Analyzing Topic Trends in HRIS User Feedback Using BERTopic for Data-Driven System Evaluation

Ade Iskandar, Eko Hariyanto

Abstract

This study examines temporal changes in Human Resource Information System (HRIS) user feedback and differences between complaints and suggestions. We audited 5,228 raw records and analyzed 4,509 anonymized, deduplicated records dated 7 December 2022 to 5 October 2026. We modeled topics with BERTopic using multilingual sentence embeddings. Category labels were normalized into complaint (3,136) and suggestion (1,373). The study identified 22 topics and 1,022 outliers (22.7%). Its largest cluster contained 1,729 records (38.3%) and mixed clock and attendance terms, so it was not assigned a narrow issue label. Among more interpretable clusters, inability to record attendance (topic 1) increased from 5.8% of feedback in 2023 to 9.0% in 2025, while face detection (topic 10) increased from 0.7% to 3.8%; slow loading (topic 2) declined from 3.3% to 1.3%. These descriptive shares suggest issues for review, not evidence that a product change caused them. The broad largest cluster and lack of human label validation limit interpretation.

Keywords: HRIS; user feedback; BERTopic; temporal topic trends; system evaluation

Ade Iskandar
Univeristas Pembangunan Panca Budi, Indonesia
e-mail: suratiskandar@gmail.com

Eko Harianto
e-mail: eko.hariyanto@dosen.pancabudi.ac.id .
3rd International Conference Epicentrum of Economic Global Framework (ICEEGLOF)
Theme: Cross-Sector and Cross-Disciplinary Strategies for Achieving the SDGS
https://proceeding.pancabudi.ac.id/index.php/ICEEGLOF

Introduction

Digital HRIS platforms support day-to-day employee administration, yet the experience of individual users is not fully captured by system-availability indicators or aggregate transaction counts. Employees report attendance failures, location and facial-verification problems, leave requests, scheduling issues, and requests for enhancements in unstructured language. Such feedback contains operational knowledge that may complement formal testing and service monitoring. A systematic review of software-app review mining identifies user-reported defects, requested features, and product evolution as important software-engineering applications [7]. Accordingly, feedback should be examined as evidence of what users report rather than treated as a direct measure of the technical incidence or severity of failures.

Manual reading becomes increasingly difficult as reports accumulate over several years. Simple category totals cannot reveal whether a particular concern has newly emerged, persisted, or declined. The temporal dimension matters to product owners because a topic that dominates historic records may differ from an issue that is growing in more recent periods. Comparing relative topic shares across consistent time windows offers a descriptive view that is less sensitive than raw counts to fluctuations in submission volume. Nevertheless, such shares remain conditional on submitted feedback, not on the complete population of active HRIS users.

Topic modeling provides a way to organize large collections of unstructured text. Traditional probabilistic topic modeling, exemplified by latent Dirichlet allocation [8], offers a useful conceptual foundation, while embedding-based approaches can capture semantic proximity among texts that do not share identical terms. BERTopic combines document embeddings, dimensionality reduction, clustering, and class-based TF-IDF to construct interpretable topic representations [1]. This is relevant for short, informal HRIS feedback in which users may describe closely related difficulties with different vocabulary. It does not imply that BERTopic necessarily produces valid or superior topics on this particular corpus without independent evaluation.

This study analyzes 4,509 preprocessed feedback records from one HRIS, spanning December 2022 to October 2026. It asks three questions: (RQ1) What interpretable topics can be identified? (RQ2) How do the proportions of selected topics change across complete calendar years and quarterly periods? (RQ3) How do topic distributions differ across the feedback categories available in the analyzed workbook? The contribution is a reproducible, descriptive temporal and category-based view for data-driven HRIS review, rather than a causal assessment of software releases or a benchmark claiming that BERTopic outperforms other methods.

Literature Review

Text mining of software feedback has become an established means of extracting requirements and fault reports from otherwise unstructured accounts. A systematic literature review [7] synthesizes research on mining application reviews for software engineering and highlights feature analysis, defect detection, and product evolution. HRIS feedback has similar informational potential, although organizational feedback channels and employee populations differ from public application stores. The present study therefore uses review-mining literature as methodological context, without assuming equivalent reporting behavior.

Topic modeling summarizes recurring themes across documents. Latent Dirichlet allocation models documents probabilistically as mixtures of latent topics [8]; by contrast, BERTopic produces document embeddings, clusters semantically related items, and derives topic descriptions through class-based TF-IDF [1]. Sentence-BERT research provides a foundation for deriving sentence-level semantic representations that can be compared efficiently [9]. UMAP contributes nonlinear dimensionality reduction [10], while density-based clustering can separate compact groups from observations that do not clearly belong to a cluster.

Each algorithmic stage introduces modeling choices that may influence which issues appear as coherent topics, particularly in short texts and highly repetitive attendance terminology.

For longitudinal analysis, a stable global topic assignment permits comparable counting of records per topic across calendar periods. BERTopic's topics-over-time analysis can additionally examine how topic representations change, but frequency trends must be interpreted carefully: a rising share means a larger fraction of submitted records belongs to a topic, not that the underlying incident rate rose. Topic labels also remain interpretive. Research on topic coherence demonstrates why interpretability should be evaluated rather than inferred solely from model output [11], [12]. In this dataset, a broad attendance-related cluster and a sizable outlier proportion make such caution particularly important.

Prior studies show several ways to analyze short user comments. Information extraction from social-media text has been used for student-risk classification [2], while Naive Bayes and LSTM have been applied to student feedback about services and learning [3]. NLP and SVM have been used to classify sentiment in e-commerce reviews [4], and lexicon-based and BERT methods have been compared on mobile-banking reviews [5]. A comparison of LDA and BERTopic on mobile-banking reviews illustrates topic modeling of short user comments [6]. Building on this literature, the present study examines how HRIS feedback topics vary over time and between complaint and suggestion categories.

The research gap is therefore practical and analytical: this paper goes beyond one-time categorization by linking discovered HRIS feedback topics to consistent time periods and the categories actually present in the processed workbook. The study does not claim a novel clustering algorithm; its intended contribution is an auditable reporting framework that helps product owners identify which reported concerns warrant further manual inspection and possible corroboration against release history and usage data.

Research Methodology

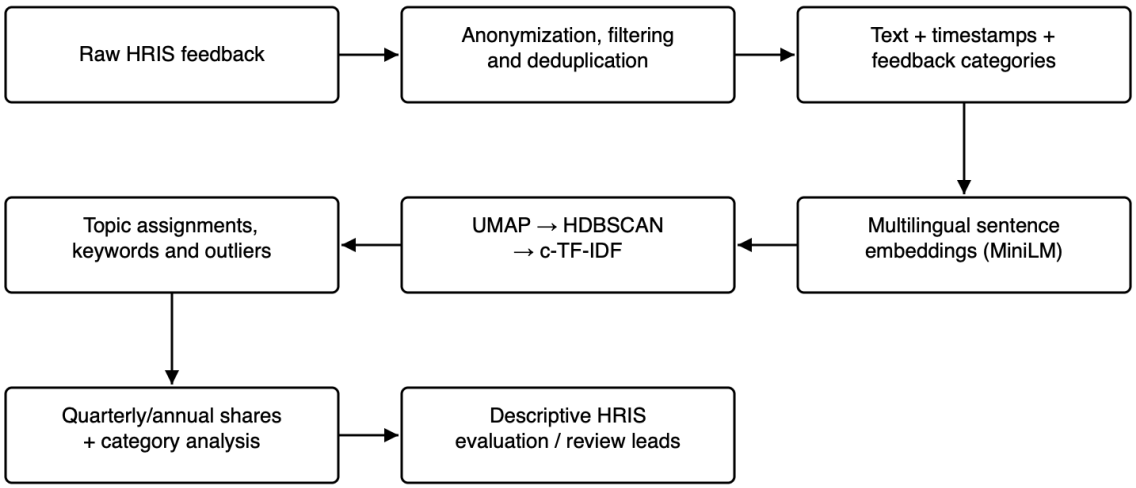


Figure 1. Research Methodology Workflow Using BERTopic

The source workbook contains 5,228 user-feedback records from Nusawork, the HRIS examined in this study. Preprocessing masked URLs, email addresses, monetary amounts, selected personal names, phone and identity patterns, company names, and application-version metadata in descriptions. Feedback with empty or fewer-than-three-word cleaned text was excluded; a small predefined junk list was filtered and duplicate cleaned_text values were removed. The processed workbook contains 4,509 unique feedback IDs and cleaned texts; 719 rows were removed by the combined filtering and deduplication steps. The processed workbook retains original columns, but only cleaned_text, created_at, category, and feedback_id were used analytically. No record-level text was exported.

Category variants Keluhan/Complain/complain were standardized as complaint and Saran/Suggestion/suggestion as suggestion. BERTopic was applied to the 4,509 processed feedback texts. Its workflow in this study was: feedback document → multilingual sentence embedding (MiniLM) → UMAP → HDBSCAN → c-TF-IDF → topic and keywords [1]. The embeddings capture semantic similarity; UMAP reduces their dimensions; HDBSCAN groups similar feedback and can mark ambiguous records as outliers; and c-TF-IDF identifies terms that distinguish each topic. These are stages of the single BERTopic model. The BERTopic analysis used BERTopic 0.17.4 with UMAP set to 15 neighbors, five dimensions, cosine distance, and seed 42; HDBSCAN used a minimum cluster size of 30 and minimum samples of 10; the topic representation used unigrams and bigrams with minimum document frequency 5. After topic assignment, each topic was counted per quarter and divided by all feedback in that quarter, including outliers in the denominator. Annual comparisons use the complete years 2023–2025; 2022 and 2026 are partial years. Complaint and suggestion patterns were compared using within-category topic shares. Changes in representative topic terms were inspected across 16 time intervals. Topic names are provisional interpretations and have not been validated by human annotators.

Results

The processed data span 7 December 2022 to 5 October 2026. Complaints account for 3,136 records (69.6%) and suggestions for 1,373 (30.4%). The annual counts are 838 (2022, partial), 808 (2023), 981 (2024), 1,304 (2025), and 578 (2026, partial). The BERTopic model assigned 3,487 records to 22 topics and left 1,022 (22.7%) as outliers. Topic 0 alone contains 1,729 records (38.3% of all feedback) and has a broad clock/attendance vocabulary. The next two clusters contain 351 (7.8%) and 107 (2.4%) records, respectively.

Table 1. Shares of selected interpretable topics in complete calendar years (% of all processed feedback)

Topic and interpretation	2023	2024	2025
1 Unable to record attendance	5.8	8.2	9.0
2 Slow loading	3.3	3.6	1.3
9 Login	2.4	1.9	1.8
10 Face detection	0.7	1.8	3.8

Topic 1 rose by 3.2 percentage points from 2023 to 2025, while topic 10 rose by 3.1 points. Topic 2 fell by 2.0 points and topic 9 by 0.5 points. These are changes in the fraction of submitted feedback, not incidence rates among all HRIS users. The final quarter of 2026 contains only seven records through 5 October and is excluded from year-to-year trend claims.

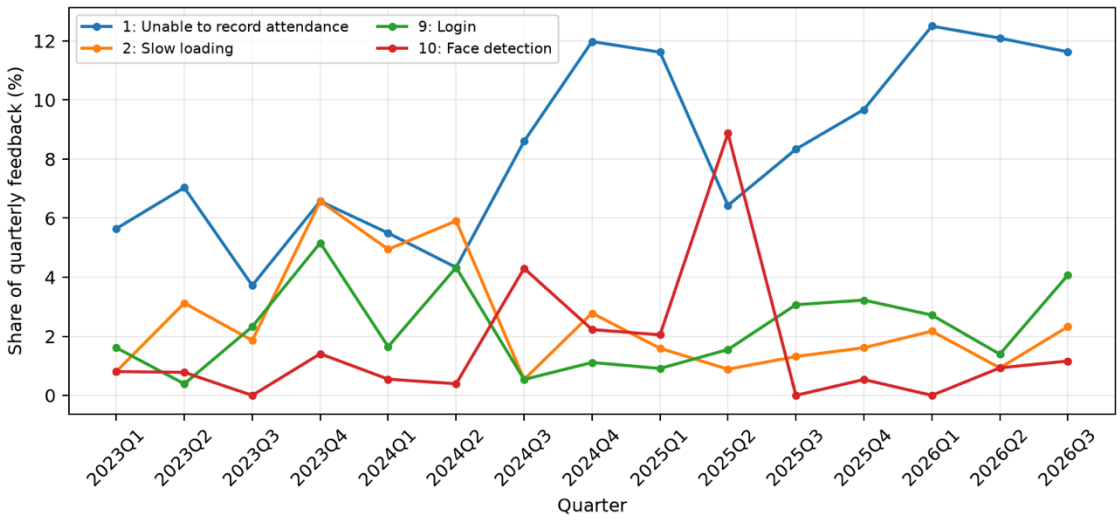


Figure 1. Quarterly shares of four selected interpretable topics, 2023Q1–2026Q3.

Table 2. Category comparison for selected interpretable topics (% within category)

Topic and interpretation	Complaint (n=3,136)	Suggestion (n=1,373)
1 Unable to record attendance	9.2	4.7
2 Slow loading	3.1	0.7
9 Login	2.2	0.8
10 Face detection	2.3	0.7

The selected interpretable topics occur proportionally more often in records labeled complaints than in suggestions. Topic 0, however, covers similar shares of complaints (38.3%) and suggestions (38.5%) and cannot be treated as one specific issue. Outlier rates are 20.2% of complaints and 28.3% of suggestions. These comparisons describe model assignments, not the full range of themes.

Discussion

The interpretable inability-to-record-attendance cluster rises in share from 2023 to 2025, as does the face-detection cluster. The slow-loading cluster declines, while the login cluster changes less. The largest cluster is too broad to support a claim about a single clock-in/clock-out trend. Product review should therefore examine attendance capture and face recognition separately, while manually inspecting the large mixed cluster before prioritizing it. Observed shares could reflect product behavior, changing user populations, submission channels, or category practices; the dataset does not identify a cause.

Although the outlier share is 22.7%, the largest cluster covers 38.3% of feedback and mixes common clock and attendance terms. This limits topic specificity and may mask several distinct issues. The processed file still contains original identity-bearing columns, and pattern-based anonymization cannot guarantee that every personal name was removed from cleaned_text. Representative quotations therefore require review before release. Human validation of topic assignments, sensitivity analysis across environments and BERTopic settings, and normalization by active-user counts are needed before treating these shares as performance indicators.

Conclusion

The BERTopic analysis identifies 22 topics and descriptive changes in interpretable HRIS feedback clusters: inability to record attendance and face detection rise in share across 2023–

2025, while slow loading declines. Complaint and suggestion labels have different topic distributions. The largest cluster remains broad and requires manual review before it can support a specific product conclusion. These findings can guide feedback triage but do not measure defect incidence or prove the impact of an intervention. Future work should validate labels with human annotators, examine the large cluster and outliers, and compare trends with usage and release histories.

References

[1] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv:2203.05794, 2022. https://arxiv.org/abs/2203.05794

[2] E. Hariyanto, S. Wahyuni, and S. Batubara, "Study of Potential Classification of Lost Students in College Based on Information Extraction on Text-Based Social Media; Case Study of Panca Budi Pembangunan University," International Journal of Research and Review, vol. 8, no. 11, pp. 325–331, 2021. doi:10.52403/ijrr.20211140.

[3] A. Rahman, Khairul, M. Iqbal, and A. P. U. Siahaan, "Sentiment Analysis of Student Feedback for Learning Improvement using Naive Bayes and LSTM (Long Short-Term Memory) Methods," International Conference Epicentrum of Economic Global Framework, pp. 2695–2701, 2026. [Online]. Available: https://proceeding.pancabudi.ac.id/index.php/ICEEGLOF/article/view/1850.

[4] J. I. W. Siregar, A. P. U. Siahaan, M. Iqbal, D. Nasution, and R. F. Wijaya, "Sentiment Analysis Classification of E-commerce User Reviews Using Natural Language Processing (NLP) and Support Vector Machine (SVM) Methods," Bulletin of Information Technology, vol. 6, no. 2, pp. 196–205, 2025. doi:10.47065/bit.v6i2.2018.

[5] S. M. Rambe, Z. Sitorus, and M. Iqbal, "Analisis Sentimen Pengguna terhadap Layanan Mobile Banking di Indonesia Menggunakan Lexicon-Based dan BERT," Jurnal Sistem Komputer dan Informatika (JSON), vol. 8, no. 1, pp. 320–329, 2026. doi:10.30865/json.v8i1.10188.

[6] D. Halim, I. Budi, B. F. Mubina, A. B. Santoso, and P. K. Putra, "Comparative Topic Modelling of Mobile Banking User Reviews Using LDA and BERTopic: A Case Study of wondr by BNI," The Indonesian Journal of Computer Science, vol. 15, no. 2, pp. 2236–2246, 2026. doi:10.33022/ijcs.v15i2.5118.

[7] J. Dąbrowski, E. Letier, A. Perini, and A. Susi, "Analysing app reviews for software engineering: A systematic literature review," Empirical Software Engineering, vol. 27, art. 43, 2022. doi:10.1007/s10664-021-10065-7.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003. https://www.jmlr.org/papers/v3/blei03a.html.

[9] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. EMNLP-IJCNLP, 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.

[10] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," arXiv:1802.03426, 2018. https://arxiv.org/abs/1802.03426.

[11] M. Röder, A. Both, and A. Hinneburg, "Exploring the Space of Topic Coherence Measures," in Proc. WSDM, 2015, pp. 399–408. doi:10.1145/2684822.2685324.

[12] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing Semantic Coherence in Topic Models," in Proc. EMNLP, 2011, pp. 262–272. https://aclanthology.org/D11-1024/.