

---

## Comparative Topic Modelling of Mobile Banking User Reviews Using LDA and BERTopic: A Case Study of wondr by BNI

Dicky Halim<sup>1</sup>, Indra Budi<sup>2</sup>, Basma Fathan Mubina<sup>3</sup>, Aris Budi Santoso<sup>4</sup>, Prabu Kresna Putra<sup>5</sup>

dicky.halim@office.ui.ac.id<sup>1</sup>, indra@cs.ui.ac.id<sup>2</sup>, basma.fathan@cs.ui.ac.id<sup>3</sup>,

aris.budi@office.ui.ac.id<sup>4</sup>, prabu.kresna@office.ui.ac.id<sup>5</sup>

<sup>1,2,3,4,5</sup> Faculty of Computer Science, Universitas Indonesia, Jakarta, Indonesia

---

### Article Information

Received : 7 April 2026

Revised : 13 April 2026

Accepted : 23 April 2026

---

### Keywords

Mobile Banking, Topic Modelling, LDA, BERTopic

---

### Abstract

This study explores user reviews of the Wondr mobile banking application to identify factors that influence user experience and service quality. The dataset, obtained from the Google Play Store, was processed through several preprocessing steps, including normalization, stopword removal, and stemming. Two topic modelling methods were applied: Latent Dirichlet Allocation (LDA) as a probabilistic baseline and BERTopic as an embedding-based approach. The LDA model was evaluated using coherence scores to determine the most suitable number of topics, while BERTopic was assessed based on topic distribution, interpretability, and additional coherence analysis.

The results show that BERTopic produces more semantically meaningful and contextually rich topics, particularly in capturing short-text user reviews. Although BERTopic achieves lower overall coherence compared to LDA, certain topics demonstrate high semantic consistency, especially for well-defined issues such as login verification problems. The analysis reveals that most user feedback is concentrated on positive user experience, while critical issues related to login verification and system errors remain significant concerns. These findings provide actionable insights for improving mobile banking services and demonstrate the effectiveness of embedding-based topic modeling in financial text analytics. These findings highlight a trade-off between statistical consistency and semantic richness in topic modeling approaches. The results provide actionable insights for improving mobile banking services and demonstrate the effectiveness of combining probabilistic and embedding-based methods in financial text analytics.

## A. Introduction

Over the past decade, the financial sector has undergone rapid transformation, largely influenced by ongoing developments in digital technology. One clear example of this shift is the rise of mobile banking applications, which have simplified how people handle everyday financial activities. Tasks that once required a visit to a bank can now be completed in seconds through a smartphone. This rapid growth of digital banking has fundamentally reshaped service delivery and structural operations within the financial industry [1]. As a result, users have grown accustomed to this convenience and increasingly expect applications to function smoothly, securely, and without disruption. In this environment, user reviews on platforms such as Google Play have become an important source of practical feedback, reflecting how these applications perform in real-life situations rather than controlled settings [2]. The analysis of such user-generated content from app stores is increasingly vital for banks to assess their competitive positioning and service quality [3].

At the same time, these reviews introduce a different kind of challenge. They are written in free-form languages, differ greatly in style, and accumulate quickly in large numbers. Because of this, manually reviewing them is not only time-consuming but also inefficient. To deal with this issue, many researchers now turn to natural language processing (NLP). One commonly used approach is topic modelling, which helps identify patterns or recurring themes within large collections of text. This method has proven useful in areas such as customer feedback analysis and business intelligence, where understanding user perspectives can support better decision-making [4]. Among the available techniques, Latent Dirichlet Allocation (LDA) is frequently used because it offers a structured, probabilistic way to detect hidden topics in a set of documents [5].

Even so, LDA is not without drawbacks. Because it relies on a bag-of-words approach, it treats words independently and does not consider their order or deeper contextual meaning [6]. This methodological constraint often leads to lower coherence when applied to the fragmented and informal nature of digital banking reviews [7]. Such limitations become especially noticeable when dealing with short texts like mobile app reviews, where meaning often depends on context and subtle phrasing. In response to these challenges, newer approaches have been developed. One example is BERTopic, a method that builds transformer-based models to better capture how words relate to each other in context [8]. Instead of relying solely on word frequency, it uses document embeddings along with dimensionality reduction and clustering techniques to form topics. In practice, this often leads to topics that are easier to interpret and more consistent, especially when working with short or less structured text such as user reviews [9]. Recent implementations of BERTopic on Indonesian financial services reviews have shown its superior ability to capture semantic nuances compared to traditional frequency-based models [10]. Studies have also explored enhancements in topic interpretability by integrating contextual understanding and advanced modeling strategies, further improving the quality of generated topics [11].

Previous studies comparing these approaches point to a general pattern. LDA tends to produce stable and interpretable results, while BERTopic is often more effective at capturing meaning and generating clearer topic distinctions [12].

Additionally, transformer-based approaches have been found to produce more meaningful topic groupings compared to traditional probabilistic models [13]. Recent developments have also introduced automated topic modeling pipelines that integrate clustering, topic extraction, and trend analysis, enabling more scalable and adaptive analysis of large textual datasets [14]. However, much of this research has focused on areas like tourism, news, or biomedical data. There is still limited work examining mobile banking reviews, particularly in the Indonesian context, where user experience analysis remains critical for regional technological adoption [15].

To address this gap, this study examines user reviews of wondr mobile banking application. The objective is to identify key themes related to user experience, system performance, and service quality. This study also compares how LDA and BERTopic perform in extracting useful insights from user-generated content. To evaluate the models, coherence scores are used for quantitative assessment, while interpretability is considered through qualitative analysis. The findings of this study are expected to provide practical insights for improving mobile banking services, while also contributing to ongoing research on topic modelling in financial text analysis. More specifically, this study adds value by offering an empirical comparison between LDA and BERTopic, expanding the evaluation approach through the inclusion of coherence analysis for embedding-based models, and drawing attention to the balance between statistical consistency and semantic interpretability.

## **B. Research Method**

### **B.1. Research Framework**

This study utilizes a text mining approach to analyze user reviews of the Wondr mobile banking application. The research is organized into four main stages, namely data collection, text preprocessing, topic modelling, and model evaluation. Two topic modelling techniques—Latent Dirichlet Allocation (LDA) and BERTopic—are applied to enable a comparison between conventional probabilistic methods and more recent embedding-based approaches. The framework is designed to uncover underlying topics within user reviews and highlight key issues related to user experience, system performance, and service quality. The outputs from both models are then assessed to determine how effectively they capture meaningful insights from unstructured textual data.

### **B.2. Data Collection**

The data used in this study consists of user reviews obtained from the Google Play Store for the Wondr mobile banking application. The collection process was carried out using a Python-based scraping technique, with a focus on reviews from the most recent six months to ensure that the data reflects current user experiences. Each review includes textual content, a rating score, and a timestamp. After collecting, the dataset was cleaned by removing incomplete or missing entries to maintain consistency and data quality. The resulting dataset is then used for the subsequent preprocessing and modelling stages.

### **B.3. Text Preprocessing**

Before analysis, the collected reviews undergo a preprocessing stage to convert raw text into a structured format suitable for topic modelling. The process begins with text cleaning, where irrelevant elements such as URLs, numbers, and punctuation are removed, and all text is converted to lowercase to ensure consistency. This is followed by normalization, which aims to standardize informal or inconsistent language commonly found in user-generated content.

Next, stopword removal is performed using a combination of Indonesian stopwords from the NLTK library and a manually curated list to filter out words that do not carry significant meaning. The text is then tokenized, breaking sentences into individual words to facilitate further processing. Stemming is applied using the Sastrawi library to reduce words to their root forms, helping to minimize redundancy in the vocabulary. Additionally, bigram construction is implemented using the Gensim library to capture commonly occurring word pairs that represent meaningful expressions. The final output of this stage is a cleaned and structured corpus, along with tokenized text data, which serves as input for the topic modelling stage.

### **B.4. Topic Modelling using LDA**

Latent Dirichlet Allocation (LDA) is employed as a baseline technique for topic modelling and is implemented using the Gensim library. It is a probabilistic approach in which each document is viewed as a combination of multiple topics, and each topic is characterized by a distribution of words. In this study, the preprocessed text corpus is first converted into a dictionary and a bag-of-words format to enable efficient processing.

To identify the optimal number of topics, several LDA models are trained with different topic configurations. The coherence score is used as the main evaluation metric, as it reflects the semantic consistency of the top words within each topic. The model that achieves the highest coherence score is selected as the final model. The resulting topics are then interpreted based on their keyword distributions to understand the main themes present in the user reviews.

### **B.5. Topic Modelling using BERTopic**

BERTopic is applied as an alternative approach that focuses on capturing contextual meaning within text. In contrast to LDA, this approach relies on transformer-based models to produce dense vector representations of documents, enabling a more accurate capture of semantic relationships between words. In this study, a pre-trained SentenceTransformer model is used to convert user reviews into embeddings.

The resulting embeddings are then processed using Uniform Manifold Approximation and Projection (UMAP) to reduce their dimensionality, while still maintaining the underlying semantic structure and improving computational performance. After that, clustering is carried out using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), which groups similar documents without the need to specify the number of topics in advance. Each resulting cluster is subsequently represented using class-based TF-IDF to identify key terms that best describe each topic.

To enable comparison with LDA, a topic reduction step is applied to merge similar topics into a more concise set. In addition, outlier documents identified during clustering are excluded to improve the clarity and interpretability of the final topics.

## B.6. Model Evaluation

The performance of the topic modelling approaches is evaluated using both quantitative and qualitative methods to ensure a balanced comparison. For the LDA model, the coherence score is used to determine the optimal number of topics and to assess the semantic consistency of the generated topics. For BERTopic, evaluation primarily focuses on topic distribution and interpretability, as embedding-based approaches rely on semantic similarity rather than word co-occurrence. Topic distribution analysis is used to examine how documents are distributed across topics, providing insight into the prominence of each theme.

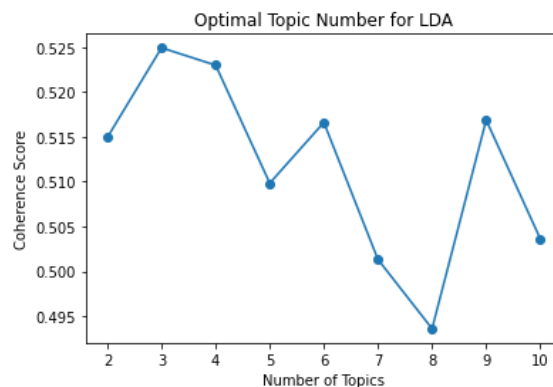
In addition, coherence scores are also computed for BERTopic as a complementary metric to provide a quantitative perspective. However, these scores are interpreted with caution, as coherence measures based on word co-occurrence may not fully capture the semantic structure of embedding-based topic models.

Furthermore, qualitative evaluation is conducted by reviewing the interpretability and distinctiveness of the topics produced by both models. This involves examining the top keywords associated with each topic as well as representative user reviews to confirm the relevance and meaning of each topic. By combining these evaluation methods, the study ensures a more comprehensive and fair comparison between LDA and BERTopic, considering both statistical performance and practical usefulness.

## C. Result and Discussion

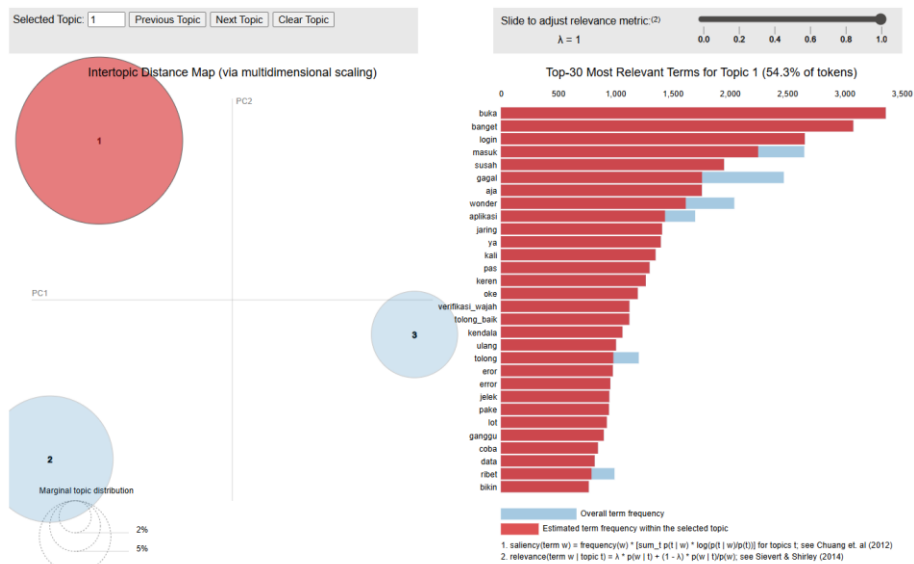
### C.1. LDA Topic Modelling Results

The LDA model was implemented to identify latent topics within user reviews of wondr mobile banking application. The appropriate number of topics was identified by testing several models and comparing their coherence scores, as illustrated in Figure 1. The experimental results indicate that the model reached its highest coherence score when three topics were used, suggesting that this configuration produces the most semantically consistent set of topics.



**Figure 1.** Coherence Score Plot (LDA)

The analysis highlights three main topics from user feedback, as illustrated in Figure 2. The first topic relates to issues with login, system performance, and face recognition. Many users report difficulties when accessing the application, particularly during authentication, which can lead to frustration and a negative initial experience. The second topic concerns problems encountered during financial transactions. Some users mention that the application is fast and convenient to use. However, others report issues such as failed transfers and difficulties when checking account balances. This suggests that, despite its strengths, the reliability of transaction-related features still requires attention. The third topic reflects overall user satisfaction. Several reviews contain positive expressions such as good, excellent, and helpful indicating that many users are satisfied with the application. Taken together, the LDA model can capture both positive and negative aspects of user feedback, although the identified topics tend to be distributed relatively evenly.



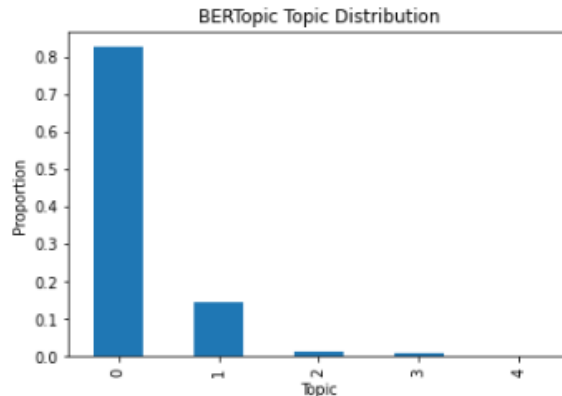
**Figure 2.** Visualization of LDA Topic Modelling

## C.2. BERTopic Modelling Results

The BERTopic model was applied to extract semantically meaningful topics from user reviews using an embedding-based approach. Unlike LDA, BERTopic does not require a predefined number of topics, as it determines topic structures dynamically based on semantic similarity within the data. To enable comparison with LDA, a topic reduction step was applied, resulting in five final topics after merging similar clusters.

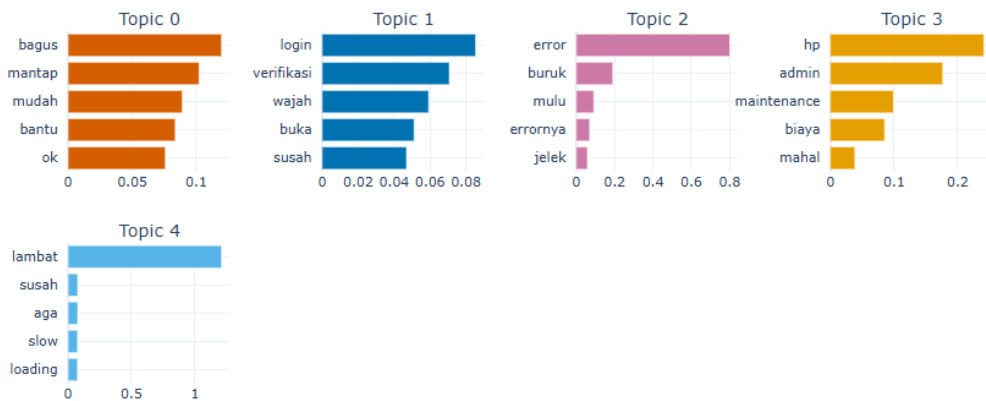
The topic distribution is illustrated in Figure 3, showing a highly skewed pattern. The majority of user reviews (82.6%) are concentrated in Topic 0, indicating that most feedback revolves around a dominant theme related to positive user experience, including ease of use, speed, and overall satisfaction. Topic 1 accounts for 14.7% of the data and primarily captures issues related to login and verification processes. The remaining topics (Topics 2, 3, and 4) represent only a small proportion of the data, each contributing less than 2% of the total reviews.

These topics reflect more specific concerns such as system errors, application performance, and operational issues.



**Figure 3.** Topic distribution generated by BERTopic model

Further interpretation of each topic is supported by the keyword representation shown in Figure 4. Topic 0 is characterized by words such as *bagus*, *mantap*, *mudah*, and *cepat*, indicating positive user sentiment toward the application. Topic 1 includes keywords such as *login*, *verifikasi*, and *wajah*, highlighting authentication-related issues. Topic 2 reflects negative experiences related to system errors, with keywords such as *error*, *buruk*, and *jelek*. Topic 3 captures operational concerns such as device compatibility and administrative processes, while Topic 4 is associated with performance issues, particularly related to slow loading and application responsiveness.



**Figure 4.** Top keywords for each topic identified by BERTopic

To complement the qualitative interpretation, coherence scores were also calculated for the BERTopic model as shown in Table 1. The overall coherence score is 0.475, which is lower than that of the LDA model. However, the per-topic coherence analysis reveals significant variation across topics. Topic 1 achieves the highest coherence score (0.698), indicating strong semantic consistency in capturing login and verification issues. Topics 2 and 4 also show moderate coherence scores (0.503 and 0.547, respectively), while Topics 0 and 3 exhibit lower coherence values. This variation suggests that BERTopic is effective in identifying highly coherent topics for dominant or well-defined themes, but may produce less consistent topic structures for smaller or less distinct clusters.

**Table 1.** Coherence score BERTopic

Topics -n	Coherence score
0	0,355
1	0,698
2	0,503
3	0,274
4	0,547
<b>Overall</b>	<b>0,475</b>

Overall, the findings suggest that BERTopic is particularly strong in reflecting the natural patterns of user feedback and in identifying key issues with clearer semantic meaning. However, the results also show a trade-off: while the model captures richer contextual information, it tends to produce uneven topic distributions and varying coherence levels, highlighting the balance between semantic depth and statistical consistency in embedding-based topic modelling approaches.

### C.3. Comparison between LDA and BERTopic

The comparison between LDA and BERTopic reveals notable differences in topic structure, interpretability, and data representation. LDA produces a fixed number of topics based on predefined parameters, resulting in a more evenly distributed topic structure. This allows LDA to provide a comprehensive segmentation of user feedback, including both major and minor themes. In addition, LDA achieves a higher overall coherence score, indicating more consistent statistical relationships among topic words.

In contrast, BERTopic adopts a data-driven approach, allowing the model to dynamically determine topic structures based on semantic similarity. As a result, BERTopic produces a highly skewed topic distribution, where one dominant topic represents most of user feedback. This indicates that BERTopic is more effective in capturing the natural distribution of user opinions rather than forcing equal topic representation.

From a quantitative perspective, the overall coherence score of BERTopic is lower than that of LDA. However, further analysis at the individual topic level shows that certain BERTopic topics achieve high coherence scores, particularly those representing well-defined and dominant themes such as login and verification issues. This suggests that while BERTopic may not optimize global coherence, it is capable of generating highly coherent topics for specific and meaningful clusters.

In terms of interpretability, BERTopic performs better in capturing the contextual relationships between words. The topics generated by BERTopic are more semantically coherent and easier to interpret, particularly for short-text data such as user reviews. Additionally, BERTopic enables the extraction of representative documents, providing deeper insights into user concerns.

However, LDA remains useful in providing a balanced and structured overview of topics, making it easier to identify a wider range of issues across the dataset. Therefore, both models offer complementary advantages: LDA provides broader topic coverage and higher statistical consistency, while BERTopic delivers more contextually meaningful and semantically rich insights.

#### **C.4. Discussion and Implications**

The findings of this study suggest that user feedback on the wondr mobile banking application is generally positive, as reflected in the dominant topic identified by BERTopic. However, recurring issues related to login and authentication indicate that important usability challenges still need to be addressed. Since authentication acts as the primary gateway to the application, enhancing this feature should be prioritized to improve the overall user experience.

In addition, although system-related errors appear less frequently in the data, the strong negative reactions associated with these issues underline the importance of maintaining system stability and reliability. Even a relatively small number of critical errors can significantly affect user satisfaction and trust.

From a methodological standpoint, the results show that embedding-based approaches such as BERTopic are more effective in capturing contextual meaning and semantic relationships, particularly when working with short-text data. Although BERTopic tends to produce lower overall coherence scores compared to LDA, a closer look at individual topics reveals that some achieve high semantic consistency, especially those related to clearly defined and frequently discussed issues like login and verification. This suggests that BERTopic is still capable of identifying meaningful and actionable insights, even when overall statistical coherence varies.

At the same time, traditional models such as LDA remain useful for providing a more structured and evenly distributed representation of topics, often with higher overall coherence. Therefore, combining both approaches offers a more balanced understanding of user feedback by integrating statistical reliability with richer semantic interpretation, ultimately supporting more effective data-driven decision-making in digital banking services.

#### **D. Conclusion**

This study examines user reviews of the wondr mobile banking application by applying two topic modelling techniques, namely Latent Dirichlet Allocation (LDA) and BERTopic, to identify key themes related to user experience, system performance, and service quality. The results demonstrate that both models can extract meaningful topics from user-generated reviews, but they differ significantly in their representation and interpretability of the data.

The LDA model provides a structured and balanced distribution of topics, enabling the identification of a wide range of issues, including login problems, transaction reliability, and general user satisfaction. In contrast, BERTopic captures the natural distribution of user feedback more effectively, revealing that most reviews are concentrated on a dominant topic related to positive user experience. At the same time, BERTopic highlights critical issues such as login and verification problems, as well as system errors, which, although less frequent, have high impact on user perception.

From a methodological perspective, the findings indicate that embedding-based topic modeling approaches offer superior capability in capturing semantic relationships and contextual meaning, particularly for short and noisy text data such as mobile application reviews. Although BERTopic exhibits lower overall coherence compared to LDA, further analysis shows that certain topics achieve high coherence

scores, indicating strong semantic consistency for well-defined themes. Meanwhile, traditional probabilistic models remain useful for providing a more comprehensive and evenly distributed overview of topics. Therefore, the combination of both approaches provides a more holistic understanding of user feedback by balancing statistical consistency and semantic interpretability.

From a practical standpoint, the results suggest that improving authentication processes and enhancing system stability should be prioritized to increase user satisfaction. While the overall perception of the application is positive, addressing critical pain points can significantly improve user experience and strengthen user trust in digital banking services.

This study is limited to user reviews collected within a specific time and focuses on a single mobile banking application. Future studies could build on this work by incorporating approaches such as sentiment analysis, examining changes in user feedback over time, or utilizing additional data sources, including application ratings. Furthermore, advanced topic modeling techniques and large language models can be explored to enhance topic interpretability and analytical depth.

### **E. Acknowledgment**

Throughout the completion of this research, the author received valuable guidance and support from academic supervisors as well as other related parties. Their assistance and encouragement played an important role in ensuring that the study could be carried out successfully. The authors would like to extend sincere appreciation to Universitas Indonesia for its ongoing support throughout the research process.

### **F. References**

- [1] V. Anggraini, I. Budi, A. B. Santoso, and P. K. Putra, "Measuring mobile banking service quality using Topic Modeling and Term Ranking: A case study of an Indonesian digital bank," *Indonesian Journal of Computer Science*, vol. 13, no. 6, 2024, doi: 10.33022/ijcs.v13i6.4517.
- [2] L. Çallı, "Exploring mobile banking adoption and service quality features through user-generated content: The application of a topic modeling approach," *International Journal of Bank Marketing*, vol. 41, no. 2, pp. 428–447, 2023, doi: 10.1108/IJBM-10-2021-0489.
- [3] Y. S. Amirkhalili and H. Y. Wong, "Banking on Feedback: Text Analysis of Mobile Banking iOS and Google App Reviews," *arXiv preprint*, 2025, doi: 10.48550/arxiv.2503.11861.
- [4] A. F. A. Mahmoud et al., "A Comparative Evaluation of LDA, NMF, and BERTopic: Analyzing Perplexity and Coherence Metrics," *Ingénierie des Systèmes d'Information*, vol. 30, no. 12, pp. 3163–3169, 2025, doi: 10.18280/isi.301208.
- [5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
- [6] T. Mihajlov and M. Ikonc Nešić, "Topic Modeling of the SrpELTeC Corpus: A Comparison of NMF, LDA, and BERTopic," *Proc. FedCSIS*, pp. 649–653, 2024.

- [7] D. Suryadi and K. O. Padlan, "Leveraging Topic Modeling and Sentiment Analysis to Improve Digital Bank Applications," in *Proc. ICDABI / IEEE*, 2024, doi: 10.1109/icdabi63787.2024.10799999.
- [8] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2022.
- [9] L. Ma et al., "AI-powered topic modeling: comparing LDA and BERTopic in analyzing opioid-related cardiovascular risks in women," *Experimental Biology and Medicine*, vol. 250, 2025, doi: 10.3389/ebm.2025.10389.
- [10] A. Hanifah, M. Saputra, and R. Y. Fa'rifah, "Exploring User Sentiments and Adoption Barriers in E-Wallet Services in Indonesia using BERTopic on Play Store Reviews," in *Proc. ICERA / IEEE*, 2025, doi: 10.1109/icera66156.2025.11086632.
- [11] C. Alba, "ConText Mining: Complementing Topic Models with Few-Shot In-Context Learning to Generate Interpretable Topics," *IEEE CI-NLPSoMe Companion*, 2025.
- [12] Y. Liu, "Comparison of LDA and BERTopic in News Topic Modeling: A Case Study of The New York Times' Reports on China," *Pacific International Journal*, vol. 7, no. 3, 2024, doi: 10.55014/pij.v7i3.616.
- [13] M. Hanafi, I. N. Nugraha, and S. Adi, "Adoption of Various Topic Modelling Algorithm to Analysis Indonesian Tourism Customer Feedback," *Proc. ICORIS*, 2024, doi: 10.1109/ICORIS63540.2024.10903770.
- [14] A. Abolfadl et al., "AutoCluster, AutoTopicModeling, AutoTrendAnalysis: A Complete AutoML Pipeline for Predicting Emerging Trends," *Proc. ICCA*, 2025, doi: 10.1109/ICCA66035.2025.11430858. u et al., "Large scale incremental learning," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 374–382, 2019, doi: 10.1109/CVPR.2019.00046.
- [15] Sulistiyani D, Nurchayati D, and N. D. Handani D, "User experience of mobile banking application in Indonesia: New technology of banking," *Global Business and Finance Review (GBFR)*, vol. 29, no. 2, pp. 127–142, 2024, doi: 10.17549/gbfr.2024.29.2.127.