An Integrated Topic–Sentiment Analysis of User Reviews in Vidio Application Using BERTopic and IndoRoBERTa

Nalendra Whisnu Pinilih [1]*, Ika Novita Dewi [2]*, Farrikh Alzami [3]*
*Faculty of Computer Science, Universitas Dian Nuswantoro
112202206931@mhs.dinus.ac.id [1], ikadewi@dsn.dinus.ac.id [2], alzami@dsn.dinus.ac.id [3]

Article Info

Article history:

Received 2026-05-07
Revised 2026-05-24
Accepted 2026-06-11

Keyword:

Topic Modeling,
Sentiment Analysis,
BERTopic,
IndoRoBERTa,
User Reviews.

ABSTRACT

User reviews on digital platforms provide valuable insights into user experience; however, the large volume and unstructured nature of such data make systematic analysis challenging. In the case of the Vidio application, user feedback frequently reflects concerns related to advertisements, subscription systems, and technical performance. Nevertheless, existing researches often apply sentiment analysis and topic modeling separately, limiting the ability to understand how specific discussion themes are associated with user sentiment. To address this limitation, this research proposes an integrated topic–sentiment analysis approach for analyzing user reviews of the Vidio application collected from the Google Play Store. After filtering and quality control, 8,854 reviews were retained for further analysis using BERTopic for topic modeling and IndoRoBERTa for sentiment classification. The topic modeling process was optimized through parameter tuning, resulting in an improvement of the coherence score from 0.4076 to 0.6878, indicating better semantic consistency among the identified topics. Meanwhile, the sentiment classification model achieved an accuracy of 72%, although its performance was affected by class imbalance, particularly in identifying neutral sentiment. The analysis identified seven primary topics, where advertising-related issues emerged as the dominant topic and were strongly associated with negative sentiment, followed by concerns regarding subscription mechanisms and login accessibility. In contrast, content-related topics, particularly sports broadcasts, were consistently associated with positive sentiment. Furthermore, statistical evaluation confirmed a significant relationship between topic categories and sentiment distribution. Overall, the findings demonstrate that integrating topic modeling and sentiment analysis provides a more comprehensive understanding of user opinions and can support improvements in application quality and user experience.

I. Introduction

The rapid growth of the video-on-demand (VoD) industry in Indonesia reflects the increasing consumption of digital content by society[1]. Streaming platforms are not only required to provide diverse entertainment content but also to ensure high-quality digital services, including system stability, application performance, and optimal user experience [1]. One prominent platform is Vidio, which has a substantial user base and offers premium licensed sports content such as Liga 1 and the UEFA Champions League[2]. User interactions with such platforms are frequently expressed through reviews on the Google Play Store, which contain valuable information regarding application performance, service quality, and user satisfaction[3]. However, the large volume of reviews as illustrated in Figure 1, combined with unstructured textual characteristics such as informal language, abbreviations, and code-mixing, makes manual analysis inefficient and prone to bias due to inconsistency among annotators [2], [3]. Therefore, automated approaches based on Natural Language Processing (NLP) are required to extract meaningful insights from user-generated content [4].

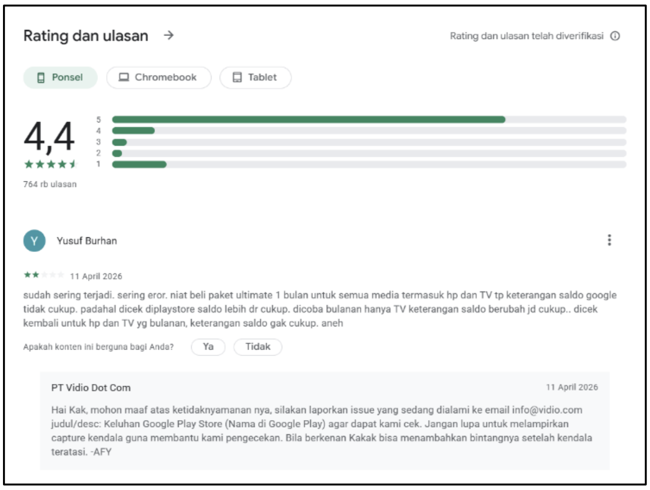

Figure 1. User reviews Video App on Google Play Store

Two widely used approaches in NLP for analyzing user opinions are sentiment analysis and topic modeling. Sentiment analysis aims to classify the polarity of opinions into categories such as positive, negative, or neutral [5], while topic modeling is used to identify latent themes within textual data [1]. Despite their effectiveness, these approaches are often applied separately, leading to a fragmented understanding of user feedback. Sentiment analysis provides information about emotional polarity without identifying the specific aspects being evaluated, whereas topic modeling reveals thematic structures without capturing user attitudes toward those topics [4], [6]. Consequently, it remains difficult to determine which specific issues are associated with particular sentiment patterns.

Previous research has applied various methods to analyze user reviews on digital platforms. Traditional approaches such as Latent Dirichlet Allocation (LDA) have been widely used for topic modeling [1], while classification methods such as Naïve Bayes, Support Vector Machine, and transformer-based models like IndoBERT have been employed for sentiment analysis [6], [7], [8]. However, distribution-based methods such as LDA often struggle to capture complex semantic relationships in natural language [5]. Recent developments in embedding-based approaches, such as BERTopic, have shown improved performance in generating coherent and interpretable topics by leveraging contextual embeddings [7], [10], [11]. Meanwhile, transformer-based models such as IndoRoBERTa have demonstrated strong capability in Indonesian sentiment analysis, although their application in integrated analytical frameworks remains limited [12].

Based on these limitations, this research proposes an integrated topic–sentiment analysis approach by combining BERTopic and IndoRoBERTa to analyze user reviews of the Vidio application, enabling the mapping of thematic structures to sentiment polarity for a more comprehensive understanding of user perceptions. Specifically, the research aims to identify dominant discussion topics, analyze sentiment distribution across topics, and examine the relationship between topics and user sentiment in order to generate more contextual insights into user experience. To strengthen semantic interpretability, the research evaluates topic quality through parameter optimization and coherence analysis, while also statistically examining the relationship between discussion topics and sentiment distribution. The findings are expected to provide practical insights for improving application services and contribute to the development of NLP-based opinion analysis, particularly within the Indonesian digital streaming context.

II. METHODS

This research employs an integrated topic–sentiment analysis approach to analyze user reviews of the Vidio application. The methodology consists of data collection, data preparation, text preprocessing, topic modeling using BERTopic, sentiment classification using IndoRoBERTa, model evaluation, integration, and statistical analysis. The overall workflow is illustrated in Figure 2.

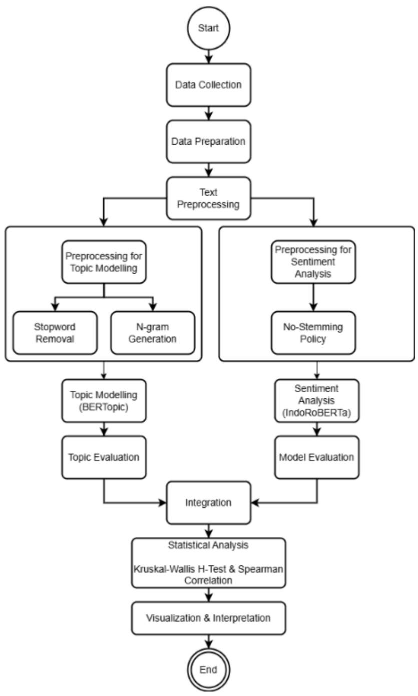

Figure 2. Research workflow

The preprocessing stage is designed using two separate pipelines. For topic modeling, preprocessing includes stopword removal and n-gram generation to extract meaningful terms. For sentiment analysis, preprocessing focuses on preserving polarity through no-stemming policy.

Topic modeling is performed using BERTopic and evaluated using coherence score, while sentiment classification is conducted using IndoRoBERTa and evaluated using accuracy and F1-score. The results are then integrated to map sentiment distribution across topics and analyzed using statistical methods, including the Kruskal–Wallis test and Spearman correlation.

A. Data Collection

The dataset used in this research consists of publicly available user reviews of the Vidio application collected from the Google Play Store using the *google-play-scraper* library. Reviews were retrieved using the application identifier *com.vidio.android* and limited to Indonesian-language content. The collection process was conducted over the period from May 2, 2025, to April 2, 2026, resulting in 15,000 reviews. All data were obtained from publicly accessible sources and do not contain personally identifiable information. An example of the collected data is presented in Table I.

TABLE I
SAMPLE OF COLLECTED USER REVIEWS

Text	Rating	Date
Terimakasih untuk yang membuat aplikasi vidio ini The best pisan	5	2025-04-01 07:07:35
coba aja nggk ada vidio premium,pasti saya nonton terus dramanya,kenapa harus ada pembayaran sih,coba dong di hilangin	1	2025-04-01 06:52:52
aplikasi video ini seru karena bisa nonton apa saja dan murah	5	2025-04-01 04:24:00

Each review consists of three main attributes, namely review text, user rating (on a scale of 1–5), and review date. To ensure data quality, duplicate entries and reviews with very short or empty textual content (fewer than three words) were removed. In addition to duplicate removal and minimum text-length filtering, several quality control procedures were applied to reduce noise in the dataset. Reviews containing empty or non-informative textual content, excessive symbol repetition, or irrelevant entries with insufficient semantic information were excluded from further analysis. The filtering process was conducted to improve textual consistency and ensure that the retained reviews contained meaningful information suitable for topic modeling and sentiment classification. After all filtering and quality control procedures were completed, a total of 8,854 reviews were retained for further analysis.

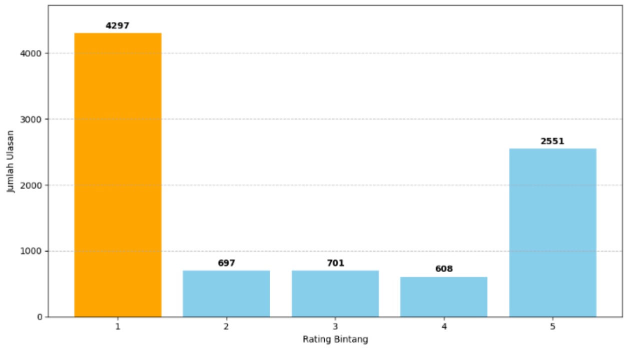

Figure 3. Distribution of user ratings in the filtered dataset

The distribution of user ratings after filtering is presented in Figure 3. Based on the figure 3, rating 1 has 4,297 reviews, rating 2 has 697 reviews, rating 3 has 701 reviews, rating 4 has 608 reviews, and rating 5 has 2,551 reviews. This shows that the dataset is not evenly distributed across rating categories, with a higher concentration of reviews at the lowest rating level compared to other categories.

B. Data Preparation

The data preparation stage aims to construct a reliable and high-quality dataset prior to text processing and model development. This stage focuses on reducing data-level noise and generating labeled data required for supervised sentiment classification [13].

First, a filtering process was applied to eliminate irrelevant and low-quality entries. Duplicate reviews were removed to prevent redundancy, while reviews containing insufficient textual content, such as empty entries or texts with fewer than three words, were excluded due to their limited semantic information. This step ensures that each document contains meaningful linguistic content suitable for further analysis.

Second, sentiment labels were generated based on user-provided ratings to enable supervised learning. A rule-based mapping strategy was employed, where reviews with ratings of 1–2 were assigned to the negative class, rating 3 to the neutral class, and ratings of 4–5 to the positive class. This labeling approach leverages user evaluations as a proxy for sentiment polarity, allowing large-scale annotation without manual labeling. Although rating-based labeling may introduce noise due to subjective user judgments and inconsistencies between rating and textual sentiment, it remains a widely adopted approach in large-scale sentiment analysis due to its practicality and scalability. The distribution of sentiment is presented in Table II.

TABLE II
DISTRIBUTION OF SENTIMENT CLASSES IN THE DATASET

Class	Total Count
Positive	2052
Neutral	830
Negative	5972

The sentiment distribution indicates a substantial class imbalance, where negative reviews dominate the dataset compared to neutral and positive classes. This imbalance reflects the natural tendency of users to provide reviews primarily when experiencing dissatisfaction with the application. Such uneven distribution may influence the learning process of the sentiment classification model, particularly reducing its ability to accurately recognize minority classes such as neutral sentiment.

C. General Text Normalization

General text normalization is performed to standardize textual data and reduce noise prior to task-specific preprocessing. This stage includes case folding to ensure uniform text representation, URL removal to eliminate non-semantic elements, slang normalization to convert informal expressions into standard language, and repeated character handling to reduce unnecessary token variations while preserving semantic meaning.

1) *Case Folding*: Case folding converts all characters in the document into lowercase to ensure uniform text representation [8], as expressed in Equation (1).

$$D_1 = \text{lower}(D) \tag{1}$$

where D_1 represents the document after case folding.

2) *URL Removal*: URL removal eliminates hyperlinks that do not contribute semantic information to the analysis [14], as defined in Equation (2).

$$D_2 = D_1 - U \tag{2}$$

where U denotes the set of URL tokens and D_2 is the resulting document.

3) *Slang Normalization:* Slang normalization transforms informal or non-standard words into their standardized forms using a dictionary-based mapping function [3], as shown in Equation (3).

$$w' = F_{norm}(w) \tag{3}$$

where w' is a non-standard word and w is its normalized form.

4) *Repeated Character Handling*: Repeated character handling reduces excessive character repetitions to prevent redundant token variations [1], as defined in Equation (4).

$$w' = F_{rep}(w, n) \tag{4}$$

where n represents the number of repeated characters and w' is the normalized word.

D. Topic Pre-Processing

The preprocessing pipeline for topic modeling is designed to extract semantically meaningful features that enhance topic coherence and interpretability. Unlike sentiment analysis, which preserves full contextual information, this stage focuses on reducing noise and retaining informative lexical units that represent the main themes within the corpus.

1) *Stopwords Removal*: Stopwords removal is applied to eliminate common words that do not contribute significant semantic meaning, such as conjunctions and function words. In addition to standard Indonesian stopwords, domain-specific terms that frequently appear across documents but provide limited discriminative value are also removed [6]. This process improves topic clarity by reducing irrelevant token dominance. This process is formulated in Equation (6).

$$D' = D - S \tag{6}$$

where D represents the original document, S denotes the set of stopwords, and D′ is the resulting document after stopwords removal.

2) *N-Gram Generation:* To capture contextual relationships between frequently co-occurring words, n-gram generation is applied, particularly in the form of bigrams and trigrams. This allows the model to represent phrases as single semantic units, improving topic specificity and coherence [15]. The n-gram construction is defined as (7).

$$G = \{(w_i, w_{i+1}, \dots, w_{i+n-1}) \mid 1 \leq i < n\} \tag{7}$$

where G denotes the set of generated n-grams and w_i represents the i-th word in the document.

E. Sentiment Preprocessing

The preprocessing pipeline for sentiment analysis is designed to preserve contextual and emotional information within the text to ensure accurate sentiment classification [16]. Unlike topic modeling, this stage minimizes text transformation in order to maintain semantic integrity and polarity cues embedded in the original sentence.

1) *Zero-Stemming Policy*, meaning that no stemming or lemmatization is performed during preprocessing. This approach preserves morphological forms that may carry important sentiment information [17]. This ensures that sentiment information embedded in inflected verbs and adjectives remains intact, as shown in Equation (8).

$$w' = w \tag{8}$$

Where *w* is original token and w' token after preprocessing.

F. Topic Modeling using BERTopic

Topic modeling in this research is conducted using the BERTopic method, an embedding-based approach that leverages contextual representations from transformer models to generate coherent and interpretable topics [4]. Unlike traditional probabilistic methods, BERTopic utilizes semantic embeddings to capture contextual relationships between documents, enabling more accurate topic representation. The overall BERTopic workflow is illustrated in Figure 4.

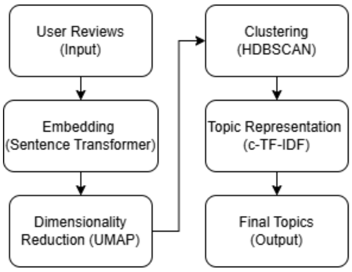

Figure 4. BERTopic workflow for topic modeling

The topic modeling process consists of four main stages. First, document embedding transforms each review into a high-dimensional semantic vector representation using a pre-trained sentence transformer model. This transformation can be represented as Equation (9).

$$v_i = f_{emb}(d_i) \quad (9)$$

where d_i denotes the i-th document and v_i is its corresponding embedding vector.

Second, dimensionality reduction is performed using the Uniform Manifold Approximation and Projection (UMAP) algorithm to project high-dimensional embeddings into a lower-dimensional space while preserving local and global structure. This step improves clustering efficiency and can be expressed as Equation (10).

$$z_i = f_{umap}(v_i) \quad (10)$$

where z_i represents the reduced-dimensional representation of document i.

Third, document clustering is conducted using the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm. This method groups semantically similar documents into clusters without requiring a predefined number of topics as shown in Equation (11).

$$C = f_{cluster}(z) \quad (11)$$

where *C* denotes the set of document clusters.

Finally, topic representation is generated using the class-based Term Frequency–Inverse Document Frequency (c-TF-IDF) method to extract representative keywords for each cluster [11] as calculated in Equation (12).

$$c - TF - IDF(t, c) = \frac{f_{t,c}}{\sum_{t' \in c} f_{t',c}} \cdot log \frac{N}{n_t} \quad (12)$$

where $f_{t,c}$ is the frequency of term *t* in cluster *c*, *N* is the total number of documents, and n_t is the number of documents containing term *t*. Through this multi-stage process, BERTopic generates a set of topics that represent the dominant themes within the user reviews.

Parameter optimization was conducted experimentally by evaluating several BERTopic configurations involving UMAP, HDBSCAN, and vectorizer settings. The tuning process focused on improving topic coherence while maintaining interpretable and semantically distinct topic clusters. Several parameter combinations were tested by adjusting the number of neighbors in UMAP, cluster density thresholds in HDBSCAN, and n-gram extraction ranges in the vectorizer component. The final configuration was selected because it produced the highest coherence score and generated more stable and interpretable topic representations compared to the baseline configuration. The parameters used in this research are presented in Table II.

TABLE II
BERTOPIC PARAMETER CONFIGURATION

Component	Parameter	Setting
Embedding	Model Name	paraphrase-multilingual-MiniLM-L12-v2
UMAP	n_neighbors	50
	n_components	10
	min_dist	0.1
	metric	cosine
HDBSCAN	min_cluster_size	30
	min_samples	10
	cluster_selection_method	eom
Vectorizer	ngram_range	(2, 3)
	stop_words	Indonesian + domain-specific stopwords
	min_df	2
c-TF-IDF	reduce_frequent_words	True
Fine-tuning	nr_topics	auto

The selected configuration using n_neighbors = 50, min_cluster_size = 30, and ngram_range = (2,3) was found to provide better semantic grouping and reduce fragmented topic clusters. In particular, the use of bi-gram and tri-gram representations improved contextual phrase extraction, while the HDBSCAN configuration enabled more stable clustering for short-text user reviews.

G. Sentiment Analysis using IndoRoBERTa

Sentiment analysis in this research is conducted using the IndoRoBERTa model, a transformer-based language model specifically designed for Indonesian text [12]. IndoRoBERTa leverages contextual embeddings to capture semantic and syntactic relationships within a sentence, enabling accurate classification of sentiment polarity. The overall workflow of sentiment classification using IndoRoBERTa is illustrated in Figure 5.

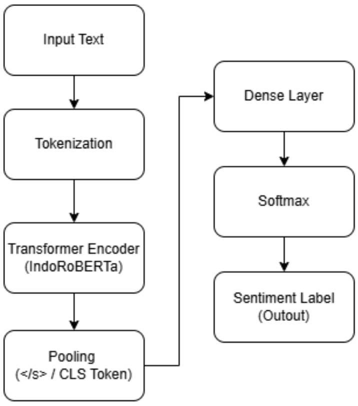

Figure 5. Sentiment classification workflow using IndoRoBERTa

Each preprocessed review is first transformed into a sequence of tokens and encoded into vector representations suitable for model input. This process can be formally defined as Equation (13).

$$x_i = f_{token}(d_i) \quad (13)$$

where d_i represents the input document and x_i denotes the tokenized representation.

The encoded input is then processed through multiple transformer layers, which apply self-attention mechanisms to capture contextual dependencies between tokens. The contextual embedding is obtained as Equation (14).

$$h_i = f_{enc}(x_i) \quad (14)$$

where h_i is the contextual representation of the document.

Finally, a classification layer is applied to predict the sentiment label as Equation (15).

$$\widehat{y_i} = f_{cls}(h_i) \quad (15)$$

where $\widehat{y_i} \in \{positive, neutral, negative\}$.

The predicted sentiment labels are then used to analyze sentiment distribution across topics generated by the BERTopic model [18]. This integration enables the identification of how specific themes are associated with user sentiment. The parameters used in this research are presented in Table III.

TABLE III
INDOROBERTA PARAMETER CONFIGURATION

Component	Parameter	Setting
Model Base	Model Name	indonesian-roberta-base-sentiment-classifier
Input Handling	max_length	512
	truncation	True
Inference	batch_size	32

H. Integration and Statistical Evaluation

This stage aims to integrate the results of topic modeling and sentiment analysis to examine the relationship between identified topics and user sentiment. The integration process is conducted by mapping each document to its corresponding topic label and predicted sentiment label, forming a combined analytical dataset.

Formally, each document d_i is associated with a topic t_i and a sentiment label s_i as defined in Equation (16).

$$d_i \rightarrow (t_i, s_i) \quad (16)$$

where t_i represents the topic assignment generated by BERTopic, and $s_i \in \{potitive, neutral, negative\}$ denotes the sentiment classification obtained from IndoRoBERTa.

In addition, statistical analysis is performed to examine the relationship between identified topics and sentiment distribution.

1) *Coherence Score:* is used to evaluate the semantic consistency of words within each topic, reflecting the interpretability and quality of the generated topics. A higher coherence score indicates that the words within a topic are more semantically related [11]. The coherence score is defined as Equation (17).

$$C = \sum_{i<j} \text{sim}(w_i, w_j) \quad (17)$$

where w_i, w_j represent pairs of words within a topic, *sim(.)* denotes the semantic similarity function between words, and *C* represents the topic coherence value.

2) *Kruskal–Wallis Test:* To determine whether sentiment distributions differ significantly across topics, the Kruskal–Wallis test is applied as a non-parametric statistical method. This test evaluates whether multiple independent samples originate from the same distribution [19]. The Kruskal–Wallis test was selected because sentiment distributions across topics do not necessarily satisfy normality assumptions and involve multiple independent topic groups. The test statistic is defined as Equation (18).

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i R_i^2 - 3(N+1) \quad (18)$$

where *N* represents the total number of observations, *k* denotes the number of groups or topics, n_i indicates the number of samples in the *i*-th group, and R_i represents the average rank of the *i*-th group.

3) *Spearman Rank Correlation:* To measure the relationship between topic prevalence and sentiment intensity, the Spearman rank correlation coefficient is employed [20]. This method assesses the strength and direction of a monotonic relationship between two variables. Spearman rank correlation was employed because it is suitable for measuring monotonic relationships between non-parametric variables and does not require linear relationships between observations. The coefficient is defined as Equation (19).

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (19)$$

where ρ represents the Spearman correlation coefficient, d_i denotes the difference in ranks between the *i*-th pair of observations, and n represents the number of data pairs.

I. Visualization and Interpretation

The final stage of this research involves visualizing and interpreting the integrated results of topic modeling and sentiment analysis to understand the relationship between discussion themes and user responses. The integrated dataset, which combines topic assignments and sentiment labels, is analyzed to identify patterns of sentiment distribution across topics.

Several visualization techniques are employed to support interpretation. First, sentiment distribution plots are used to illustrate the proportion of positive, neutral, and negative sentiments within each topic. Second, topic–sentiment mapping is presented to highlight the association between specific topics and dominant sentiment trends. Third, keyword-based topic representations are visualized to facilitate the interpretation of thematic structures. In addition, word frequency visualizations and word clouds are used to highlight dominant terms within each sentiment category.

III. Results and Discussion

A. Text Pre-processing Results

The text data preprocessing phase was executed systematically to eliminate noise and standardize the textual format, thereby enhancing the accuracy of the subsequent topic and sentiment analyses. The outcomes of this preprocessing stage are detailed in Table IV.

TABEL IV
RESULTS OF THE TEXT PREPROCESSING STAGE

Original Text	Preprocessed Text
bagus tapi siaran RCTI nya tidak ada	bagus tapi siaran rcti nya tidak ada
Terimakasih untuk yang membuat aplikasi vidio ini The best pisan	terimakasih untuk yang membuat aplikasi vidio ini the best pisan
coba aja nggk ada vidio premium, pasti saya nonton terus dramanya, kenapa harus ada pembayaran sih, coba dong di hilangin	coba aja gak ada vidio premium, pasti saya nonton terus dramanya, kenapa harus ada pembayaran sih, coba dong di hilangin

In the topic modeling stage using BERTopic, the initially preprocessed text is strictly filtered again through a stopword removal phase. This process combines the default Sastrawi dictionary with a domain-specific list of custom stopwords (such as "aplikasi", "vidio", "nya", "the", "best", "aja", and "sih") to eliminate frequently occurring words that lack strong informative value for topic formation. Once the text is cleared of stopwords, textual feature representations are extracted using the *CountVectorizer* component with the *ngram_range = (2, 3)* setting. Instead of treating the text as a collection of isolated single words (unigrams), the model specifically captures phrases consisting of two to three words simultaneously such as the phrases *"bagus siaran rcti"* or *"premium nonton dramanya"*. The use of this n-gram range is highly crucial to ensure that the context and specific meanings between words remain intact when the model forms topic clusters from user reviews. These results can be seen in Table V.

TABEL V
TEXT TRANSFORMATION AFTER STOPWORD FILTERING AND N-GRAM EXTRACTION

Preprocessed Text	Text After Stopword Removal	Generated Bi-Gram and Tri-Gram Features
bagus tapi siaran rcti nya tidak ada	bagus siaran rcti tidak	bagus siaran, siaran rcti, rcti tidak, bagus siaran rcti, siaran rcti tidak
terimakasih untuk yang membuat aplikasi vidio ini the best pisan	terimakasih pisan	terimakasih pisan
coba aja gak ada vidio premium, pasti saya nonton terus dramanya, kenapa harus ada pembayaran sih, coba dong di hilangin	premium nonton dramanya pembayaran hilangin	premium nonton, nonton dramanya, pembayaran hilangin, premium nonton dramanya

During the sentiment classification phase utilizing the pre-trained Transformer model (RoBERTa), a *no-stemming policy* was deliberately adopted to preserve the morphological and semantic integrity of the textual data. Unlike traditional NLP techniques that strip affixes to reduce words to their root forms, this research intentionally omits the stemming process. Transformer-based architectures rely heavily on contextualized embeddings, meaning they require original word forms and their surrounding linguistic context to accurately extract nuanced semantic representations.

B. Topic Modelling Results

The performance of the topic modeling process is evaluated using the coherence score, which measures the semantic consistency among words within each topic. A higher coherence score indicates that the extracted topics are more meaningful and interpretable.

Based on the experimental results, the baseline BERTopic model achieves a coherence score of 0.4076, indicating moderate topic quality. After applying parameter optimization, the coherence score improves significantly to 0.6878, demonstrating that the optimized configuration produces more coherent and semantically meaningful topics.

In the context of short-text user reviews, a coherence score of 0.6878 can be considered relatively good, particularly because application reviews typically contain informal language, abbreviations, noisy expressions, and highly varied discussion contexts. Previous researches on digital platform review analysis generally report moderate coherence values due to the semantic sparsity and unstructured nature of user-generated content. Comparatively, this result outperforms the research conducted by Wahyuni et al. [21], which reported a coherence score of 0.46 for their optimized BERTopic model. Therefore, the obtained coherence score indicates that the optimized BERTopic configuration in this study was capable of generating more semantically consistent and highly interpretable topic clusters. The comparison between the baseline and optimized models is presented in Table VI.

TABEL VI
TOPIC MODELING PERFORMANCE RESULTS

Model	Coherence Score
Baseline BERTopic	0.4076
Optimized BERTopic	0.6878

The optimized BERTopic model generates a set of dominant topics representing user discussions in the dataset. Each topic is characterized by a group of representative keywords derived using the c-TF-IDF method. The extracted topic representations generated by the optimized BERTopic model are presented in Table VII, which lists the most representative keywords for each identified topic. These keyword groupings reveal the dominant themes discussed by users, allowing the topics to be interpreted into meaningful categories.

TABEL VII
TOPIC REPRESENTATION OF THE OPTIMIZED MODEL

Topic ID	Topic Interpretation	Number of Reviews	Representative Keywords
0	Advertising Issues	3,917	Kebanyakan iklan, terlalu banyak iklan, nonton iklan
1	User Satisfaction	1,834	Sangat bagus, sangat baik, Premier League, memuaskan
2	Film Content	802	Film terbaru, drama Korea, Liga Indonesia, update film
3	Subscription Packages	453	Paket Platinum, langganan platinum, Platinum Extra
4	Login Accessibility	282	Nomor HP, no HP, login pake, pake no HP
5	Interface Features	148	Ukuran subtitle, fitur download, besar kecil
6	National TV Channels	137	Cenel RCTI, nonton RCTI, siaran bola, siaran TV
-1	*Outliers* (Unclassified)	1.231	Google Play, suruh beli, live streaming, loading

Based on the table VII, the identified topics reflect key aspects of user experience. The dominant topics observed in the dataset include: a) Content availability and variety; b) Subscription pricing and payment issues; c) Streaming quality and buffering; d) Advertising interruptions; e) Application performance and usability.

The semantic relationships between the identified topics are visualized in Figure 6, which presents the intertopic distance map generated by the BERTopic model. In this visualization, each topic is represented as a cluster, and the distance between clusters reflects the degree of semantic similarity between topics. Topics that are positioned closer together indicate overlapping or related discussion themes, while those that are farther apart represent more distinct topics. This visualization confirms that the optimized model is able to separate topics with clear boundaries while still preserving meaningful relationships among them, thereby supporting the overall coherence and interpretability of the extracted topics.

Topic interpretation was conducted manually by examining the representative keywords generated through the c-TF-IDF process together with a subset of user reviews contained within each cluster. The topic labels were assigned based on the dominant semantic meaning and contextual similarity among the representative terms and associated review contents. This approach was intended to ensure that each identified topic reflected a coherent and interpretable discussion theme. To improve interpretability, manual validation was also performed by reviewing sample documents from each topic cluster to verify the semantic consistency between representative keywords and the underlying review content. This validation process helped reduce ambiguity in topic labeling and supported the reliability of topic interpretation.

Figure 6 Intertopic Distance Map

The representative keywords and their importance scores are illustrated in Figure 7. Figure 6 shows that each topic forms a distinct cluster, where the relative size indicates the number of associated reviews and the distance between clusters reflects semantic similarity. Figure 7 complements this by presenting the most representative keywords for each topic, enabling clearer interpretation of the underlying discussion themes. Based on these visualizations, the identified topics can be interpreted as follows:

Topic 0 (Advertisement Issues): This topic represents the largest cluster, indicating that it contains the highest volume of user discussions. The topic is dominated by keywords such as “banyak iklan”, “terlalu banyak iklan”, and “iklan banyak”, reflecting user dissatisfaction with excessive advertisements in the application.

Topic 1 (User Satisfaction): This topic forms a strong positive cluster with 1,834 reviews. The dominant keywords, including “sangat bagus”, “bagus banget”, and “sangat puas”, indicate high user satisfaction, particularly related to content quality such as sports broadcasts.

Topic 2 (Film and Content Availability): This topic captures discussions related to entertainment content, with keywords such as “film”, “film terbaru”, and “drama Korea”. It reflects user expectations regarding the availability and variety of content.

Topic 3 (Subscription Packages): This topic focuses on subscription-related issues, as indicated by keywords such as “paket platinum”, “langganan platinum”, and “platinum nonton”. It reflects user experiences with paid service tiers.

Topic 4 (Login Accessibility): This topic highlights technical difficulties related to user authentication, particularly using mobile phone numbers, as shown by keywords such as “login pakai no hp” and “nomor hp”.

Topic 5 (Interface Features): This topic captures usability concerns, with keywords such as “fitur download”, “ukuran subtitle”, and “kurang nyaman”, indicating issues related to application functionality and user interface.

Topic 6 (National TV Channels): This topic addresses access to live television content, with keywords such as “nonton rcti” and “channel tv”, reflecting user concerns about the availability of specific broadcast channels.

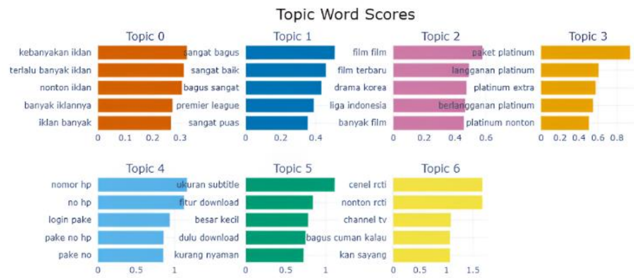

Figure 7 Representative Keyword Scores for Each Identified Topic Generated by BERTopic

The hierarchical clustering analysis, as illustrated in Figure 8, reveals the semantic relationships between the identified topics based on their similarity. The results show that Topic 4 (Login Accessibility) and Topic 5 (Interface Features) are closely clustered, indicating that these topics share strong semantic similarities. This suggests that users tend to perceive login difficulties and feature-related issues as part of a broader category of technical problems affecting application usability.

In contrast, topics related to content and user satisfaction are positioned farther apart in the hierarchy, indicating that they represent distinct aspects of user experience. This separation confirms that the model successfully differentiates between functional issues and content-related discussions.

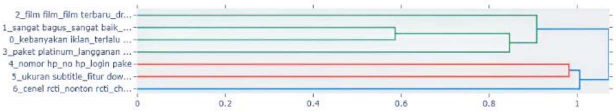

Figure 8 Hierarchical Clustering of Identified Topics Based on Semantic Similarity

C. *Sentiment Analysis Results*

The performance of the sentiment classification model is evaluated using a confusion matrix, as presented in Table VIII, which compares the actual sentiment labels with the predicted labels generated by the IndoRoBERTa model. The matrix provides a comprehensive view of classification performance across all sentiment classes, highlighting both correct predictions and misclassifications. As shown in Table VIII, the model demonstrates strong performance in identifying clearly expressed positive and negative sentiments, as indicated by higher values along the diagonal. However, misclassifications are more prominent in the neutral class, where instances are frequently predicted as either positive or negative. This suggests that the model has difficulty capturing subtle or ambiguous sentiment expressions, which is likely influenced by the imbalanced distribution of the dataset.

TABEL VIII
CONFUSION MATRIX OF SENTIMENT CLASSIFICATION RESULTS

Actual \ Predicted	**Positive**	**Neutral**	**Negative**
Positive	1840	365	954
Neutral	81	72	548
Negative	131	393	4470

The performance of the baseline sentiment classification model is presented in Table IX, which summarizes precision, recall, F1-score, and overall accuracy for the three sentiment classes. The model achieves an overall accuracy of 72% with a weighted F1-score of 0.7192, indicating reasonably good performance in classifying user reviews. The results show that the model performs well on the negative class, achieving an F1-score of 0.82 and a high recall of 0.90, indicating that most negative reviews are correctly identified. Similarly, the positive class achieves a high precision of 0.90, although its recall is lower (0.58), suggesting that some positive instances are misclassified.

In contrast, the neutral class exhibits significantly lower performance, with an F1-score of only 0.09. This indicates that the model struggles to correctly identify neutral sentiment, often misclassifying it as either positive or negative. This limitation is primarily attributed to the imbalanced dataset, where neutral samples (701 instances) are substantially fewer compared to negative (4,994) and positive (3,159) samples. Additionally, neutral expressions tend to be more ambiguous and less explicitly defined, making them inherently more difficult for the model to classify. Overall, these results demonstrate that while the baseline IndoRoBERTa model is effective for identifying clear positive and negative sentiments, it has limitations in handling subtle or mixed sentiment expressions.

In this research, no explicit resampling or class balancing techniques, such as oversampling, undersampling, or cost-sensitive learning, were applied during model training. Consequently, the classifier tended to favor the dominant negative class, which contributed to lower performance in minority sentiment categories, particularly the neutral class.

Although the model achieved an overall accuracy of 72%, this performance should be interpreted cautiously due to the imbalanced sentiment distribution and the low classification performance observed in the neutral class. The results indicate that the model is more reliable for detecting explicitly positive and negative opinions than for identifying nuanced or ambiguous sentiment expressions commonly found in user-generated reviews. Therefore, the findings of this study should be viewed as providing indicative insights rather than fully precise sentiment categorization.

TABLE IX
PERFORMANCE EVALUATION OF THE BASELINE INDOROBERTA SENTIMENT CLASSIFICATION MODEL

Class	Precision	Recall	F1-Score	Support
Negative	0.75	0.90	0.82	4994
Neutral	0.09	0.10	0.09	701
Positive	0.90	0.58	0.71	3159
Accuracy			0.72	8854
Macro Avg	0.58	0.53	0.54	8854
Weighted Avg	0.75	0.72	0.72	8854

A comparative baseline experiment using LSTM was conducted to evaluate the effectiveness of the proposed IndoRoBERTa approach. The LSTM model achieved an accuracy of 56.41%; however, the classification results were heavily biased toward the majority negative class, with the neutral and positive classes obtaining zero precision and recall scores. This indicates that the LSTM model was unable to adequately capture contextual and semantic variations within the imbalanced Indonesian review dataset. In contrast, IndoRoBERTa demonstrated better contextual representation capability and more balanced sentiment classification performance, supporting the suitability of transformer-based architectures for Indonesian-language sentiment analysis involving informal user-generated text.

Sample sentiment classification results are presented in Table X to illustrate the behavior of the IndoRoBERTa model in predicting sentiment labels. The examples show that the model is able to correctly classify clearly expressed sentiments, particularly for strongly positive and negative reviews. For instance, the review “terimakasih untuk yang membuat aplikasi vidio ini the best pisan” is accurately classified as positive, while strongly negative feedback related to premium content is correctly identified as negative. However, the table also highlights cases of misclassification, such as the review “bagus tapi siaran rcti nya tidak ada”, which is labeled as positive but predicted as negative. This indicates that the model tends to prioritize negative contextual cues when conflicting sentiments appear within the same sentence. These findings suggest that while the model performs well on explicit sentiment expressions, it has limitations in handling mixed or nuanced sentiments.

Despite the proficiency of the transformer-based IndoRoBERTa architecture in capturing contextual semantic dependencies, its performance may deteriorate when encountering implicit sentiments, sarcasm, irony, or highly nuanced expressions prevalent in online user reviews. These complex linguistic structures introduce conflicting contextual signals, presenting substantial challenges for document-level sentiment classification models to achieve high accuracy.

TABLE X
SAMPLE SENTIMENT CLASSIFICATION RESULTS USING INDOROBERTA

Review Text	Actual Sentiment Label	Predicteed Sentiment Label
bagus tapi siaran rcti nya tidak ada	positive	negative
terimakasih untuk yang membuat aplikasi vidio ini the best pisan	positive	positive
coba aja gak ada vidio premium, pasti saya nonton terus dramanya, kenapa harus ada pembayaran sih, coba dong di hilangin	negative	negative

The dominant words in each sentiment category are visualized using word clouds in Figure 9, Figure 10, and Figure 11. As shown in Figure 9, positive sentiment is characterized by words such as “bagus”, “suka”, and “mantap”, indicating user satisfaction toward content and viewing experience. Meanwhile, Figure 10 shows that neutral sentiment contains more balanced and descriptive terms, including “langganan”, “iklan”, and “film”, reflecting informational or mixed opinions. In contrast, Figure 11 demonstrates that negative sentiment is dominated by words such as “iklan”, “langganan”, and “error”, highlighting complaints related to advertisements, subscription systems, and technical issues. These visualizations support the sentiment analysis results by illustrating the dominant expressions associated with each sentiment category.

Figure 9. Positive Sentiment WordCloud

Figure 10. Neutral Sentiment WordCloud

Figure 11. Negative Sentiment WorldCloud

D. Integrated Topic–Sentiment Analysis

The results of the integrated topic modeling and sentiment analysis are presented in Table XI, which summarizes the distribution of positive, neutral, and negative sentiments across each identified topic. The results show that Advertising Issues is dominated by negative sentiment with 2,869 reviews, indicating that advertisements are a major source of user dissatisfaction. In contrast, the User Satisfaction topic contains the highest proportion of positive sentiment, reflecting favorable user responses toward content quality and viewing experience. Topics related to subscription services, login accessibility, and interface features also exhibit predominantly negative sentiment, suggesting that technical and service-related issues significantly influence negative user perceptions. Overall, the table demonstrates that sentiment distribution varies substantially depending on the contextual nature of each topic.

TABEL XI
RESULTS OF INTEGRATED TOPIC MODELLING & SENTIMENT ANALYSIS

Topic Interpretation	Sentiment			Total
	Positive	Neutral	Negative	
Advertising Issues	685	363	2,869	3,917
Outliers (Unclassified)	451	219	1,641	2,311
User Satisfaction	780	162	892	1,834
Film Content	127	38	215	380
Subscription Packages	7	26	212	245
Login Accessibility	0	11	71	82
Interface Features	2	2	44	48
National TV Channels	0	9	28	37

Sample results of the integrated topic modeling and sentiment classification are presented in Table XII. The table illustrates how each review is assigned to a specific topic interpretation and corresponding predicted sentiment label, enabling a more contextual analysis of user feedback. The examples demonstrate that negative sentiment is associated with service accessibility issues, positive sentiment reflects user appreciation toward the application, while neutral sentiment generally appears in reviews containing informational or less emotionally explicit expressions. These results show that the integration of BERTopic and IndoRoBERTa provides a more comprehensive understanding of user opinions by connecting discussion themes with sentiment polarity.

TABEL XII
SAMPLE OUTPUT OF INTEGRATED TOPIC MODELING AND SENTIMENT CLASSIFICATION

Review Text	Topic Interpretation	Predicted Sentiment
bagus tapi siaran rcti nya tidak ada	National TV Channels	negative
terimakasih untuk yang membuat aplikasi vidio ini the best pisan	Advertising Issues	positive
ukuran teks kenapa tidak bisa diubah mohon info	Interface Features	Neutral

The distribution of sentiment across topics is illustrated in Figure 12, showing clear variations in sentiment patterns among the identified discussion themes. Positive sentiment is predominantly concentrated in Topic 1, indicating strong user satisfaction related to content quality and viewing experience. In contrast, Topic 2 is dominated by negative sentiment, reflecting user dissatisfaction associated with advertising issues. Other topics, including subscription packages, login accessibility, and interface features, also exhibit higher proportions of negative sentiment compared to positive sentiment. Overall, the figure demonstrates that user sentiment is strongly influenced by the contextual nature of each topic, where content-related discussions tend to generate positive responses, while technical and service-related issues are more likely to produce negative perceptions.

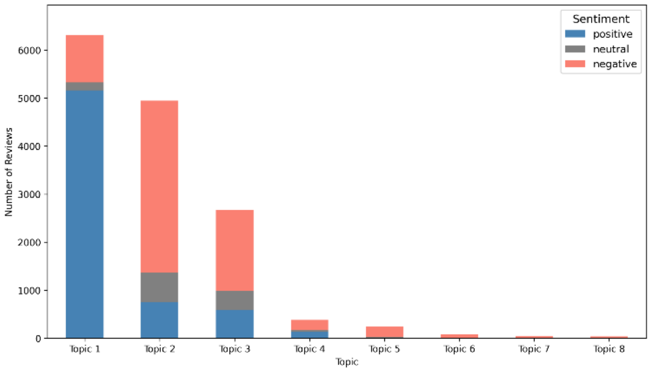

Figure 12. Distribution of Sentiment Per Topic

The relationship between topics and sentiment distribution is visualized in Figure 13 using a heatmap representation. The figure clearly shows that negative sentiment is highly concentrated in Topic 0, which corresponds to advertising-related issues, while positive sentiment is predominantly associated with Topic 1, representing user satisfaction. Other topics, including subscription packages, login accessibility, and interface features, exhibit relatively lower sentiment frequencies but remain dominated by negative sentiment. The heatmap provides a clearer comparison of sentiment intensity across topics, confirming that user perceptions vary significantly depending on the discussion theme.

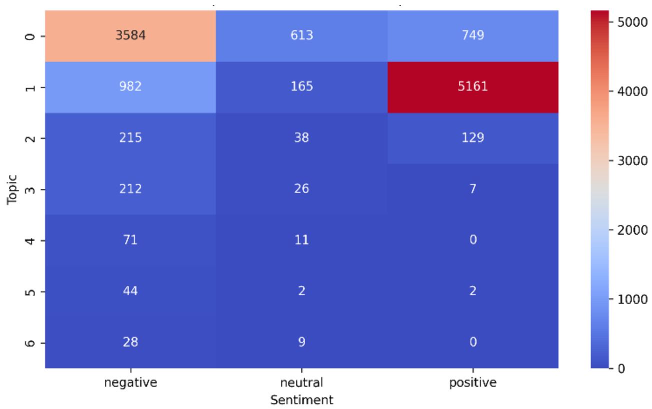

Figure 13. Heatmap Visualization of Sentiment Distribution Across Topics

Figure 14. Percentage Distribution of Sentiment Across Topics

The percentage distribution of sentiment within each topic is presented in Figure 14, highlighting the proportion of positive, neutral, and negative sentiments across the identified discussion themes. The figure shows that Topic 1 is strongly dominated by positive sentiment, indicating high user satisfaction, whereas topics related to advertising issues, subscription packages, login accessibility, and interface features are predominantly negative. Neutral sentiment appears in relatively smaller proportions across most topics, reflecting the tendency of users to express strong opinions rather than moderate evaluations. Overall, the figure demonstrates that sentiment polarity varies considerably depending on the contextual focus of each topic.

The integrated analysis provides several practical implications for application developers. Topics dominated by negative sentiment, particularly advertising issues, subscription packages, and login accessibility, indicate areas that require technical and service-level improvements. For example, reducing excessive advertisement frequency, improving authentication stability for mobile number login, simplifying subscription management, and optimizing streaming accessibility may help improve overall user satisfaction. In addition, interface-related feedback suggests the importance of improving subtitle customization, feature usability, and application responsiveness to enhance user experience.

E. Statistical Evaluation Results

The results of the statistical evaluation are summarized in Table XIII, which present the outcomes of the Spearman rank correlation and Kruskal–Wallis test. The Spearman correlation coefficient of 0.7910 with a significance level of $p < 0.001$ indicates a strong positive association between sentiment scores and topic groupings. However, as topic labels represent categorical variables, this relationship should be interpreted as an association rather than a causal or linear relationship. In addition, the Kruskal–Wallis test yields a statistic value of H = 5072.07 with p < 0.001, confirming a statistically significant difference in sentiment distribution across topics. These findings demonstrate that user sentiment varies meaningfully depending on the discussion theme, supporting the results obtained from the integrated topic–sentiment analysis.

TABEL XIII
RESULTS OF STATISTICAL EVALUATION

Test Method	Statistic Value	P-Value	Interpretation
Spearman Rank Correlation	0.7910	< 0.001	Significant positive association between variables
Kruskal–Wallis Test	5072.07	< 0.001	Significant difference in sentiment across topics

The statistical results strengthen the integrated analysis by quantitatively confirming that sentiment variation is strongly associated with discussion topics. In particular, topics related to advertisements, subscription services, and technical accessibility tend to exhibit significantly higher negative sentiment concentrations compared to content-related topics. These findings indicate that user sentiment is context-dependent and varies meaningfully according to the thematic focus of each topic cluster. However, the correlation result should be interpreted as an associative relationship rather than a causal relationship, since topic labels represent categorical thematic groupings rather than continuous explanatory variables.

F. Limitation of the study

This research has several limitations that should be considered when interpreting the results. First, the imbalanced distribution of sentiment classes introduces bias in the sentiment classification model, particularly reducing its ability to accurately identify neutral sentiment. Second, the use of rating-based labeling may introduce labeling noise, as numerical ratings do not always fully represent the actual sentiment expressed in the review text, especially in reviews containing mixed or ambiguous opinions. In addition, the implementation of document-level sentiment analysis limits the model's ability to capture multiple aspect-specific sentiments within a single review. The dataset is also restricted to user reviews collected from the Google Play Store may contain several forms of data bias and noise. Some reviews may include spam-like content, duplicated opinions, emotionally exaggerated expressions, or very brief comments that provide limited contextual information. Furthermore, online review platforms tend to be dominated by highly active users or users experiencing extreme satisfaction or dissatisfaction, which may lead to opinion bias and reduce the representativeness of general user perceptions. Although duplicate filtering and quality control procedures were applied, it is difficult to completely eliminate all potential noise and bias from user-generated reviews. Therefore, the findings of this research may not be fully generalizable to other digital streaming platforms with different service characteristics, user demographics, content ecosystems, and interaction patterns. Although several identified issues, such as advertisements, subscription systems, and streaming accessibility, may also appear in other streaming applications, the sentiment distribution and dominant discussion topics may vary across platforms. Consequently, the interpretation of the results should remain context-specific to the Vidio application environment.

Future research is recommended to address these limitations by applying techniques for handling class imbalance, such as resampling methods or cost-sensitive learning approaches, and by fine-tuning sentiment classification models using domain-specific datasets. The adoption of aspect-based sentiment analysis is also suggested to obtain more granular insights into user opinions across different application features. Furthermore, expanding the data sources and incorporating additional evaluation and validation methods would improve the robustness, reliability, and generalizability of future researches.

IV. CONCLUSION

This research proposes an integrated approach for analyzing user reviews of the Vidio application by combining BERTopic-based topic modeling and IndoRoBERTa-based sentiment classification. A total of 15,000 reviews were collected from the Google Play Store, of which 8,854 reviews were retained after filtering and preprocessing. The topic modeling process was optimized through parameter tuning, resulting in an improvement of the coherence score from 0.4076 to 0.6878, indicating better semantic consistency among the identified topics. The analysis successfully identified several major discussion themes, including user satisfaction, advertising issues, film content, subscription packages, login accessibility, interface features, and national TV channels. In addition, the sentiment classification model achieved an accuracy of 72%, demonstrating reasonably effective performance for identifying dominant positive and negative sentiment patterns, although limitations remain in recognizing neutral and ambiguous expressions. in identifying positive and negative sentiments, although limitations were observed in recognizing neutral sentiment due to class imbalance within the dataset.

The integrated analysis shows that user satisfaction is primarily associated with content quality and viewing experience, while negative sentiment is strongly related to advertisements, subscription systems, and technical accessibility issues. Statistical evaluation using Spearman rank correlation and the Kruskal–Wallis test further confirms that sentiment distribution differs significantly across topics, indicating that user perception is strongly influenced by the contextual characteristics of each discussion theme. Overall, the findings demonstrate that the integration of topic modeling and sentiment analysis provides a comprehensive understanding of user opinions and can serve as valuable input for improving application quality, service performance, and overall user experience.

REFERENCES

[1] G. Rosalinda, R. Santoso, and P. Kartikasari, "Pemodelan Topik Ulasan Aplikasi Netflix pada Google Play Store Menggunakan Latent Dirichlet Allocation," *J. Gaussian*, vol. 11, no. 4, pp. 554–561, Feb. 2023, doi: 10.14710/j.gauss.11.4.554-561.

[2] R. P. Setiawan, B. Irawan, and W. P. Prihartono, "Analisis Sentimen Ulasan Growtopia di Google Play Store Menggunakan Naïve Bayes Classifier untuk Identifikasi Kebutuhan Pengguna," *J. Inform. Dan Tek. Elektro Terap.*, vol. 13, no. 2, Apr. 2025, doi: 10.23960/jitet.v13i2.6415.

[3] R. A. K. N. Bintang and N. T. Romadloni, "Perbandingan Kinerja Algoritma Klasifikasi pada Review Pengguna Aplikasi Netflix," *J. Inform. Dan Tek. Elektro Terap.*, vol. 13, no. 2, Apr. 2025, doi: 10.23960/jitet.v13i2.6303.

[4] N. A. Adrielvino and A. T. Ayunda, "Penerapan Indobert dan Bertopic Dalam Absa untuk Evaluasi Kualitas Aplikasi E-Government Indonesia | Rabit : Jurnal Teknologi dan Sistem Informasi Univrab," Jan. 2026, Accessed: Mar. 16, 2026. [Online]. Available: https://jurnal.univrab.ac.id/index.php/rabit/article/view/7143

[5] A. Wirayudha, M. Murniyati, and R. Rosdiana, "Analisis Sentimen Terhadap Ulasan Access By KAI Pada Google Play Store Menggunakan Metode Indobert," *Portal Ris. Dan Inov. Sist. Perangkat Lunak*, vol. 3, no. 1, pp. 9–20, Jan. 2025, doi: 10.59696/prinsip.v3i1.69.

[6] E. Nurmawati and A. Amanda, "Analisis Sentimen dan Pemodelan Topik pada Tweet Terkait Data Badan Pusat Statistik," *J. Sist. Inf. Dan Inform.*, vol. 6, no. 2, pp. 165–176, Aug. 2023, doi: 10.47080/simika.v6i2.2789.

[7] M. U. Yanuar and W. Wibowo, "Pemodelan Topik dan Analisis Sentimen pada Ulasan Pengguna Aplikasi Trans Jatim: Topic Modeling and Sentiment Analysis on Trans Jatim Application User Reviews," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 6, no. 1, pp. 156–166, Jan. 2026, doi: 10.57152/malcom.v6i1.2410.

[8] Tarwoto, R. Nugroho, N. Azka, and W. S. R. Graha, "Analisis Sentimen Ulasan Aplikasi Mobile JKN di Google PlayStore Menggunakan IndoBERT," *J. JTIK J. Teknol. Inf. Dan Komun.*, vol. 9, no. 2, pp. 495–505, Apr. 2025, doi: 10.35870/jtik.v9i2.3340.

[9] K. C. Pradhisa and R. Fajriyah, "Analisis Sentimen Ulasan Pengguna E-commerce di Google Play Store Menggunakan Metode IndoBERT," *Build. Inform. Technol. Sci. BITS*, vol. 6, no. 1, pp. 92-104–92−104, Jun. 2024, doi: 10.47065/bits.v6i1.5247.

[10] A. F. Anugrah and R. D. Agatha, "Utilizing IndoBERT and BERTopic to Explore Public Opinion on BPS Instagram Posts," *J. Appl. Inform. Comput.*, vol. 9, no. 5, pp. 2836–2844, Oct. 2025, doi: 10.30871/jaic.v9i5.10327.

[11] D. Aryani, I. L. Kharisma, A. Sujjada, and K. Kamdan, "Topic Modeling of the 2024 Election Using the BERTopic Method on Detik.com News Articles," *Inf. J. Ilm. Bid. Teknol. Inf. Dan Komun.*, vol. 9, no. 2, pp. 171–180, Aug. 2024, doi: 10.25139/inform.v9i2.8429.

[12] F. M. Apriansyah, T. I. Ramadhan, C. R. Hidayat, and A. K. Wijaya, "Perbandingan IndoBERT dan IndoRoBERTa Untuk Analisis Sentimen Pada Film Dokumenter Dirty Vote," *J. Inform. J. Pengemb. IT*, vol. 10, no. 3, pp. 593–605, Jul. 2025, doi: 10.30591/jpit.v10i3.8607.

[13] Y. Asri, D. Kuswardani, W. N. Suliyanti, Y. O. Manullang, and A. R. Ansyari, "Sentiment analysis based on Indonesian language lexicon and IndoBERT on user reviews PLN mobile application," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 38, no. 1, pp. 677–688, Apr. 2025, doi: 10.11591/ijeecs.v38.i1.pp677-688.

[14] I. Mursidah, R. Sanjaya, B. Yulianto, D. Sweetania, and P. Sularsih, "Klasifikasi Sentimen Google Play Store Aplikasi ChatGPT Berbahasa Indonesia Berbasis IndoBERT," *J. Minfo Polgan*, vol. 14, no. 2, pp. 3349–3359, Dec. 2025, doi: 10.33395/jmp.v14i2.15751.

[15] H. P. Doloksaribu and Y. T. Samuel, "Komparasi Algoritma Data Mining untuk Analisis Sentimen Aplikasi Pedulilindungi," *J. Teknol. Inf. J. Keilmuan Dan Apl. Bid. Tek. Inform.*, vol. 16, no. 1, pp. 1–11, Jan. 2022, doi: 10.47111/jti.v16i1.3747.

[16] A. A. Permana, M. W. Prayuda, R. Taufiq, and D. A. Kristiyanti, "Analisis Sentimen Terhadap Vaksin Covid-19 Menggunakan Algoritma Naïve Bayes Classifier," *J. Minfo Polgan*, vol. 11, no. 2, pp. 129–137, Sep. 2022, doi: 10.33395/jmp.v11i2.12346.

[17] R. Hans, "Mengenal Lemmatization dalam Machine Learning NLP." Accessed: Mar. 10, 2026. [Online]. Available: https://dqlab.id/mengenal-lemmatization-dalam-machine-learning-nlp

[18] C. Brando, S. Anggai, and T. Tukiyat, "Analisis Sentimen Ulasan Pengguna Aplikasi Info BMKG pada Google Play Store Menggunakan Model Transformer BERT dan RoBERTa," *J. SISKOM-KB Sist. Komput. Dan Kecerdasan Buatan*, vol. 9, no. 1, pp. 40–49, Sep. 2025, doi: 10.47970/siskom-kb.v9i1.872.

[19] W. Wahyudin, "Aplikasi Topic Modeling pada Pemberitaan Portal Berita Online Selama Masa Psbb Pertama," *Semin. Nas. Off. Stat.*, vol. 2020, no. 1, pp. 309–318, 2020, doi: 10.34123/semnasoffstat.v2020i1.579.

[20] J. Luo *et al.*, "Analyzing sentiments in peer review reports: Evidence from two science funding agencies," *Quant. Sci. Stud.*, vol. 2, no. 4, pp. 1271–1295, Dec. 2021, doi: 10.1162/qss_a_00156.

[21] W. Wahyuni, T. P. Lestari, M. Apriliana, and R. Gumelta, "Implementation of BERTopic for Topic Modeling Analysis of the Free Nutritious Meal Program Based on YouTube Comments," *J. Appl. Inform. Comput.*, vol. 9, no. 4, pp. 1964–1971, Aug. 2025, doi: 10.30871/jaic.v9i4.9754.