International Journal of Education, Vocational and Social Science
https://e-journal.citakonsulfindo.or.id/index.php/IJEVSS
E-ISSN : 2963 4954
IJEVSS

Comparing Latent Dirichlet Allocation and Non-Negative Matrix Factorization for Topic Modeling of User Complaints in the Access by KAI Application

Valen Tino Rosi
Bina Darma University, Indonesia
Ari Muzakir
Bina Darma University, Indonesia
A. Haidar Mirza
Bina Darma University, Indonesia
Fitri Purwaningtias
Bina Darma University, Indonesia

Corresponding Author: 221410142@student.binadarma.ac.id

Abstract

User reviews on the Google Play Store hold valuable information about technical complaints, yet they arrive as unstructured short text that is impractical to examine manually. This study compares Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) for modeling complaint topics in 51,770 one- and two-star reviews of the Access by KAI application. Both algorithms were trained over the range of two to ten topics on an identical preprocessing pipeline and were evaluated using two coherence metrics and Topic Diversity. Robustness was examined through four test scenarios covering preprocessing quality, random initialization, and topic granularity, and the resulting topics were validated by two independent annotators. NMF outperformed LDA across all nine topic-number scenarios and on both coherence metrics, reaching a mean coherence of 0.5853 against 0.5273 for LDA, while LDA retained higher Topic Diversity. The advantage held after the normalization dictionary was expanded from 144 to 18,099 entries and across five random seeds. Human validation produced a mean readability score of 3.63 for NMF against 3.50 for LDA, with a Cohen's Kappa of 0.750. A document-level cross-mapping showed that each LDA topic corresponds to a dominant NMF topic at an average rate of 67.0%, indicating that both algorithms recover the same underlying themes.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

The four-topic NMF model is therefore recommended for complaint categorization on Indonesian short text.

Keywords: Topic Modeling, Latent Dirichlet Allocation, Non-Negative Matrix Factorization, Short Text, Topic Coherence.

INTRODUCTION

PT Kereta Api Indonesia (Persero) develops the Access by KAI application as an integrated platform for ticket ordering, schedule checking, and access to various supplementary services. Given its role for millions of active users, periodic evaluation of service quality is essential to maintain customer satisfaction and service competitiveness (Dwi Cahya et al., 2025). In their interaction with the application, users proactively provide feedback through reviews and ratings on the Google Play Store. Such reviews constitute a valuable data asset because they contain opinions, reports of technical problems, and development suggestions expressed directly by the user side (Mustaqim & Suryono, 2025).

The main challenge lies in the unstructured format of this data and in its volume, which grows daily. Manually identifying problems across tens of thousands of review entries is highly inefficient, prone to bias, and time consuming (Indrawan et al., 2023). The characteristics of the corpus compound this challenge. The final corpus in this study consists of 51,770 reviews with an average of only 10.62 tokens per document after preprocessing, and 57.6% of these documents contain nine tokens or fewer. This condition yields a highly sparse document-term matrix, a situation known to hinder topic modeling algorithms that estimate per-document probability distributions (Albalawi et al., 2020).

Topic modeling is an unsupervised machine learning technique capable of discovering hidden thematic structures in a document collection without requiring manual labeling (Blei et al., 2003; Jelodar et al., 2018). The two most widely used algorithms are Latent Dirichlet Allocation, which operates through probability distributions, and Non-Negative Matrix Factorization, which operates through matrix factorization. Studies comparing the two have largely been conducted on English corpora (Egger & Yu, 2022), whereas studies on Indonesian data have generally applied only a single algorithm without a comparator (Kusumaningtyas et al., 2024; Rosalinda et al., 2023). Comprehensively comparing these two algorithms on low-rated Indonesian reviews is therefore still relevant.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

Two gaps emerge from this review. First, comparisons of LDA and NMF on low-rated Indonesian reviews remain very limited. Second, and more fundamentally, prior studies evaluate both algorithms solely through coherence metrics without examining the consistency of document assignment. Coherence, however, measures only the quality of the keyword list, not the decisiveness of the model in assigning documents to topics. This study addresses both gaps.

This study therefore offers three contributions. First, a direct comparison of LDA and NMF on 51,770 Indonesian complaint reviews evaluated using two coherence metrics and Topic Diversity. Second, verification of the robustness of the findings through four separate test scenarios covering preprocessing quality, random initialization, and topic granularity. Third, a document-level cross-mapping analysis that measures how decisively each algorithm assigns documents to topics, complementing the coherence-based evaluation commonly used.

LITERATURE REVIEW

Latent Dirichlet Allocation is a generative probabilistic model that assumes each document is a mixture of topics and each topic is a distribution over words (Blei et al., 2003). The model infers latent topics by iteratively assigning words to topics and adjusting those assignments to maximize the likelihood of the observed data. Because the estimation is performed per document, LDA requires a sufficient number of word occurrences within each document to produce a reliable distribution, a requirement that short text often fails to meet (Jelodar et al., 2018).

Non-Negative Matrix Factorization takes a different route. It decomposes the document-term matrix into two lower-dimensional matrices whose elements are all non-negative (Lee & Seung, 2000). Because the decomposition works on word co-occurrence patterns across the entire corpus rather than on per-document estimation, NMF is generally more stable on sparse data. Its non-negativity constraint also makes the resulting components easier to interpret as topics.

Several studies have compared the two algorithms on short text. (Egger & Yu, 2022) compared four topic modeling algorithms on Twitter posts and found that LDA performs better on long text, whereas NMF yields more stable coherence scores on unstructured short text. (Nanyonga et al., 2023) obtained higher coherence for NMF than for LDA on aviation accident reports. (Zoya et al., 2021) reported a similar pattern on Urdu tweets, suggesting that the advantage of NMF on short text is not confined to a single language. (Albalawi et al., 2020)

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

specifically highlighted the challenge of data sparsity in short text and showed that different algorithms perform better under different data conditions.

In the Indonesian context, topic modeling research generally stops at a single algorithm. (Kusumaningtyas et al., 2024) applied LDA to negative reviews of an M-Banking application and demonstrated that the algorithm effectively uncovers technical complaint trends, but without a comparator algorithm. (Rosalinda et al., 2023) applied LDA to Netflix application reviews using a similar approach, and (Nugraha et al., 2023) carried out sentiment analysis of Satu Sehat application reviews with a comparable single-algorithm design. (Pratiwi & Tania, 2025) are an exception, as they compared BERTopic, LDA, and NMF on GoPartner reviews and found NMF to be highly competitive in both speed and the extraction of specific keywords from short text.

Evaluation of topic models relies primarily on coherence, which measures the semantic similarity between the highest-weighted words in a topic. (Röder et al., 2015) introduced the C_v measure, which combines a sliding window, one-set segmentation, and normalized pointwise mutual information. (Bouma, 2009) formulated the normalized pointwise mutual information underlying the C_{NPMI} measure. (Lau et al., 2014) showed that automatic coherence measures correlate strongly with human judgment, which justifies their use as the primary criterion. Because coherence alone can reward models that repeat the same words across topics, it is commonly complemented by Topic Diversity, the proportion of unique words among the top words of all topics (Dieng et al., 2020).

RESEARCH METHOD

This study follows the Cross-Industry Standard Process for Data Mining, adapted into six stages: problem understanding, data collection, preprocessing, modeling, evaluation, and deployment as a prototype. The research is quantitative and comparative, with the object of comparison being the performance of two topic modeling algorithms on the same corpus.

Data Collection

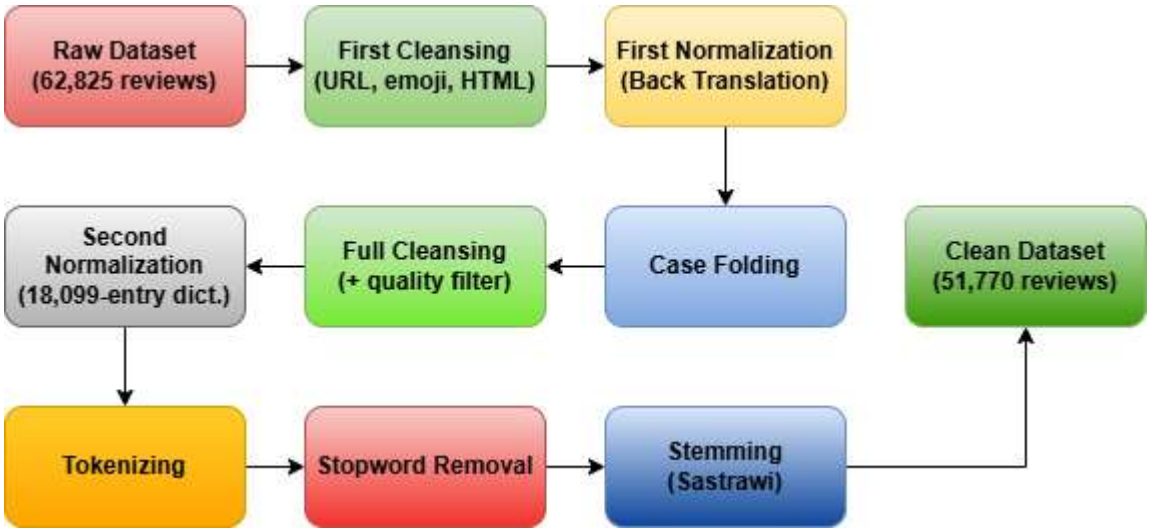
Access by KAI application reviews were retrieved from the Google Play Store using web scraping, filtered to one- and two-star ratings within a time restriction, yielding 62,825 raw reviews. Restricting the analysis to low ratings was applied because this study focuses on complaint identification rather than on general sentiment analysis.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

Preprocessing

The data passed through eight sequential stages, namely initial cleansing, back-translation normalization, case folding, full cleansing, dictionary-based normalization, tokenization, stopword removal, and stemming, as shown in Figure 1. Of the 62,825 raw reviews, the first cleansing stage removed URLs, HTML tags, emoji, and control characters without discarding any rows, so the corpus remained intact through back translation and case folding. The first reduction occurred at the full cleansing stage, where the back-translated text was cleaned again and 5,953 reviews left with fewer than four words were filtered out as too short to model. Dictionary normalization, which combined a curated dictionary compiled by the authors with a large-scale normalization dictionary totaling 18,099 entries, did not reduce the number of documents. Stopword removal then discarded a further 151 reviews whose text became empty because every word was a stopword, leaving 56,721 reviews; subsequent filtering of documents with fewer than three tokens left a final corpus of 51,770 reviews. Careful preprocessing matters here, as the quality of normalization has been shown to substantially affect the downstream analysis of Indonesian user-generated text (Khairunnisa et al., 2021).

Figure 1. The eight-stage preprocessing pipeline.



The final corpus contains 549,917 tokens drawn from 7,971 unique words, of which 48.0% occur only once across the entire corpus. The average document length is 10.62 tokens, and

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

57.6% of documents contain nine tokens or fewer. These characteristics render the corpus highly sparse and form the theoretical basis for comparing the two algorithms in this study.

Feature Extraction

LDA uses a Bag of Words representation via CountVectorizer, whereas NMF uses Term Frequency-Inverse Document Frequency weighting. Both vectorizations were configured with identical parameters, namely a minimum document frequency threshold of five and a maximum threshold of 0.5, and they share the same vocabulary so that the comparison remains fair. The choice of representation is not neutral, since different text representation schemes have been shown to yield different performance on Indonesian review data (Heri Suroyo & Pratama, 2025). This process produces a document-term matrix of 51,770 documents and 2,381 features.

Modeling

Both algorithms were trained over a topic range of two to ten. Because LDA is non-deterministic, each scenario was rerun five times with random state values of 42, 123, 456, 789, and 1011, and the mean was reported together with its standard deviation. NMF with nndsvda initialization is deterministic and therefore required only a single run. This treatment ensures that any reported advantage is the average of five trials rather than the outcome of a single run that might happen to be favorable.

Evaluation

Evaluation was carried out through four complementary test scenarios. The first scenario compares both algorithms across all topic numbers using Coherence C_v as the primary criterion, reinforced by C_{NPMI} and Topic Diversity. The second scenario examines the effect of preprocessing quality by comparing two normalization dictionary configurations. The third scenario tests stability against random initialization, and the fourth compares topic granularity to determine the final model.

Human validation was performed by two independent annotators on the ten highest-ranked keywords of each topic. The identity of the algorithm was concealed and the topic order was shuffled so that the annotators did not know which model was being evaluated. Each annotator provided a thematic label and a readability score on a scale of one to four, with the level of agreement measured using Cohen's Kappa (Cohen, 1968). The kappa value was

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

interpreted using the criteria of (Landis & Koch, 1977) and of (Altman, 1991), and was read together with the raw agreement percentage as recommended for reliability studies with limited sample sizes (McHugh, 2012).

RESULTS AND DISCUSSION

Coherence Comparison

Table 1. Coherence and Topic Diversity Values for LDA and NMF at $k = 2$ to 10

k	C_v LDA	C_v NMF	C_{NPMI} LDA	C_{NPMI} NMF	Div LDA	Div NMF
2	0.5158	0.5775	0.0488	0.0837	0.960	0.900
3	0.5217	0.6148	0.0456	0.0954	0.915	0.853
4	0.5429	0.5990	0.0516	0.0955	0.872	0.830
5	0.5242	0.5815	0.0437	0.0851	0.854	0.776
6	0.5195	0.5748	0.0422	0.0775	0.832	0.753
7	0.5325	0.5834	0.0477	0.0819	0.811	0.697
8	0.5412	0.5777	0.0496	0.0803	0.799	0.660
9	0.5218	0.5834	0.0414	0.0835	0.796	0.609
10	0.5261	0.5754	0.0451	0.0776	0.790	0.576

The measurements in Table 1 and Figure 2 show that NMF outperforms LDA across all nine topic-number scenarios, on both the C_v and C_{NPMI} metrics. C_v values for NMF range from 0.5748 to 0.6148, whereas those for LDA range from 0.5158 to 0.5429. The C_{NPMI} values obtained by NMF, ranging from 0.0775 to 0.0955, are consistently almost twice those of LDA, which range from 0.0414 to 0.0516. Given that C_{NPMI} values in topic modeling practice are typically small positive numbers, a gap of this size constitutes a meaningful difference.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

International Journal of Education, Vocational and Social Science
https://e-journal.citakonsulfindo.or.id/index.php/IJEVSS
E-ISSN : 2963 4954
IJEVSS

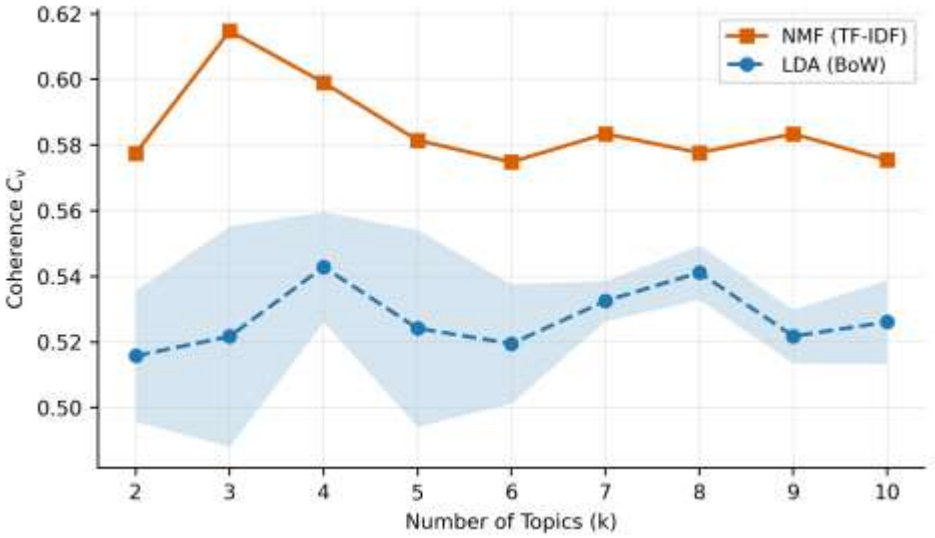


Figure 2. Coherence C_v values for LDA and NMF at $k = 2$ to 10.

The opposite pattern occurs for Topic Diversity. LDA obtains higher values across all scenarios, and the NMF value declines sharply as the number of topics increases, from 0.900 at two topics to 0.576 at ten topics. This indicates a trade-off: NMF produces more coherent topics, but the more topics that are imposed, the more words recur across topics.

Determining the Number of Topics

The highest C_v value for NMF occurs at three topics, at 0.6148. Nevertheless, the four-topic configuration was chosen as the final model for three reasons. First, the C_v value at four topics, 0.5990, differs by only 0.0158 and is therefore practically equivalent. Second, this configuration in fact attains the highest C_{NPMI} value among all scenarios, at 0.0955. Third, in the three-topic configuration a single topic absorbs 51.2% of the documents while conflating two distinct themes, namely ticket ordering and application performance, so that complaints regarding application stability have no category of their own.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

International Journal of Education, Vocational and Social Science
https://e-journal.citakonsulfindo.or.id/index.php/IJEVSS
E-ISSN : 2963 4954
IJEVSS

Robustness of the Findings

Table 2. Effect of Expanding the Normalization Dictionary on the Mean Metric Values

Metric	144 entries	18,099 entries	Difference
C_v NMF	0.5807	0.5853	+0.0046
C_{NPMI} NMF	0.0850	0.0845	−0.0005
C_v LDA	0.5330	0.5273	−0.0058
C_{NPMI} LDA	0.0471	0.0462	−0.0009
Div NMF	0.7358	0.7394	+0.0036
Div LDA	0.8500	0.8477	−0.0023

Expanding the normalization dictionary from 144 to 18,099 entries altered the coherence values by an average of 0.0142 points, with a largest change of 0.0355 points. Nevertheless, the ranking between the models did not change at all: NMF continued to outperform LDA across all nine scenarios. This finding indicates that the advantage of NMF is robust to changes in the preprocessing configuration.

Testing with five random initialization values produced C_v standard deviations for LDA ranging from 0.0061 to 0.0336, with an average coefficient of variation of 3.26%. The upper bound of the LDA range across all configurations remains below the C_v value of NMF, so the advantage of NMF is statistically consistent and not the result of a single fortunate run.

Topic Structure of the Final Model

Table 3. Label, Document Share, and Ten Highest-Ranked Keywords of the Final NMF Model at k = 4

Topic	Label	Share	Ten Highest-Ranked Keywords
1	Ticket Ordering and Availability	25.5%	tiket, pesan, beli, kereta, lokal, muncul, jadwal, batal, habis, susah
2	Login and Account Registration	21.9%	daftar, login, email, akun, masuk, nomor, susah, password, benar, verifikasi
3	Payment and Balance	18.3%	bayar, saldo, pakai, gagal, link, laku, lewat, isi, metode, potong
4	Application Performance and Bugs	34.3%	update, lambat, terus, buka, buruk, error, bagus, versi, mohon, waktu

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

Table 3 shows that the four-topic configuration successfully separates ticket ordering from application performance into two distinct categories. The document distribution is far more balanced, namely 25.5% for Ticket Ordering and Availability, 21.9% for Login and Account Registration, 18.3% for Payment and Balance, and 34.3% for Application Performance and Bugs. Several high-frequency filler words with low discriminative power, such as mohon and terus, remain prominent and constitute one of the limitations noted at the end of this article.

Human Validation

Table 4. Human Validation Results and Inter-Annotator Reliability

Measure	Model A (LDA)	Model B (NMF)
Mean score, Annotator 1	3.50	3.75
Mean score, Annotator 2	3.50	3.50
Combined mean	3.50	3.63
Label agreement	4 of 4	4 of 4
Cohen's Kappa	0.750	—
Weighted Kappa (linear)	0.750	—
Exact agreement	87.5%	—
Pearson correlation	0.775	—

Two independent annotators evaluated eight topics blindly, namely four topics from each algorithm in the four-topic configuration. NMF obtained a mean readability score of 3.63 out of a maximum of four, whereas LDA obtained 3.50. The difference of 0.13 points indicates that at an equivalent level of granularity both models produce topics that are equally easy for humans to read, with a slight advantage for NMF.

Inter-annotator agreement reached a Cohen's Kappa of 0.750, a value that falls within the substantial agreement range according to the criteria of Landis and Koch (1977) and of Altman (1991), with an exact agreement of 87.5%, and no single rating differed by more than one level. The linear weighted kappa is likewise 0.750, because both annotators used only the two highest categories on the scale so that all disagreements carry the same distance. All four NMF topic labels were semantically agreed upon by both annotators. Because the number of rated items is small, this kappa value is better read together with the raw agreement percentage (McHugh, 2012).

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

International Journal of Education, Vocational and Social Science
https://e-journal.citakonsulfindo.or.id/index.php/IJEVSS
E-ISSN : 2963 4954
IJEVSS

Cross-Mapping of Documents

To obtain quantitative evidence regarding the decisiveness of document assignment, each document mapped to an LDA topic was traced to its assignment among the NMF topics. Figure 3 shows that the four LDA topics each have a dominant and mutually distinct NMF counterpart, with dominance levels of 63.0%, 75.2%, 70.6%, and 59.1%. This one-to-one mapping constitutes important cross-validity evidence, since two algorithms operating on different mathematical principles recover the same four core themes from the same corpus.

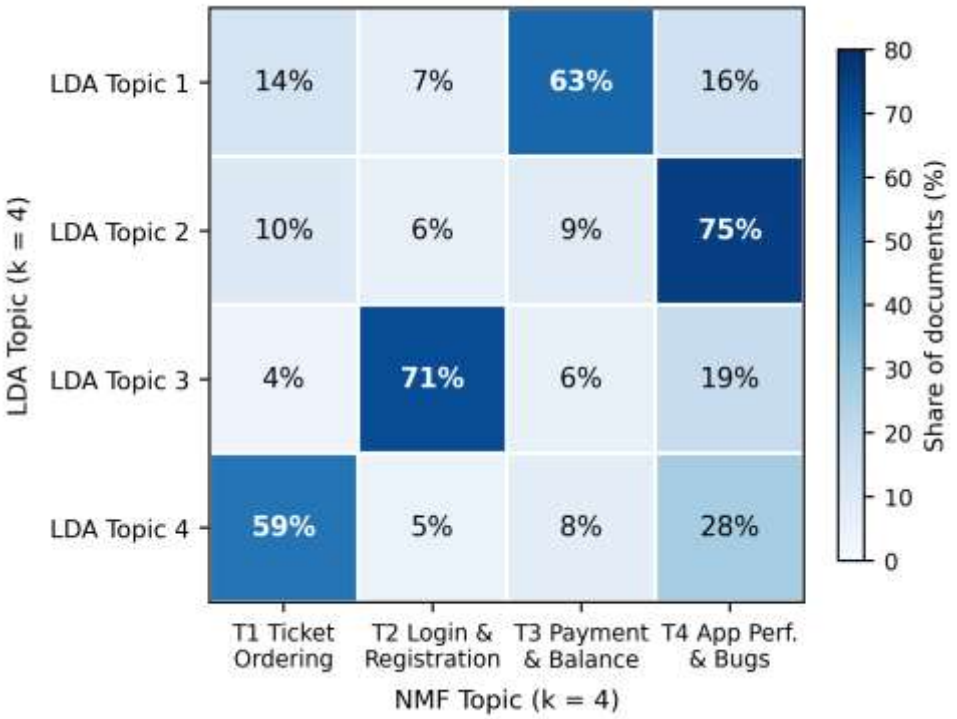


Figure 3. Distribution of documents from each LDA topic into the NMF topics.

Nevertheless, this agreement is far from perfect. On average only 67.0% of documents occupy the dominant cell, so approximately one in three documents is assigned to a different category by the two models. The most striking case occurs in the fourth LDA topic. All ten of its keywords relate to ticket ordering, yet 27.7% of its documents are mapped by NMF into the Application Performance and Bugs category. Conversely, the first LDA topic is already mixed at

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

the level of its keyword list, since it contains bayar and saldo alongside tiket and duduk, and consequently 37.0% of its documents are scattered across three other categories.

This finding is independently reinforced by the human validation results. The first LDA topic was the only topic to which both annotators assigned a label containing two themes at once, namely payment and seat selection, and it also received the lowest readability score within the LDA model. The conflation of themes in that topic is thus confirmed through three mutually independent sources: the keyword list, the document distribution, and human interpretation.

DISCUSSION

The advantage of NMF can be explained through the characteristics of the data and the architecture of the algorithm. Reviews on the Google Play Store are short text and therefore yield a highly sparse document-term matrix. Under such conditions, LDA, which estimates a topic probability distribution for each document, lacks statistical evidence, since a ten-token document does not provide enough word occurrences to estimate a distribution reliably. NMF, by contrast, works through factorization of the TF-IDF matrix and does not depend on per-document estimation but rather on word co-occurrence patterns across the entire corpus. This characteristic theoretically explains why NMF prevails across all tested scenarios.

The direction of these findings is consistent with prior work comparing the two algorithms on short-text data. (Egger & Yu, 2022), (Nanyonga et al., 2023), and (Zoya et al., 2021) reported the same pattern in three separate domains and languages. The advantage of NMF demonstrated here is thus not an isolated result, but has been independently reported across different corpora. The principal difference of this study lies in the document-level cross-mapping analysis, which provides a new dimension of evaluation beyond coherence metrics.

Prototype Implementation

As a realization of the deployment stage, the final model was operationalized into a web-based automatic labeling prototype. The aim is to demonstrate that topic modeling output, initially a list of abstract keywords, can be translated into categories directly usable by non-technical stakeholders. The four-topic NMF model together with the trained TF-IDF vectorizer was exported and used during inference. Because the vectorizer recognizes only the 2,381 features formed during training, identical preprocessing between the training and inference

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

stages is an absolute requirement; even the smallest difference in the normalization dictionary would cause a vocabulary mismatch and ultimately produce invalid predictions.

The prototype provides two modes of operation. In single-review mode, a review entered by the user is mapped to one of the four categories and displayed together with a confidence level and a text transformation panel, so that users can verify that preprocessing proceeds as designed. In bulk mode, the system processes a file of reviews and presents an aggregate analysis comprising the category distribution, the trend in complaint volume over time, and the most frequently occurring keywords in each category. Reviews that remain too short after preprocessing receive no category, a direct consequence of the short-text characteristics discussed earlier.

Limitations

This study has three principal limitations. The number of topics evaluated by humans is relatively small, and one of the annotators is the researcher, so the agreement value is better regarded as a preliminary indication. In addition, topic assignment is hard, so reviews falling outside the four categories are necessarily forced into one of them. Finally, several high-frequency filler words with low discriminative power remain prominent in the resulting topics.

CONCLUSION

This study compared LDA and NMF for modeling complaint topics in 51,770 low-rated reviews of the Access by KAI application. NMF with TF-IDF weighting proved more effective than LDA with a Bag of Words representation, prevailing on both coherence metrics across all nine topic-number scenarios, and this advantage persisted under changes to the preprocessing configuration as well as under random initialization. Human validation by two independent annotators supports this finding, with readability scores of 3.63 against 3.50 and a Cohen's Kappa of 0.750. The cross-mapping analysis shows that although both algorithms recover the same four themes, LDA assigns documents less decisively. The final NMF model with four topics maps the complaints into ticket ordering, account registration, payment, and application performance.

Three directions for further research can be pursued. First, human validation should involve annotators drawn entirely from outside the study with broader topic coverage, so that agreement measured using Cohen's Kappa or Fleiss' Kappa yields more convincing results. Second, hard topic assignment could be replaced by a threshold-based approach so that reviews

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

outside the four categories are not forced into one of them. Third, the comparison could be extended to embedding-based models such as BERTopic to test whether the advantage of NMF on Indonesian short text persists.

REFERENCES

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Frontiers in Artificial Intelligence*, *3*, 42. https://doi.org/10.3389/frai.2020.00042

Altman, D. G. (1991). *Practical statistics for medical research*. Chapman and Hall.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bouma, G. (2009). Normalized (Pointwise) Mutual Information in Collocation Extraction. *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, 31–40.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220.

Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2020). Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, *8*, 439–453. https://doi.org/10.1162/tacl_a_00325

Dwi Cahya, F. D., Satyareni, D. H., & Farhan, A. (2025). Analisis Kinerja Aplikasi Access by KAI Menggunakan Metode End-User Computing Satisfaction Dan Delone & McLean. *Jurnal Ilmiah Sistem Informasi dan Teknik Informatika (JISTI)*, *8*(2), 153–161. https://doi.org/10.57093/jisti.v8i2.286

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, *7*, 886498. https://doi.org/10.3389/fsoc.2022.886498

Heri Suroyo, & Pratama, E. J. (2025). Comparison of Text Representation Methods for Sentiment Analysis Using Support Vector Machine. *Journal of Advances in Information and Industrial Technology*, *7*(1), 21–30. https://doi.org/10.52435/jaiit.v7i1.610

Indrawan, H., Irawan, B., & Suprapti, T. (2023). *KLASIFIKASI ULASAN PENGGUNA APLIKASI ACCESS BY KAI BERBASIS ASPEK DENGAN ALGORITMA NAÏVE BAYES DAN SVM*. *7*(6).

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2018). *Latent Dirichlet Allocation (LDA) and Topic modeling: Models, applications, a survey* (arXiv:1711.04305). arXiv. https://doi.org/10.48550/arXiv.1711.04305

Khairunnisa, S., Adiwijaya, A., & Faraby, S. A. (2021). Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19). *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(2), 406. https://doi.org/10.30865/mib.v5i2.2835

Kusumaningtyas, K., Dwijayanti, I., Rizqi Lahitani, A., & Habibi, M. (2024). Analisis Tren Topik dalam Ulasan Negatif Aplikasi M-Banking Menggunakan Latent Dirichlet Allocation. *JURNAL FASILKOM*, *14*(3), 549–555. https://doi.org/10.37859/jf.v14i3.8035

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 530–539. https://doi.org/10.3115/v1/E14-1056

Lee, D. D., & Seung, H. S. (2000). Algorithms for Non-negative Matrix Factorization. *Advances in Neural Information Processing Systems 13*, 556–562.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Mustaqim, I. Z., & Suryono, R. R. (2025). A Systematic Literature Review of Topic Modeling Techniques in User Reviews. *Journal of Information Systems Engineering and Business Intelligence*, *11*(2), 238–253. https://doi.org/10.20473/jisebi.11.2.238-253

Nanyonga, A., Wasswa, H., & Wild, G. (2023). Topic Modeling Analysis of Aviation Accident Reports: A Comparative Study between LDA and NMF Models. *2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, 1–2. https://doi.org/10.1109/SMARTGENCON60755.2023.10442471

Nugraha, M. T., Sulistiyowati, N., & Enri, U. (2023). *ANALISIS SENTIMEN ULASAN APLIKASI SATU SEHAT PADA GOOGLE PLAY STORE MENGGUNAKAN NAÏVE BAYES CLASSIFIER*. *7*(5).

Pratiwi, M. D., & Tania, K. D. (2025). Knowledge Discovery Through Topic Modeling on GoPartner User Reviews Using BERTopic, LDA, and NMF. *Journal of Applied Informatics and Computing*, *9*(1), 1–7. https://doi.org/10.30871/jaic.v9i1.8782

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. https://doi.org/10.1145/2684822.2685324

Rosalinda, G., Santoso, R., & Kartikasari, P. (2023). PEMODELAN TOPIK ULASAN APLIKASI NETFLIX PADA GOOGLE PLAY STORE MENGGUNAKAN LATENT DIRICHLET ALLOCATION. *Jurnal Gaussian*, *11*(4), 554–561. https://doi.org/10.14710/j.gauss.11.4.554-561

Zoya, Latif, S., Shafait, F., & Latif, R. (2021). Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling. *IEEE Access*, *9*, 127531–127547. https://doi.org/10.1109/ACCESS.2021.3112620

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License