully fine-tuned model could potentially achieve higher classification accuracy, particularly for nuanced or ambiguous reviews. Future research should consider fine-tuning IndoBERT on a labeled dataset specific to the application domain to improve classification precision and reduce potential misclassifications. Second, the dataset was limited to reviews from a single application (Digital Korlantas POLRI) on a single platform (Google Play Store). While this focused approach enabled in-depth analysis, the generalizability of the findings to other applications or platforms requires further investigation. Future studies should apply the framework to multiple applications across different public service domains to validate the methodology and identify domain-specific patterns in user dissatisfaction. Third, the GPT-2 model's recommendation generation was trained on relatively few comment-response pairs (approximately 960 pairs). While the model performed well on automatic and human evaluation metrics, performance could potentially be improved with a larger training dataset. Future research should collect a larger corpus of official responses to enable more robust fine-tuning and potentially explore larger language models such as GPT-3 or GPT-4 for recommendation generation, provided that computational resources and data privacy considerations can be addressed ([Naveed et al., 2025](#); [Zhao et al., 2024](#); [Islam et al., 2023](#)). Fourth, the human evaluation of generated recommendations involved only 10 respondents. While the evaluation provided valuable insights, a larger and more diverse evaluation panel would strengthen the validity of the findings. Future research should conduct larger-scale human evaluations, potentially involving both users and service managers, to assess the practical utility of generated recommendations from multiple perspectives. Fifth, while the study identifies root causes of user dissatisfaction, it does not implement or evaluate the effectiveness of the recommended interventions. Future research should extend this work by implementing the proposed recommendations and measuring their impact on user satisfaction through longitudinal analysis, providing empirical evidence of the framework's effectiveness in driving service improvement. Sixth, the study's reliance on a single data source (Google Play reviews) may not capture the full spectrum of user experiences, as dissatisfied users may be more likely to leave reviews, potentially introducing selection bias. Future research should consider integrating multiple data sources, such as surveys, interviews, and user analytics, to provide a more comprehensive understanding of user satisfaction. Despite these limitations, the proposed framework offers a robust and scalable approach for analyzing user feedback and generating actionable recommendations in digital public service contexts, with significant potential for adaptation and extension in future research.

5. Conclusions

This study has demonstrated the potential of an integrated analytical framework for identifying and addressing user dissatisfaction in digital public service applications, combining sentiment analysis using IndoBERT, topic modeling with Guided LDA, Root Cause Analysis through fishbone diagrams, and recommendation generation via a fine-tuned GPT-2 model. The Guided LDA model achieved a coherence score of 0.51, indicating its capability to generate interpretable topics aligned with the EUCS framework, while the fine-tuned GPT-2 model showed promising performance in generating contextually relevant recommendations with a perplexity of 1.50 and human evaluation scores above 4.0 for relevance, clarity, and usefulness. The findings suggest that Accuracy-related issues and Ease of Use represent the most critical pain points for Digital Korlantas

POLRI users, with technical verification processes, particularly e-KTP integration, facial recognition, and data validation emerging as priority areas for intervention. The developed prototype illustrates how analytical insights can be translated into practical service improvement tools. However, the framework's generalizability requires further validation across multiple applications and platforms. Future research should focus on fine-tuning the sentiment classification model on domain-specific data, expanding the training dataset for recommendation generation, and conducting longitudinal studies to evaluate the impact of implemented recommendations on user satisfaction. Despite these limitations, the proposed framework shows promise as a scalable approach for enhancing user experience in digital government platforms, contributing to the growing body of literature on data-driven public service improvement.

Acknowledgments

The authors would like to express their sincere gratitude to Dr. I Kadek Dwi Nuryana, S.T., M.Kom. for their valuable guidance, continuous support, and insightful feedback throughout this research. In addition, the author wishes to thank Information System Study Program, Universitas Negeri Surabaya for providing the necessary resources and assistance during the completion of this study.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions

The authors were responsible for all aspects of the research presented in this paper, including conceptualization, methodology, analysis, writing, and review. The authors have read and agreed to the published version of the manuscript.

Ethics Statement

The data supporting the findings of this study are integrated into and accessible through the online application Digital Korlantas POLRI.

AI Use Declaration

The authors declare that no generative AI tools were used in the writing or analysis of this manuscript. The author takes full responsibility for the content of the publication.

References

- Abrori, N. Y. A. (2024). Optimasi layanan Warung Ayam Bakar Lientang dengan root cause analysis berdasarkan hasil analisis sentimen berbasis aspek. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*. <https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/13615>
- Alam, H. K., & Hanny Purnamasari. (2024). EFEKTIVITAS DIGITALISASI PELAYANAN PUBLIK MELALUI APLIKASI DIGITAL KORLANTAS POLRI. *Moderat : Jurnal Ilmiah Ilmu Pemerintahan*, 10(3), 511–522. <https://doi.org/10.25157/moderat.v10i3.3668>
- Darwis, D., Siskawati, N., & Abidin, Z. (2021). Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional. *Jurnal Tekno Kompak*, 15(1), 131. <https://doi.org/10.33365/jtk.v15i1.744>
- Graber, M. L., Cooper, S., & Kachalia, A. (2024). Root cause analysis of cases involving diagnosis. *Diagnosis*, 11(1), 1–8. <https://doi.org/10.1515/dx-2024-0102>
- Halim, N. (2020). Analisis End-User Computing Satisfaction (EUCS) pada aplikasi mobile Universitas Bina Darma. *SISTEMIK: Jurnal Sistem Informasi*, 9(1), 1–8. <https://doi.org/10.32520/stmsi.v9i1.625>
- Hyun, J. H., & Oh, Y. (2024). Latent Dirichlet Allocation (LDA) topic models for space syntax studies on spatial experience. *City, Territory and Architecture*, 11(1), 1–20. <https://doi.org/10.1186/s40410-023-00223-3>
- Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., & Pedrycz, W. (2024). A comprehensive survey on applications of transformers for deep learning tasks. *Expert Systems with Applications*, 241, 122666. <https://doi.org/10.1016/j.eswa.2023.122666>
- Lee, J. H., & Oh, Y. (2024). Latent Dirichlet Allocation (LDA) topic models for Space Syntax studies on spatial experience. *City Territory and Architecture*, 11(1), 1–20. <https://doi.org/10.1186/s40410-023-00223-3>
- Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2), 100017. <https://doi.org/10.1016/j.metrad.2023.100017>
- Utami, M., Rusdi, M., Arif, I. A., Atmansyah, L., & Indar, N. I. N. (2023). Efektivitas Pelayanan Digital di Kepolisian RI: Studi Kasus Kota Makassar. *Development Policy and Management Review (DPMR)*, 63–79. <https://doi.org/10.61731/dpmr.v3i1.29951>
- Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2025). A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5), 1–72. <https://doi.org/10.1145/3744746>
- Park, S. K. (2021). Root cause analysis based on relations among sentiment words. *Cognitive Computation*, 13, 903–918. <https://doi.org/10.1007/s12559-021-09872-3>
- POLRI, K. (2021). Digital Korlantas POLRI [Mobile app]. Google Play Store. <https://play.google.com/store/apps/details?id=id.qoin.korlantas.user&hl=id>
- Rahimi, H., Mimno, D., Hoover, J., Naacke, H., Constantin, C., & Amann, B. (2024). Contextualized topic coherence metrics. In *Findings of the Association for Computational Linguistics: EACL 2024* (pp. 1760–1773). <https://doi.org/10.18653/v1/2024.findings-eacl.123>
- Sayeed, M. S., Mohan, V., & Muthu, K. S. (2023). BERT: A review of applications in sentiment analysis. *HighTech and Innovation Journal*, 4(2), 453–462. <https://doi.org/10.28991/HIJ-2023-04-02-015>
- Schröer, C., Kruse, F., & Gomez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, 181, 526–534. <https://doi.org/10.1016/j.procs.2021.01.198>

- Wu, X., Nguyen, T., & Luu, A. T. (2024). A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2). <https://doi.org/10.1007/s10462-023-10661-7>
- Zhang, H., & Shafiq, M. O. (2024). Survey of transformers and towards ensemble learning using transformers for natural language processing. *Journal of Big Data*, 11, 25. <https://doi.org/10.1186/s40537-023-00842-0>
- Zhao, Z., Fan, W., Li, J., Liu, Y., Mei, X., Wang, Y., Wen, Z., Wang, F., Zhao, X., Tang, J., & Li, Q. (2024). Recommender systems in the era of large language models (LLMs). *IEEE Transactions on Knowledge and Data Engineering*, 36(11), 6889–6907. <https://doi.org/10.1109/TKDE.2024.3392335>
- Zhou, K., Wang, J., Ashuri, B., & Chen, J. (2023). Discovering the research topics on construction safety and health using semi-supervised topic modeling. *Buildings*, 13(5), 1169. <https://doi.org/10.3390/buildings13051169>