Topic Discovery and Sentiment Mapping from Indonesian User Reviews Using NLP and Clustering Techniques

Iyus Rusmansyah[1], Deny Ariestiandy[2]
Department of Information System, Faculty of Computer Science, Citra Buana Indonesia Institute
Jl. Raya Cemerlang, Sukakarya, Warudoyong District, Sukabumi City 43135, West Java, Indonesia
Email: [1]afecto.rusmansyah24@gmail.com, [2]denyariestiandycbi@gmail.com

Abstrak. *Volume umpan balik pengguna yang tidak terstruktur dan masif pada gim seluler membuat analisis manual menjadi tidak praktis. Oleh karena itu, penelitian ini menemukan topik dominan dan memetakan sentimen dari 8.000 ulasan Honor of Kings menggunakan pipeline NLP dan clustering terintegrasi. Kami menggabungkan vektorisasi TF-IDF, clustering mini-batch k-means, dan pemodelan topik LDA, dilengkapi analisis sentimen berbasis leksikon. Hasil mengidentifikasi sepuluh segmen tematik yang mencerminkan preferensi pemain dan keluhan teknis. Penilaian clustering menghasilkan silhouette score 0,1071, yang tipikal untuk teks pendek berdimensi tinggi dan mengindikasikan batas yang lunak, sementara uji chi-square mengonfirmasi asosiasi signifikan antara klaster dan sentimen ($p < 0,001$, Cramér's $V = 0,242$). Temuan ini memberikan wawasan praktis bagi pengembang, meskipun terdapat keterbatasan berupa lemahnya pemisahan klaster dan rendahnya akurasi analisis sentimen pada teks informal.*
Kata Kunci: *analisis sentimen, clustering, Honor of Kings, LDA, ulasan pemain*

Abstract. *The massive volume of unstructured user reviews in mobile games makes manual analysis impractical. Therefore, our study discovers dominant topics and maps sentiments from 8,000 Honor of Kings reviews using an integrated NLP and clustering pipeline. We combine TF-IDF vectorization, mini-batch k-means clustering, and LDA topic modeling, complemented by lexicon-based sentiment analysis. Results identify ten thematic segments reflecting player preferences and technical complaints. Clustering assessment yielded a silhouette score of 0.1071, typical for high-dimensional short-text and indicating soft boundaries, while a chi-square test confirmed a significant association between clusters and sentiment ($p < 0.001$, Cramér's $V = 0.242$). These findings offer practical insights for developers, despite limitations regarding weak cluster separation and low sentiment accuracy in informal text.*
Keywords: *sentiment analysis, clustering, Honor of Kings, LDA, player reviews*

1. Introduction

The rapid growth of mobile gaming in Indonesia has generated substantial user-generated content in the form of app store reviews, which serve as valuable feedback for developers [1], [2]. Honor of Kings, a popular multiplayer online battle arena (MOBA) game, has attracted thousands of Indonesian players, whose reviews offer rich insights into gameplay experiences, technical issues, and feature preferences. However, extracting actionable insights from these reviews poses significant challenges due to their unstructured nature, the informal linguistic patterns characteristic of Indonesian social media discourse, and the sheer volume, which renders manual analysis infeasible. Furthermore, while computational methods have been applied to analyze user feedback, integrating thematic discovery with sentiment evaluation within a single analytical framework remains an area that warrants further exploration, particularly in capturing how specific discussion themes influence player satisfaction.

Natural language processing (NLP) combined with unsupervised learning techniques offers a scalable approach to extract meaningful patterns from textual feedback [3]. Topic modeling has been widely utilized to uncover latent themes within reviews, while sentiment analysis provides quantifiable measures of user satisfaction, and clustering enables the grouping of similar reviews for targeted interpretation [4]. Prior research has successfully applied sentiment classification to Indonesian mobile applications [5], [6] and employed hybrid feature selection

methods for review categorization [7]. These studies have laid important groundwork in automated text analysis and demonstrated the feasibility of applying NLP techniques to user-generated content. Building on this foundation, there is an opportunity to further explore integrated methodologies that simultaneously examine topic distributions and sentiment patterns, particularly within the context of Indonesian gaming reviews. Additionally, adapting preprocessing techniques to better accommodate domain-specific terminology, informal expressions, and code-mixed language commonly found in gaming communities may enhance the interpretability and contextual relevance of analytical outcomes [8], [9].

To address these considerations, we propose an integrated NLP and clustering pipeline for topic discovery and sentiment mapping on Indonesian Honor of Kings reviews. To the best of the authors' knowledge, no prior study has simultaneously combined domain-adaptive preprocessing, unsupervised review clustering, and joint topic sentiment mapping with statistical validation within a single pipeline applied to Indonesian mobile gaming reviews. This study extends existing approaches through three methodological components: (1) implementing a domain-adaptive preprocessing framework that incorporates gaming-specific stopword removal and Indonesian stemming to better handle informal review language; (2) applying an unsupervised clustering approach that combines TF-IDF vectorization with mini-batch k-means to identify natural segmentation patterns in player feedback [10], validated through silhouette analysis; and (3) developing a joint topic-sentiment mapping methodology that associates emotional polarity with specific gameplay or technical themes using LDA topic modeling, supported by statistical validation through chi-square tests and Cramér' s V effect size.

The contributions of our study are threefold: (1) the development of a comprehensive preprocessing pipeline tailored for Indonesian gaming reviews that addresses linguistic characteristics unique to this domain; (2) empirical evidence demonstrating the applicability of unsupervised clustering for segmenting player feedback into interpretable groups without reliance on labeled data; and (3) the generation of actionable insights for game developers through sentiment-topic mapping, which highlights how specific features or technical aspects correlate with positive or negative player experiences.

2. Literature Review

2.1. App Review Mining for Software Evolution

The analysis of user-generated content from app stores has emerged as a critical practice in software engineering and product optimization [2]. Recent studies demonstrate that extracting sentiment and thematic patterns from user reviews can effectively inform development priorities and user experience enhancements [2], [6]. For instance, Pınarbasi [1] applied sentiment analysis to mobile game app reviews to evaluate emotional responses, highlighting how player feedback reflects broader satisfaction trends and feature reception. In the Indonesian context, Erlyana et al. [3] developed a scalable NLP-driven framework for multi-platform review analysis, demonstrating that automated topic-sentiment mapping can efficiently process large volumes of unstructured feedback to extract actionable insights. Furthermore, foundational surveys in text mining, such as Alghamdi and Alfalqi [4], provided a comprehensive taxonomy of topic modeling methods, including LSA, PLSA, LDA, and CTM, thereby establishing the theoretical groundwork underpinning the topic discovery component of this research.

2.2. Sentiment Analysis in Text Mining

Sentiment analysis provides a quantitative lens for interpreting subjective user feedback. While supervised classifiers [5], [11], [6], and transformer-based architectures [12] perform well, they require extensive labeled training data. Similarly, hybrid approaches improve polarity detection on informal text but often lack domain-specific adaptation [13]. Therefore, this study deliberately employs a transparent rule-based lexicon approach to bypass the labeled data bottleneck, ensure full interpretability, and allow immediate customization of gaming terminology.

While transformer-based models like IndoBERT excel in capturing contextual semantics [14], they function as 'black boxes' that obscure the rationale behind sentiment classifications. Furthermore, they require costly fine-tuning on large, domain-specific labeled datasets, which are often unavailable for niche gaming communities. Therefore, this study deliberately employs a transparent, rule-based lexicon approach to bypass the labeled-data bottleneck, ensure full interpretability for developers, and provide a robust exploratory baseline.

2.3. Topic Modeling and Latent Structure Discovery

Latent Dirichlet Allocation (LDA) has become a standard technique for uncovering thematic structures in large text corpora due to its probabilistic foundation and interpretability [4], [15]. Blei et al. [16] established the theoretical framework of LDA, demonstrating its effectiveness in discovering coherent topics across diverse domains. In text mining applications, Faisal et al. [17] successfully applied LDA to Indonesian news articles, showing that domain-adaptive preprocessing significantly improves topic coherence. More recently, hybrid approaches combining LDA with deep learning embeddings and semantic enhancement techniques have further enhanced topic quality [10], [18], [19], [12], [15], [20]. Nevertheless, topic modeling alone does not reveal how discovered themes correlate with user sentiment, a gap this research fills by jointly modeling topics and sentiment distributions [21].

While neural topic models like BERTopic [22] have gained popularity by leveraging transformer embeddings to capture deep semantics, they often obscure the explicit probabilistic relationships between terms and topics. This study deliberately retains Latent Dirichlet Allocation (LDA) because its generative framework provides mathematically transparent topic-term distributions. This transparency is essential for our joint topic-sentiment mapping, allowing researchers to directly trace how specific keywords drive thematic structures and facilitating rigorous statistical validation without relying on opaque, high-dimensional vector spaces.

2.4. Clustering Techniques for Text Segmentation

Unsupervised clustering enables the grouping of similar documents without labeled data, making it ideal for exploratory analysis of user reviews. K-Means and its variants remain widely adopted due to their simplicity and scalability [23]. For text data, TF-IDF vectorization combined with K-Means has proven effective in producing interpretable clusters, as demonstrated by Surianto [10] and Hidayat et al. [19] in their analysis of Indonesian journal articles and news clustering. Mini-batch k-means, in particular, offers computational efficiency for medium-to-large datasets while maintaining clustering quality [24]. Validation of clustering results is essential. The silhouette score provides a widely accepted metric for assessing cluster cohesion and separation [25]. The broader relevance of cohesion-separation measures for evaluating unsupervised segmentation tasks, including extensions to streaming data, is well documented [25]. In this study, the standard static silhouette score is used as the primary optimality criterion for K selection, while chi-square tests enable statistical evaluation of associations between cluster membership and categorical variables, such as sentiment labels [26].

Modern text clustering pipelines increasingly use dense embeddings from models such as Sentence-BERT (SBERT) to capture contextual nuances [27]. While SBERT excels in semantic grouping, it incurs significant computational overhead and produces dense clusters that are difficult to interpret using standard statistical association tests. Conversely, the TF-IDF and mini-batch k-means combination utilized in this study operates on a sparse, highly interpretable vector space. This sparsity aligns perfectly with our statistical validation approach, ensuring that the resulting thematic segments are computationally efficient to reproduce and statistically rigorous for mapping against categorical sentiment labels.

2.5. Indonesian Language Processing Challenges

Processing Indonesian text presents unique challenges due to its agglutinative morphology, informal orthography in social media, and frequent code-mixing with English. Stemming, or reducing words to their root forms, is particularly important for improving term

consistency in vector space models. A recent systematic literature review confirms that stemming and lemmatization remain essential preprocessing steps in Indonesian NLP, with studies consistently reporting improvements in text classification and information retrieval tasks [8]. For Indonesian digital text processing, rule-based stemming algorithms that handle complex prefix and suffix patterns have proven effective at normalizing morphological variations [9]. Additionally, handling domain-specific terminology (e.g., gaming slang like lag, bug, freeze) requires custom stopword lists and lexical resources beyond general-purpose dictionaries. Prior studies on Indonesian e-commerce reviews have demonstrated that domain-adaptive preprocessing can significantly improve downstream task performance in handling informal linguistic variations [5].

2.6. Research Gap and Positioning

While prior research has made significant contributions to app review analysis, sentiment classification, topic modeling, and text clustering [1], [2], [15], several gaps remain. First, most studies focus on either sentiment analysis or topic discovery in isolation [12], [20], limiting the ability to understand how specific themes drive positive or negative user experiences. Second, research on Indonesian-language gaming reviews remains scarce, despite the rapid growth of mobile gaming in the region. Third, few studies integrate statistical validation (e.g., chi-square, Cramér's V) to rigorously assess the association between clustering results and sentiment distributions.

Our study addresses these gaps by proposing an integrated pipeline that simultaneously performs topic discovery and sentiment mapping on Indonesian Honor of Kings reviews. By combining TF-IDF vectorization, mini-batch k-means clustering, LDA topic modeling, and lexicon-based sentiment analysis with validation through silhouette score and chi-square testing, this research provides a methodologically rigorous framework for extracting actionable insights from unstructured player feedback.

3. Methodology

3.1. Data Collection and Preprocessing

The dataset comprises 8,000 Indonesian-language reviews of Honor of Kings collected from the Google Play Store on May 21, 2026. The data were extracted using an automated Python-based web scraping script leveraging the `google_play_scraper` Python library with parameters configured for Indonesian language (`lang='id'`) and Indonesian market (`country='id'`). To ensure linguistic relevance, only reviews written in Indonesian were retained, filtered initially by the store's language metadata and further verified during the preprocessing stage. To elaborate on this verification, beyond the initial scraper metadata, we implemented a strict preprocessing pipeline that inherently acts as a secondary linguistic filter. Specifically, we removed all non-alphabetic characters (eliminating emoji-only or symbol-heavy gibberish) and applied Indonesian-specific NLP resources, namely the NLTK Indonesian stopword list and the Sastrawi rule-based stemmer designed for Indonesian agglutinative morphology. Reviews that could not be processed into meaningful Indonesian root forms or fell below the minimum length threshold (10 characters) after these transformations were systematically discarded. This multi-stage approach ensured that the final dataset retained only structurally valid Indonesian text, effectively filtering out predominantly English or non-linguistic noise, resulting in the final 7,975 valid reviews. Ethically, the dataset consists entirely of publicly available user-generated content; no Personally Identifiable Information (PII) such as usernames or email addresses was extracted, and the data collection process strictly complies with the Google Play Store's Terms of Service regarding the use of public reviews for academic research. The methodological pipeline is illustrated in Figure 1. Preprocessing was performed in sequential, modular steps: (1) text cleaning: converting to lowercase, removing URLs, mentions, hashtags, and non-alphabetic characters, and normalizing whitespace; (2) stopword filtering: removing Indonesian stopwords from NLTK corpus augmented with gaming-specific terms (e.g., game,

hero, skill); (3) stemming: applying Sastrawi stemmer [9] to reduce words to root forms; and (4) length filtering: retaining reviews with cleaned text longer than 10 characters to exclude noise.

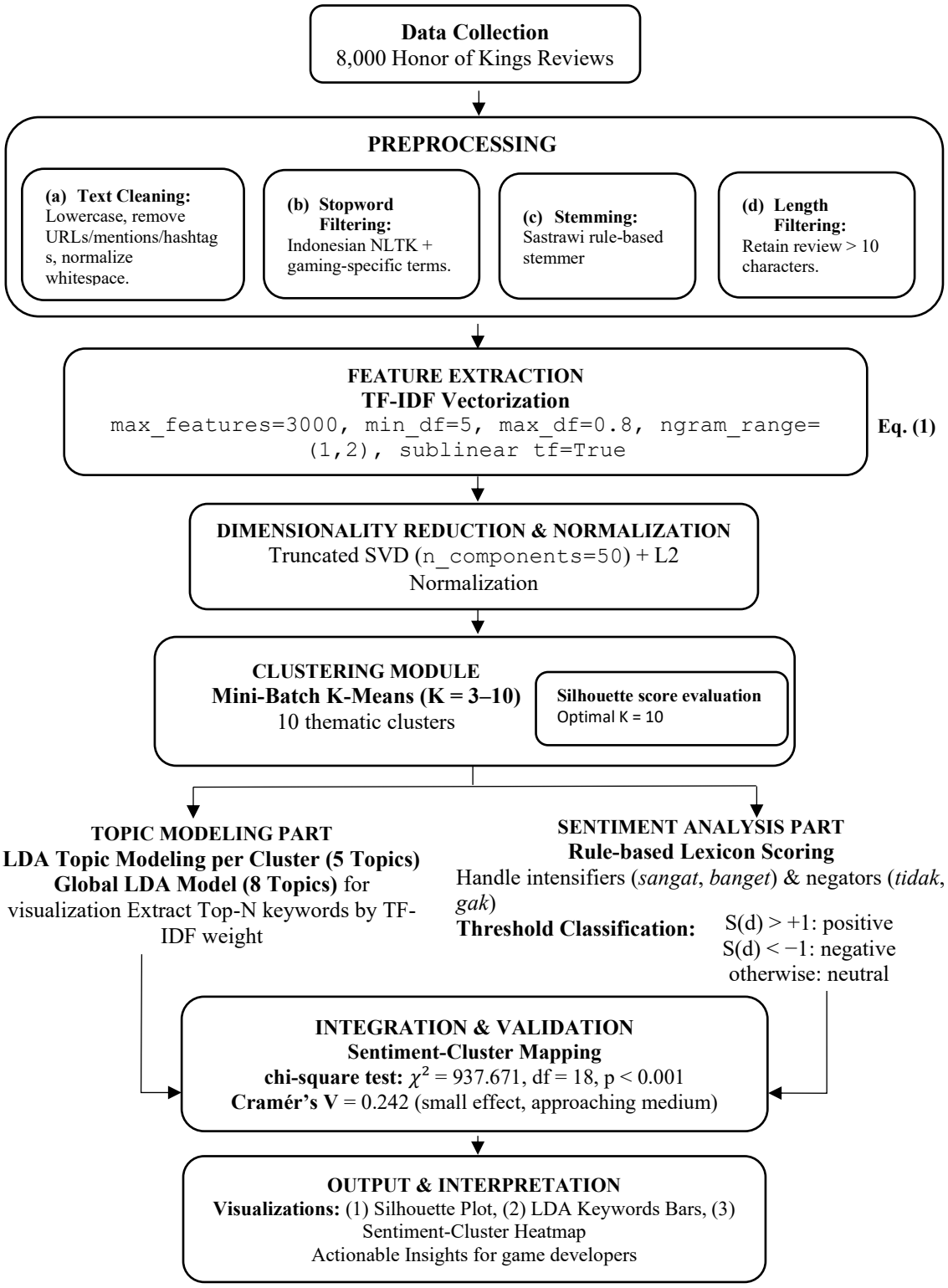


Figure 1. Methodological pipeline of the integrated NLP-clustering framework

3.2. Feature Extraction and Clustering

Cleaned reviews were transformed into numerical features using TF-IDF vectorization with parameters: `max_features=3000`, `min_df=5`, `max_df=0.8`, `ngram_range=(1,2)`, and `sublinear_tf=True` [10], [19], [21]. To mitigate the curse of dimensionality and reduce sparsity in the resulting 3,000-dimensional feature space, Truncated Singular Value Decomposition (SVD) was applied to project the TF-IDF matrix onto a 50-dimensional latent semantic space, followed by L2 normalization to align Euclidean distance computations with cosine similarity, consistent with standard practice for high-dimensional sparse text representations. Mini-batch k-means clustering was subsequently applied to this reduced and normalized feature space due to its efficiency on medium-sized datasets [24]. The optimal number of clusters (K) was determined by evaluating silhouette score (using cosine distance) across K = 3–10, selecting the value with the highest score [25].

$$\boldsymbol{TF - IDF(t,d)} \;=\; \boldsymbol{TF(t,d)\; x\; log}\left(\frac{\boldsymbol{N}}{\boldsymbol{DF(t)+1}}\right) \tag{1}$$

Where:

$TF(t,d)$: Sublinear term frequency: 1 + log (raw count) if counts > 0, else 0
(N) : Total number of documents in the corpus (N = 7,975)
$DF(t)$: Document frequency: number of documents containing term t
$+1$: Laplace smoothing to prevent division by zero for unseen terms

To transform preprocessed textual reviews into a numerical representation suitable for clustering, we employ Term Frequency–Inverse Document Frequency (TF-IDF) weighting as defined in Equation (1). This scheme quantifies the importance of a term t within a document d relative to the entire corpus, thereby down-weighting terms that appear frequently across many documents (e.g., function words or domain-general tokens) while amplifying discriminative terms that characterize specific topics [1], [3].

Following best practices for short-text clustering [4], we apply sublinear TF-scaling (i.e., TF (t,d) = 1 + log(raw count) for counts > 0) to mitigate the bias toward longer reviews, and we incorporate Laplace smoothing in the IDF denominator to ensure numerical stability for rare terms. The resulting sparse vector space captures both local (term-level) and global (corpus-level) statistical properties, providing a robust foundation for subsequent unsupervised learning stages.

$$\boldsymbol{J = \sum_{i=1}^{n}\sum_{k=1}^{K} r_{ik}\,||X_i - \mu_k||^2} \tag{2}$$

Where:

J : Objective function to be minimized
n : Total number of preprocessed reviews (n = 7,975)
K : Number of clusters (optimal K = 10)
r_{ik} : Assignment indicator: 1 if document i belongs to cluster k, 0 otherwise
X_i : SVD-reduced, L2-normalized feature vector of review i (50-dimensional)
μ_k : Centroid vector of cluster k

Mini-batch k-means clustering was applied to partition the SVD-reduced, normalized vector space into interpretable thematic groups by minimizing the within-cluster sum of squares as formulated in Equation (2) [24]. This objective function quantifies cluster cohesion through the squared Euclidean distance between each review vector and its assigned centroid.

Unlike standard k-means, the mini-batch variant updates centroids using small random subsets (mini-batches) of the data at each iteration, reducing computational complexity from $\boldsymbol{O}\,(\boldsymbol{n \times K \times I \times d})$ to $\boldsymbol{O}\,(\boldsymbol{b \times K \times I \times d})$, where b ≪ n denotes batch size and d is feature dimensionality, here corresponding to the 50-dimensional SVD-reduced space rather than the original 3,000-dimensional TF-IDF space. This efficiency is particularly advantageous for medium-scale text corpora such as the 7,975 Honor of Kings reviews analyzed herein, while

empirical studies confirm that clustering quality remains comparable to the full-batch algorithm [24]. The procedure converges when centroid displacement falls below a tolerance threshold or the maximum iteration limit is reached, yielding a partition that locally optimizes Equation (2).

3.3. Topic Modeling and Sentiment Analysis

For each identified cluster, LDA topic modeling was applied using five topics to explore cluster-specific thematic structures and identify dominant discussion patterns [21]. To provide a broader representation of thematic distributions across the entire dataset, an additional global LDA model consisting of eight topics was constructed for visualization purposes using 15 passes and `alpha='auto'`. The most representative keywords from each topic were then extracted to support thematic interpretation. Sentiment analysis employed a rule-based lexicon approach using a custom curated word set comprising 30 domain-relevant positive terms (e.g., *bagus, keren*, *mantap*) and 35 negative terms (e.g., *jelek,* lag, crash), with explicit handling of intensifiers (e.g., *sangat, banget*) and negators (e.g., *tidak, gak*) to improve polarity detection accuracy. To clarify the composition of this 65-term sentiment lexicon, we followed a domain-adaptive compilation approach. Regarding the source, we established a baseline using common Indonesian sentiment words and enriched them with gaming-specific slang (e.g., lag, bug*, mantul, jos*) extracted directly from the 8,000 Honor of Kings reviews. Regarding the filtering, we manually removed generic words that lacked distinct polarity to ensure each term carried clear emotional weight. Regarding the validation, both authors collaboratively reviewed the final list to confirm domain specificity and contextual relevance. While we acknowledge that a 65-word vocabulary is limited for highly informal text, this focused lexicon was intentionally compiled to prioritize high precision and interpretability for our exploratory analysis. The complete list of these 65 terms is provided in Tables A1–A4 (Appendix A). Each review was subsequently categorized into positive, negative, or neutral sentiment classes. Because the five sub-topics generated per cluster often contained overlapping or semantically related terms–particularly high-frequency generic words such as *nya*, *tolong*, and *bagus* that appeared across most clusters–representative keywords presented in Table 2 and discussed in Section 4.2 were manually curated across sub-topics to highlight terms most distinctive to each cluster's thematic identity, rather than relying solely on the highest-probability terms from a single topic.

This manual curation involved a systematic three-step process: (1) extracting the top 20 highest-probability terms from the five sub-topics generated per cluster; (2) filtering out generic, high-frequency function words, or overlapping slang (e.g., *nya, banget*) that appeared across multiple clusters without distinct thematic value; and (3) selecting the 5–7 most semantically distinctive terms that best capture the unique thematic identity of each cluster. This curation was performed by the first author and validated through discussion with the second author to ensure domain relevance and interpretability, prioritizing terms that are both statistically significant and semantically meaningful, thereby serving as a robust human interpretability evaluation of the topic quality to ensure the extracted themes are contextually relevant to the Indonesian gaming domain.

$$\mathbf{S(d)} = \sum_{\mathbf{w \in d}} [\boldsymbol{polarity(w)} \times \boldsymbol{intensifier(w)} \times \boldsymbol{negator(w)}] \quad (3)$$

$$\boldsymbol{sentiment}\ (\boldsymbol{d}) = \begin{cases} \boldsymbol{positive, if\ S(d) > T_{pos}} \\ \boldsymbol{negative, if\ S(d) < T_{neg}} \\ \boldsymbol{neutral, otherwise} \end{cases}$$

Where:

S(d)	: Aggregate sentiment score for document/review d
w	: Individual word token in review d
polarity(w)	: Base polarity: +1 if w ∈ positive lexicon, −1 if w ∈ negative lexicon, 0 otherwise
intensifier(w)	: Multiplier: 2.0 if preceding word is intensifier (e.g., *sangat*), else 1.0
negator(w)	: Sign flip: −1 if preceding word is negator (e.g., *tidak*), else +1
T_{pos}, T_{neg}	: are the classification thresholds, where T_{pos} = +1 and $T_{neg} = -1$

Sentiment classification employs a transparent rule-based lexicon approach as formulated in Equation (3). Each review d receives an aggregate score S(d) by summing the polarity contributions of constituent words, adjusted for contextual modifiers. Intensifiers (e.g., *sangat, banget*) amplify polarity magnitude by a factor of 2.0, while negators (e.g., *tidak, gak*) invert the polarity sign of the subsequent term [5].

The final sentiment label is assigned by comparing the positive and negative score components: a review is classified as positive if the positive score exceeds the negative score by more than one unit (i.e., $S(d) > +1$), negative if the negative score exceeds the positive score by more than one unit (i.e., $S(d) < -1$), and neutral otherwise. This margin-based thresholding reduces sensitivity to noise from isolated sentiment words and reflects the score-difference mechanism implemented in the pipeline. The design balances interpretability with reasonable accuracy for informal Indonesian text, avoiding the computational overhead and data requirements of supervised machine learning models while maintaining full transparency of the scoring mechanism.

3.4. Manual Evaluation of Lexicon-Based Sentiment

To address the potential limitations of a relatively small domain-specific lexicon (65 terms) and to evaluate its reliability, a manual evaluation was conducted. To ensure methodological rigor, the manual annotation was performed exclusively by the two authors of this study, who are native Indonesian speakers with expertise in NLP and familiarity with gaming discourse. Each author independently labeled a random sample of 100 reviews using a detailed coding guideline that defined the three sentiment categories (positive, neutral, negative) along with example phrases extracted from the dataset. Prior to the full annotation, a calibration session was conducted on 10 sample reviews to align our interpretation of the coding scheme. Inter-annotator agreement was measured using Cohen's Kappa, yielding a score of 0.68, which indicates substantial agreement. Disagreements were resolved through consensus discussion between the two authors to establish the final ground-truth labels. The predictions from the lexicon-based pipeline were then compared against these manual labels to generate a confusion matrix and standard classification metrics, as presented in Table 1. We explicitly acknowledge that having only two annotators (the authors themselves) constitutes a limitation regarding scalability and potential researcher bias, and we recommend that future work involve a larger, independent pool of annotators.

Table 1. Confusion Matrix of Lexicon-Based Sentiment vs. Manual Annotation (N = 100)

Actual (Manual)	Predicted Lexicon		
	Negative	**Neutral**	**Positive**
Negative	6	38	1
Neutral	1	18	4
Positive	1	16	15
Total	**8**	**72**	**20**

The evaluation yielded an overall accuracy of 39%, with a macro-averaged precision of 0.58, recall of 0.46, and F1-score of 0.39. The weighted F1-score was 0.37. Notably, the negative class exhibited a low recall (0.13), indicating that the small lexicon struggles to capture implicit or slang-based negative expressions in informal Indonesian, resulting in a high proportion of reviews being classified as neutral (70.3% in the full dataset). While the precision for positive and negative classes remains reasonably high (0.75), the overall metrics validate the transparency of the rule-based approach while highlighting the need for larger, domain-specific lexical resources in future work to improve recall.

Consequently, the sentiment labels generated by this pipeline should be interpreted primarily as exploratory findings that indicate macro-level polarity trends, rather than definitive ground-truth classifications for individual reviews.

3.5. Validation and Visualization

Cluster validity was assessed using the silhouette score to determine the optimal number of clusters. Association between clusters and sentiment distribution was tested using a chi-square test with Cramér's V for effect size. Visualizations included: (1) silhouette validation plot, (2) LDA topic keyword bar charts, (3) sentiment-cluster heatmap.

$$\boldsymbol{s(i)} \;=\; \boldsymbol{(b(i) \;-\; a(i)) \,/\, max\{a(i), b(i)\}} \tag{4}$$

Where:
s(i) : Silhouette coefficient for document i, bounded in [−1, 1]
a(i) : Mean intra-cluster distance (cohesion): average distance between document i and all other documents in the same cluster
b(i) : Mean nearest-cluster distance (separation): average distance between document i and all documents in the closest different cluster
max{a(i), b(i)} : Normalization factor ensuring the coefficient remains scale-invariant.

To evaluate the internal validity of the clustering partition, we employ the silhouette score as formulated in Equation (4). This metric assesses how appropriately each review is assigned to its cluster by comparing the average distance to other members within the same cluster (cohesion) against the average distance to members of the nearest neighboring cluster (separation). The global silhouette value is computed as the arithmetic mean across all documents, where values approaching +1 indicate well-separated and compact clusters, while scores near zero suggest overlapping or ambiguous groupings [25].

In high-dimensional sparse TF-IDF spaces, particularly for short informal texts, silhouette scores below 0.05 are frequently reported due to lexical variability, domain-specific slang, and inherent semantic overlap; nevertheless, the metric remains a robust internal criterion for selecting the optimal K when evaluated comparatively across candidate partitions [25], [26].
The resulting score of 0.1071 at K = 10 therefore reflects a weak but structurally meaningful segmentation consistent with prior studies on short-text clustering [10], [23].

In addition to the silhouette score, clustering quality was evaluated using the Davies-Bouldin Index (DBI) and the Calinski-Harabasz Index (CHI) for a more comprehensive assessment. DBI measures the average similarity ratio between each cluster and its most similar cluster, where lower values (closer to zero) indicate better cluster separation and less overlap. CHI evaluates the ratio of between-cluster to within-cluster dispersion, where higher values indicate denser, better-separated clusters. These complementary metrics help validate the clustering structure from different perspectives, which is particularly important when dealing with high-dimensional, sparse text data. The Elbow Method was also employed to visualize the within-cluster sum of squares (inertia) across different K values, providing complementary evidence for optimal cluster selection [28].

$$\boldsymbol{\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})2}{E_{ij}}} \tag{5}$$

Where:
χ^2 : chi-square test statistic.
O_{ij} : Observed frequency in row i (cluster) and column j (sentiment class).
E_{ij} : Expected frequency under independence = $(\frac{row\ total_i\ x\ column\ total_j}{grand\ total})$.
r : Number of rows (clusters, r = 10).
c : Number of columns (sentiment classes, c = 3).
The degrees of freedom are calculated as: **df = (r – 1) × (c – 1) = (10 – 1) × (3 – 1) = 18**

To assess the statistical association between cluster membership and sentiment distribution, we employ the chi-square test of independence as formulated in Equation (5) [26]. This test evaluates the null hypothesis that cluster assignment and sentiment polarity are independent variables. The test statistic χ^2 quantifies the discrepancy between observed frequencies O_{ij} and expected frequencies E_{ij} under the assumption of independence, where E_{ij} is computed as the product of the corresponding row and column marginal totals divided by the grand total.

In this study, the chi-square test was applied to a 10 × 3 contingency table (ten clusters × three sentiment classes: positive, neutral, negative), yielding df = (10 − 1) × (3 − 1) = 18 degrees of freedom. A significant result ($p < 0.05$) indicates that the distribution of sentiment labels varies systematically across clusters, thereby supporting the interpretability of the thematic segmentation and validating that the discovered topics carry distinct emotional profiles.

$$\boldsymbol{V = \sqrt{\frac{\chi^2}{N \times min(r-1, c-1)}}} \tag{6}$$

Where:
V : Cramér's V effect size, bounded in [0, 1]
χ^2 : chi-square statistic from Equation (5)
N : Total number of observations (N = 7,975 reviews)
r : Number of rows (clusters, r = 10)
c : Number of columns (sentiment classes, c = 3)
min (r − 1, c − 1): Smaller of (r − 1) or (c − 1); here = min (9, 2) = 2

While the chi-square test indicates whether an association between cluster membership and sentiment distribution is statistically significant, it does not convey the practical strength of that association. Therefore, we report Cramér's V as defined in Equation (6), a standardized effect size measure that adjusts for both sample size and table dimensions [26]. Values of V range from 0 (no association) to 1 (perfect association), with conventional thresholds of 0.10, 0.30, and 0.50 indicating small, medium, and large effects, respectively [29]. This metric enables meaningful comparison of association strength across studies with different sample sizes or categorical structures, and supports a more nuanced interpretation of the clustering–sentiment relationship beyond binary significance testing.

3.6. Implementation Details and Reproducibility

To ensure experimental reproducibility, all analyses were implemented in Python 3 using scikit-learn (v1.x) for vectorization and clustering, Gensim (v4.x) for topic modeling, and Sastrawi (v1.x) for Indonesian stemming. Crucial hyperparameters were strictly controlled to guarantee consistent results. TF-IDF vectorization used `max_features=3000`, `min_df=5`, `max_df=0.8`, `ngram_range=(1,2)`, and `sublinear_tf=True`. Dimensionality reduction was performed using Truncated SVD with `n_components=50`. Mini-batch k-means clustering was executed with `batch_size=1000`, `n_init=3`, `max_iter=100`, `init='k-means++'`, and a fixed `random_state=42` to ensure consistent centroid initialization and convergence across runs. For topic modeling, LDA was configured with `alpha='auto'`, `passes=15`, and a fixed `random_state=42`, utilizing `num_topics=5` for each cluster-specific model and `num_topics=8` for the global visualization model. Furthermore, silhouette scores were calculated using `metric='cosine'` and `sample_size=1000` with `random_state=42`. The preprocessing pipeline filtered reviews with a minimum length of 10 characters, retaining 7,975 out of 8,000 original reviews. For manual evaluation, a random sample of 100 reviews was annotated using `random_state=42`. All custom resources, including the 65-term sentiment lexicon and gaming-specific stopword list, are available in the supplementary materials.

4. Results and Discussion

4.1. Clustering Results and Validation

The pipeline processed 8,000 reviews, retaining 7,975 after preprocessing. Mini-batch k-means clustering identified 10 optimal clusters based on the highest silhouette score of 0.1071 at K = 10 (Figure 2). Complementary validation metrics yielded a Davies-Bouldin Index (DBI) of 3.4102 and a Calinski-Harabasz Index (CHI) of 191.30. Furthermore, the Elbow Method visualization (Figure 2a) shows a gradual decrease in inertia from K = 3 (6,390.82) to K = 10 (5,643.65), with no distinct elbow point, which is typical for text clustering scenarios [28]. While the silhouette score remains modest at 0.1071, the markedly lower DBI and higher CHI values evaluated in the same latent semantic space used for clustering indicate that the dimensionality reduction step effectively denoised the sparse lexical representation, yielding more compact and better-separated clusters than would be expected from the silhouette score alone. The combination of these metrics supports the selection of K = 10 as the optimal configuration, balancing cluster cohesion with interpretability. Cluster distribution was uneven, with the largest segments being Cluster 9 (15.5%), Cluster 1 (14.7%), and Cluster 7 (11.4%), while the smallest was Cluster 2 (2.8%). This indicates an emergent, soft segmentation with overlapping boundaries of review content, as detailed in Table 2.

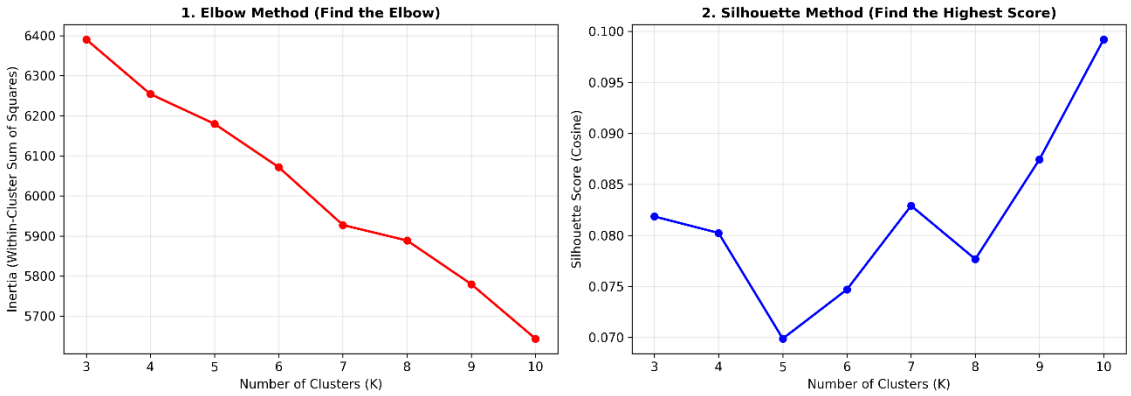


Figure 2. Cluster validation using (a) Elbow Method showing within-cluster sum of squares (inertia) and (b) silhouette score analysis for determining optimal number of clusters (K = 3 – 10). The highest silhouette score (0.1071) was achieved at K = 10, indicating the optimal cluster configuration. While the silhouette score appears modest, this is consistent with prior studies on short-text clustering in high-dimensional sparse TF-IDF spaces

Table 2. Cluster summary showing size, sentiment distribution, and primary topics for Indonesian Honor of Kings reviews (N = 7,975).

Cluster	Number of Reviews	Percentage	Dominant Sentiment	Positive %	Negative %	Neutral %	Primary Topics
0	950	11.9%	Positive	52.4%	0.6%	46.9%	*suka, hok, banget, kalo, udah*
1	1,171	14.7%	Neutral	18.6%	7.0%	74.4%	*gak, nya, bagus, udah, aja*
2	223	2.8%	Neutral	19.7%	7.6%	72.6%	*nya, bagus,* legends, *main*
3	632	7.9%	Neutral	38.0%	2.7%	59.3%	*bagus, nya,* event, *kasih*
4	848	10.6%	Neutral	15.0%	9.7%	75.4%	*nya, bintang, hok, tolong*
5	516	6.5%	Neutral	11.2%	10.5%	78.3%	*nya, gak, tolong, jalan, pas*
6	771	9.7%	Neutral	9.1%	8.3%	82.6%	*bintang, solo, bot, kasih*

Cluster	Number of Reviews	Percentage	Dominant Sentiment	Positive %	Negative %	Neutral %	Primary Topics
7	910	11.4%	Neutral	16.2%	7.6%	76.3%	*nya, musuh, main, imbang*
8	715	9.0%	Neutral	29.5%	6.2%	64.3%	bug, update, *nya, baik, fps*
9	1,239	15.5%	Neutral	12.3%	13.4%	74.3%	*jaring, nya, bagus, tolong*

Note: Table 2 and Table 3 use different numbering schemes. Table 2 presents the final cluster labels, numbered 0 – 9, whereas Table 3 presents the candidate numbers of clusters evaluated, ranging from K = 3 to K = 10.

Table 3. Cluster validation metrics across different numbers of clusters (K = 3 – 10)

Number of Clusters (K)	Inertia	silhouette score	Davies-Bouldin Index (DBI)	Calinski-Harabasz Index (CHI)
3	6,390.82	0.0818	4.4569	298.26
4	6,254.34	0.0806	4.6193	257.44
5	6,179.94	0.0695	4.1979	219.66
6	6,071.76	0.0738	4.1729	208.04
7	5,927.34	0.0846	3.9398	210.57
8	5,888.65	0.0766	3.7159	188.36
9	5,779.70	0.0919	3.4564	187.21
10	5,643.65	0.1071	3.4102	191.30

To comprehensively justify the selection of K = 10, Table 3 presents a comparison of four complementary clustering validation metrics across all candidate values. While the Calinski-Harabasz Index (CHI) peaks at lower K values–a common artifact in high-dimensional sparse text data where fewer clusters artificially inflate between-cluster dispersion–K = 10 yields the highest silhouette score (0.1071) and the lowest Davies-Bouldin Index (3.4102). This convergence indicates that K = 10 provides the optimal balance of intra-cluster cohesion and inter-cluster separation for this specific dataset, making it the most interpretable and robust configuration for our thematic segmentation.

4.2. Topic Discovery per Cluster

LDA topic modeling revealed distinct themes per cluster. Cluster 0 topics centered on general satisfaction and positive gameplay experience (*suka, banget, kalo, gamenya*), while Cluster 1 reflected broad gameplay impressions and informal feedback (*gak, bagus, aja, seru*). Cluster 2 captured general content discussion with a smaller sample (*bagus,* legends*, main, item*), Cluster 3 emphasized in-game events and visual effects (even*t, efek, segi, bagus*), and Cluster 4 addressed rating and system-related feedback (*bintang, sistem, tolong, buruk*). Additional clusters captured bug and technical-issue reports (Cluster 5: *pas, tolong, jalan,* bug), matchmaking and team-system concerns (Cluster 6: *solo, sistem,* match*,* team), enemy balance and matchmaking fairness issues (Cluster 7: *musuh, imbang,* matchmaking*, kalah*), update and graphics/FPS performance feedback (Cluster 8: update*, grafis, fps,* bug), and network and connectivity complaints (Cluster 9: *jaring, sinyal, wifi, ngelag*). These themes were extracted using LDA topic modeling [15], [16], with keyword probability distributions visualized in Figure 3 to support manual interpretation and topic coherence assessment.

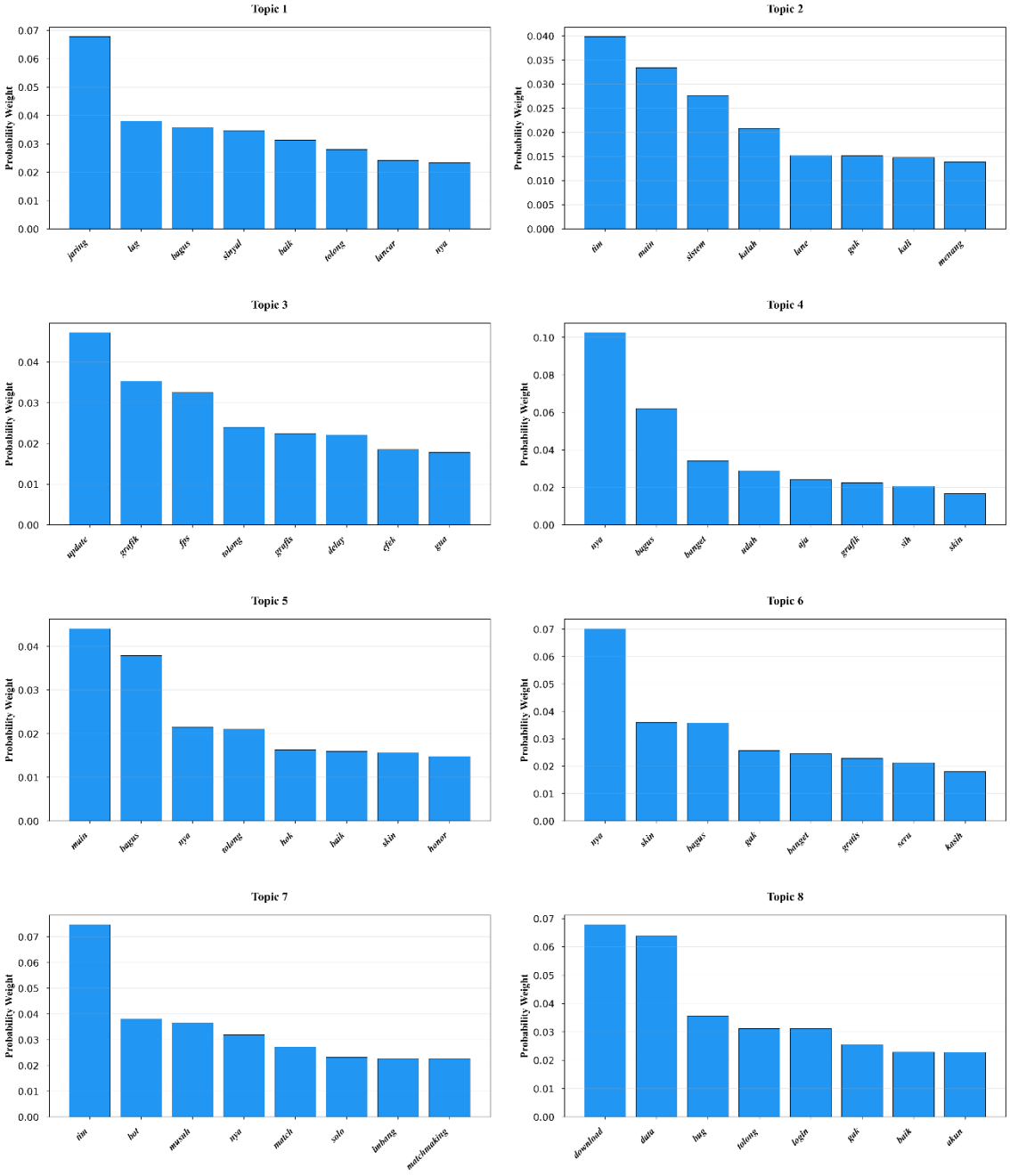


Figure 3. Top keywords per topic extracted using Latent Dirichlet Allocation (LDA) topic modeling

The visualization presents the most representative keywords across eight LDA topics extracted from Indonesian Honor of Kings reviews, revealing dominant themes that range from gameplay mechanics and matchmaking (e.g., match*, solo, kalah*) to account issues and technical performance (e.g., login*, akun, fps*).

4.3. Sentiment Mapping and Statistical Association

Sentiment analysis labeled 70.3% of reviews as neutral, 22.1% as positive, and 7.5% as negative. The sentiment distribution across clusters is visualized in Figure 4. Distribution varied significantly across clusters (chi-square test: $\chi^2 = 937.671$, df = 18, $p < 0.001$), with Cramér' s V = 0.242 indicating a small-to-medium effect size. Given the aforementioned limitations in lexicon-based recall, these statistical associations and the resulting sentiment-cluster mapping

should be interpreted as exploratory findings that highlight broad thematic trends, rather than definitive micro-level classifications. Cluster 0 showed the highest positive sentiment (52.4%), consistent with its focus on general gameplay enjoyment and visual appreciation. Conversely, negative sentiment remained relatively low across all clusters (ranging from 0.6% to 13.4%), with Cluster 9 (network/connection issues) showing the highest negative proportion, followed by Cluster 5 (download and functionality issues) at 10.5% and Cluster 4 (system/rating feedback) at 9.7%. The high prevalence of neutral sentiment suggests that many reviews contain informational or mixed-experience feedback rather than strongly polarized opinions.

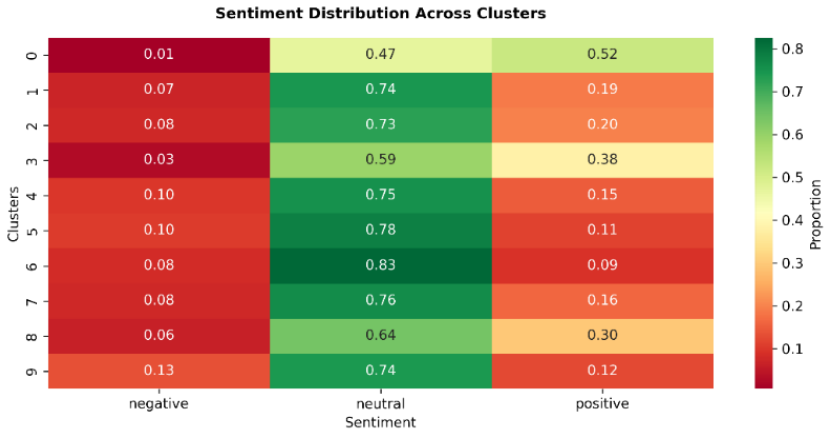


Figure 4. Heatmap showing sentiment distribution across ten clusters

The color intensity represents the proportion of negative (red), neutral (green), and positive (orange) sentiments within each cluster. The visualization reveals that neutral sentiment dominates across all clusters (46.9 – 82.6%), while positive sentiment is concentrated in Clusters 0, 3, and 8 (visual features and updates), and negative sentiment remains consistently low (< 14%) across all thematic groups.

4.4. Integrated Interpretation

The thematic segmentation reveals that Indonesian players' dissatisfaction is predominantly rooted in technical infrastructure and system mechanics rather than core gameplay design. Specifically, the clusters associated with network and connectivity issues (Cluster 9) and download and functionality issues (Cluster 5) have the highest proportions of negative sentiment (13.4% and 10.5%, respectively), closely followed by Cluster 4 (system/rating feedback, 9.7%). These clusters reflect systemic friction points, such as network instability, incomplete downloads and bugs, and rating system frustrations, that compound user dissatisfaction. This pattern aligns with prior findings in emerging mobile gaming markets, where server latency and device compatibility frequently outweigh content quality as primary drivers of churn [6], [20]. Conversely, Cluster 0, which centers on general gameplay satisfaction, exhibits the highest positive sentiment (52.4%), indicating that when technical barriers are minimized, the core gameplay experience resonates strongly with the local player base.

Furthermore, the strong positive reception of cosmetic features and reward events (Clusters 3 and 8) suggests that visual customization and free-to-play incentive structures resonate effectively with the local player base. For developers, this implies a clear prioritization strategy: stabilizing backend infrastructure and optimizing client-side performance should precede new content releases to maximize the impact on retention.

Finally, the high prevalence of neutral sentiment across most clusters (47% to 83%) likely reflects the descriptive and informational nature of Indonesian app reviews, which often blend feature requests, bug reports, and mild commentary rather than expressing strong emotional polarity. This characteristic is further corroborated by our manual evaluation (Table 1), in which the lexicon's low recall (0.13 for negative sentiments) led to many nuanced or slang-heavy reviews

being classified as neutral. This limitation should be considered when interpreting lexicon-based sentiment results, as the relatively small lexicon (65 terms) may underrepresent context-dependent slang and mixed-sentiment expressions.

4.5. Limitations and Risks

While the integrated pipeline provides actionable insights, several limitations and interpretive risks must be acknowledged. First, while the Davies-Bouldin Index (3.4102) and Calinski-Harabasz Index (191.30) indicate reasonably well-separated clusters in the SVD-reduced feature space, the modest silhouette score (0.1071) suggests that individual reviews still exhibit considerable lexical overlap at the boundary regions between thematic categories. This is characteristic of short, informal text data, where reviews frequently contain mixed content spanning multiple themes. Consequently, the thematic segmentation should be interpreted as broad, overlapping thematic tendencies rather than distinct, rigid, or mutually exclusive categories. This poses a risk of over-interpretation, as individual reviews may inherently contain keywords spanning multiple thematic areas (e.g., complaining about network lag while praising visual skins).

Second, the lexicon-based sentiment analysis relies on a relatively small custom dictionary (65 terms). As evidenced by the manual evaluation (Table 1), this introduces a recall bias, particularly for negative sentiments (0.13), as the lexicon struggles to capture implicit, context-dependent, or slang-based expressions common in informal Indonesian gaming discourse. Consequently, this results in a high prevalence of neutral classifications (70.3%). Therefore, the resulting sentiment mapping, chi-square validation, and subsequent implications for developers must be interpreted strictly as exploratory findings rather than definitive classifications.

Third, the dataset itself is subject to selection bias inherent to app store reviews. The feedback primarily represents a subset of users who are highly motivated to leave reviews, potentially overrepresenting extreme technical frustrations or highly specific feature requests, while underrepresenting the silent majority of moderately satisfied users. Finally, the domain-specific nature of gaming terminology and the informal, code-mixed nature of Indonesian social media language may limit the direct generalizability of these findings to more formal text corpora or different application domains.

5. Conclusion and Suggestions

In this study, we successfully addressed the research gap of integrating topic discovery and sentiment mapping within a single, statistically validated pipeline for Indonesian-language gaming reviews, demonstrating that domain-adaptive preprocessing combined with unsupervised clustering and lexicon-based sentiment analysis can reveal interpretable, actionable patterns from informal user feedback without relying on labeled training data. Beyond confirming thematic segmentation, our integrated interpretation revealed that player dissatisfaction in the Indonesian Honor of Kings community is driven primarily by technical and infrastructural concerns rather than core gameplay design, offering developers an exploratory basis for prioritizing backend stability and matchmaking improvements ahead of new content releases. However, this study acknowledges two main limitations: the inherent weak separation of clusters in short-text data, and the low accuracy of lexicon-based sentiment analysis in capturing informal slang. The proposed approach further demonstrates practical value as a scalable, low-cost framework that other Indonesian digital product teams can adapt for systematic review analysis. We see future development of this work extending toward hybrid or deep learning-based sentiment models to improve recall for negative expressions [20], temporal analysis to monitor sentiment shifts across game updates, and validation through direct developer interviews to bridge computational findings with practical implementation decisions.

6. Acknowledgments

The authors would like to thank the anonymous reviewers and the Faculty of Computer Science, Citra Buana Indonesia Institute, for supporting this research.

7. Appendix

Appendix A. Custom Domain-Specific Sentiment Lexicon.

The complete list of the 65 domain-relevant terms used for the rule-based sentiment analysis is detailed in Tables A1–A4.

Table A1. Positive Sentiment Terms (30 words)

Term
bagus, mantap, hebat, senang, puas, suka, recommended*, menarik, seru, lucu, epic, pro, jago,* smooth*, cepat, ringan,* update, worth, best, favorite, happy, awesome*, grafis, keren, mantul, jos, oke,* good, nice, perfect

Table A2. Negative Sentiment Terms (35 words)

Term
jelek, buruk, kecewa, marah, benci, hapus, uninstall, lag, bug, error*, lemot, berat, panas, boros, susah, ribet,* cheater*,* hacker*, ganggu, aneh,* glitch, crash, freeze, slow, bad, worst, trash*,* waste*,* boring, annoying, frustrating, pay*, mahal,* unfair, overpowered

Table A3. Intensifiers (8 words)

Term	Function
sangat	Multiplier: 2.0×
banget	Multiplier: 2.0×
sekali	Multiplier: 2.0×
amat	Multiplier: 2.0×
terlalu	Multiplier: 2.0×
super	Multiplier: 2.0×
extremely	Multiplier: 2.0×
very	Multiplier: 2.0×

Table A4. Negators (7 words)

Term	Function
tidak	Sign flip: –1×
nggak	Sign flip: –1×
gak	Sign flip: –1×
bukan	Sign flip: –1×
no	Sign flip: –1×
not	Sign flip: –1×
never	Sign flip: –1×

References

[1] F. Pınarbasi, "Evaluation of Mobile Game Apps Emotional Side: A Sentiment Analysis on Mobile Game App Reviews," *Yönetim Bilimleri Dergisi*, vol. 23, no. 58, pp. 1953–1978, Oct. 2025, doi: 10.35408/comuybd.1564966.

[2] J. Dąbrowski, E. Letier, A. Perini, and A. Susi, "Analysing app reviews for software engineering: a systematic literature review," *Empirical Software Engineering*, vol. 27, no. 2, p. 43, Mar. 2022, doi: 10.1007/s10664-021-10065-7.

[3] Y. Erlyana, H. Hartono, and J. Y. Lim, "A Scalable NLP-Driven Information System for Multi-Platform Sentiment and Topic Analysis of Indonesian Coffee Brand Reviews," *Informatica*, vol. 31, no. 2, pp. 529–538, Feb. 2026, doi: 10.18280/isi.310221.

[4] R. Alghamdi and K. Alfalqi, "A Survey of Topic Modeling in Text Mining," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 1, pp. 134–143, 2015, doi: 10.14569/IJACSA.2015.060121.

[5] V. Fitriyana, Lutfi Hakim, Dian Candra Rini Novitasari, and Ahmad Hanif Asyhar, "Analisis Sentimen Ulasan Aplikasi Jamsostek Mobile Menggunakan Metode Support Vector Machine," *Jurnal Buana Informatika*, vol. 14, no. 01, pp. 40–49, Apr. 2023, doi: 10.24002/jbi.v14i01.6909.

[6] R. Purnamasari and A. R. Thaha, "Understanding user satisfaction in digital banking: Sentiment analysis and topic modeling of Indonesian bank reviews," *Business Informatics*, vol. 19, no. 3, pp. 101–112, 2025, doi: 10.17323/2587-814X.2025.3.101.112.

[7] S. S. Abdulkhaliq and A. M. Darwesh, "Sentiment Analysis Using Hybrid Feature Selection Techniques," *University of Human Development Journal of Science and Technology*, vol. 4, no. 1, pp. 29–40, Feb. 2020, doi: 10.21928/uhdjst.v4n1y2020.pp29-40.

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

[9] P. Prihatini, I. D. Putra, I. Giriantari, and M. Sudarma, "Stemming Algorithm for Indonesian Digital News Text Processing," *International Journal of Engineering and Emerging Technology*, vol. 2, no. 2, Jul. 2017, [Online]. Available: https://garuda.kemdiktisaintek.go.id/documents/detail/1097283

[10] Dewi Fatmarani Surianto and Dewi Fatmawati Surianto, "Enhancing K-Means Clustering for Journal Articles using TF-IDF and LDA Feature Extraction," *Brilliance: Journal of Research and Science Education*, vol. 4, no. 2, pp. 964–972, Mar. 2025, doi: 10.47709/brilliance.v4i2.5547.

[11] W. L. Azzahra, J. Indra, R. Rahmat, and S. Faisal, "Sentiment Analysis of User Reviews of the AdaKami Online Loan App from the App Store Using SVM and Naive Bayes," *Journal of Applied Informatics and Computing*, vol. 9, no. 3, pp. 838–850, Jun. 2025, doi: 10.30871/jaic.v9i3.9536.

[12] V. H. Pranatawijaya, N. N. K. Sari, R. A. Rahman, E. Christian, and S. Geges, "Unveiling User Sentiment: Aspect-Based Analysis and Topic Modeling of Ride-Hailing and Google Play App Reviews," *Journal of Information Systems Engineering and Business Intelligence*, vol. 10, no. 3, pp. 328–339, Oct. 2024, doi: 10.20473/jisebi.10.3.328-339.

[13] C. G. Maurya and S. K. Jha, "Sentiment Analysis: A Hybrid Approach on Twitter Data," *Procedia Computer Science*, vol. 235, pp. 990–999, 2024, doi: 10.1016/j.procs.2024.04.094.

[14] A. A. P. Simarmata and T. B. Sasongko, "Sentiment Analysis on BRImo Application Reviews Using IndoBERT," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 3, pp. 851–862, Jun. 2025, doi: 10.30871/jaic.v9i3.8162.

[15] I. Z. Mustaqim and R. R. Suryono, "A Systematic Literature Review of Topic Modeling Techniques in User Reviews," *Journal of Information Systems Engineering and Business Intelligence*, vol. 11, no. 2, pp. 238–253, Jul. 2025, doi: 10.20473/jisebi.11.2.238-253.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[17] M. Faisal *et al.*, "Enhancing Indonesian Text Summarization with Latent Dirichlet Allocation and Maximum Marginal Relevance," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 8, pp. 652–661, 2024, doi: 10.14569/IJACSA.2024.0150852.

[18] P. Selvaprasanth, "Unsupervised Topic Discovery and Text Extraction with Generative AI and Embedding Techniques," *Journal of Science Technology and Research (JSTAR)*, vol. 6, no. 1, 2025.

[19] M. Y. Hidayat, M. A. Yaqin, and Z. Abidin, "Semantic-Enhanced News Clustering Using TF-IDF and WordNet with K-Means," *Journal of Information System and Informatics*, vol. 7, no. 4, pp. 3924–3951, Dec. 2025.

[20] S. A. Putri, K. D. Tania, and Naretha Kawadha Pasemah Gumay, "Knowledge Discovery Through Sentiment Analysis and Topic Modeling of BCA Mobile and MyBCA," *Journal*

of Mathematics, Computations, and Statistics, vol. 8, no. 2, pp. 669–682, Oct. 2025, doi: 10.35580/jmathcos.v8i2.9782.

[21] R. J. Dhanal and V. R. Ghorpade, "Aspect-based sentiment-analysis using topic modelling and machine-learning," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 6, p. 6689, Dec. 2024, doi: 10.11591/ijece.v14i6.pp6689-6698.

[22] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," Mar. 2022, *arXiv*: 2203.05794. doi: 10.48550/arXiv.2203.05794.

[23] F. Firdaus *et al.*, "Text clustering for analyzing scientific article using pre-trained language model and k-means algorithm," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 5, pp. 3728–3740, Oct. 2025, doi: 10.11591/eei.v14i5.9670.

[24] H. Marfani, H. Kamal, A. S. Hussain, and D. Syed, "A comparison of k-means and mini-batch k-means algorithm for customer segregation analysis," *Societal Transformation: AI and Big Data*, vol. 3, no. 1, pp. 27–49, Jul. 2025, doi: 10.20547/aibd.253103.

[25] F. Iglesias Vázquez and T. Zseby, "Temporal silhouette: validation of stream clustering robust to concept drift," *Machine Learning*, vol. 113, no. 4, pp. 2067–2091, Apr. 2024, doi: 10.1007/s10994-023-06462-2.

[26] C. Tang, "Review on Application of Chi-square Statistic in Text Classification in Recent Five Years," *Academic Journal of Engineering and Technology*, vol. 97, no. 1, pp. 115–118, Nov. 2024, doi: 10.54254/2755-2721/97/20241397.

[27] Y. Ortakci and B. Borhan, "Optimizing SBERT for long text clustering: two novel approaches with empirical insights," *The Journal of Supercomputing*, vol. 81, no. 8, p. 950, Jun. 2025, doi: 10.1007/s11227-025-07414-4.

[28] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *EURASIP Journal on Wireless Communications and Networking*, vol. 2021, no. 1, p. 31, Dec. 2021, doi: 10.1186/s13638-021-01910-w.

[29] J. Cohen, *Statistical power analysis for the behavioral sciences*, 2nd ed. Hillsdale, N.J: L. Erlbaum Associates, 1988.