Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

June 2026. Vol. 5. No. 3; e-ISSN: 2808-4519

Topic Modeling of Clash of Clans Player Reviews Using NLP-Based Latent Dirichlet Allocation (LDA) Machine Learning Method

Rai Markus Panamuan[1], Debi Handika[2], Muhamad Rizki Pratama[3], Weiskhy Steven Dharmawan[4], Lady Agustin Fitriana[5], Riski Annisa[6*]

[1,2,3,6]Program Studi Informatika
[5]Program Studi Sistem Informasi
[4]Program Studi Sistem Informasi Akuntansi,
raimarkuspanamuan@gmail.com[1],rxnew@gmail.com[2], aaiiqqii25@gmail.com[3],weiskhy.wvn@bsi.ac.id[4], Lady.lag@bsi.ac.id[5], riski.rnc@bsi.ac.id[6*]

Abstract

The rapid growth of the mobile gaming industry has generated millions of player reviews on platforms like the Google Play Store. Clash of Clans, developed by Supercell, is one of the world's most popular mobile strategy games, generating a vast volume of user reviews that are difficult to analyze manually. This study applies Latent Dirichlet Allocation (LDA), a generative probabilistic machine learning model based on Natural Language Processing (NLP), to identify and cluster key topics discussed in player reviews on the Google Play Store. A total of 10,000 player reviews were collected through web scraping, followed by NLP-based text preprocessing including tokenization, stopword removal, and lemmatization. The LDA model was optimized using a coherence score evaluation of 0.512, resulting in the identification of five dominant discussion topics: technical issues and bugs, game updates and balance, gameplay and strategy, monetization and in-app purchases, and social interactions and clan systems. The results show that LDA-based topic modeling provides structured and actionable insights for game developers to understand player feedback and improve game quality. This research contributes to the field of NLP-based mobile game review analysis.

***Keywords**: Clash of Clans; Latent Dirichlet Allocation; Natural Language Processing; Topic Modeling*

1. Introduction

The mobile gaming industry has experienced tremendous growth in recent years, driven by increasing smartphone penetration and internet accessibility worldwide [1]. According to the Newzoo Global Games Market Report, the global mobile gaming industry is projected to generate revenues of over $90 billion by 2023, making it the largest segment in the digital entertainment ecosystem. In Indonesia, over 175 million active smartphone users have driven significant growth in the mobile gaming user base, making Indonesia one of the largest mobile gaming markets in Southeast Asia. In this context, Clash of Clans, developed by Supercell, is one of the most popular strategy mobile games globally, with over 500 million downloads, millions of active players, and billions of user reviews on the Google Play Store.

The sheer volume of player reviews on the Google Play Store poses fundamental challenges to manual analysis, given that popular games can accumulate tens to hundreds of millions of reviews, with increasing numbers of reviews daily. Player opinions are highly diverse, encompassing aspects such as technical bug reports, responses to game updates, assessments of gameplay mechanics, criticism of the monetization system, and game balance issues, making them impossible to handle efficiently and accurately manually. The inability to process player feedback quickly and systematically makes it difficult for game developers to prioritize improvements that are truly needed by the player community [2]. Therefore, an artificial intelligence-based automated text analysis approach capable of processing reviews at scale is urgently needed.Natural Language Processing (NLP) is a subfield of artificial intelligence that focuses on developing computational methods to enable machines to understand and interpret human language automatically[3] . In the context of game review analysis, NLP has been proven effective for various tasks such as opinion mining to extract player opinions, review mining to identify feedback patterns, and app review analysis specifically designed to analyze reviews on digital application distribution platforms [4]. The application of NLP in mobile game review analysis provides real benefits for developers in identifying the most complained about features, monitoring community opinion trends over time, and supporting data-driven decision making in the product development cycle [5].

Most app review analysis research to date relies on sentiment analysis, a method that classifies review text into polarity categories such as positive, negative, or neutral [6]. While sentiment analysis is useful in measuring aggregate player satisfaction, it fails to address the more critical question for developers: what players are actually talking about in their reviews. The limitation of sentiment analysis lies in its

nature of only generating polarity labels without providing in-depth semantic information about specific topics, such as whether the main complaints relate to server issues, combat mechanics, or in-game purchase policies [7]. To address this need, a more semantically expressive approach, namely topic modeling techniques, is needed, which can uncover hidden themes behind a collection of review texts automatically and without supervision. In the LDA model, the generative process assumes that each word in a document is selected based on two Dirichlet distributions: the distribution of topics within documents and the distribution of words within topics, thus allowing flexible topic inference without pre-defined class labels [8]. LDA has been widely applied to large-scale review analysis, including identifying user discussion themes across various digital platforms, and has been shown to produce semantically coherent and interpretable topic representations.
Several previous studies have proven the effectiveness of LDA in the mobile application review domain. Kustyaningsih et al [9] applied LDA to 2,014 EdLink application reviews on the Google Play Store and successfully identified three main aspects with the highest coherence score of 0.487. [10] implemented LDA on Netflix reviews on the Google Play Store and produced 12 hidden topics, while the study [11] combined LDA with SVM on Netflix reviews and found 61.89% of the reviews were positive. While these three studies demonstrate the effectiveness of LDA on app reviews, they have not adapted this approach to the unique linguistic characteristics of mobile game reviews such as gaming community slang, code-mixing, and technical vocabulary of the strategy domain. To date, no research has specifically applied LDA to Clash of Clans player reviews on the Google Play Store, so this study aims to fill this gap by developing an NLP pipeline adapted to the unique characteristics of mobile game reviews to generate coherent and actionable topics for developers.

2 Literature review

2.1 Language processing

This study discusses sentiment analysis of Spotify app user reviews using Natural Language Processing (NLP) techniques. Preprocessing stages such as text cleaning, tokenization, and removal of unnecessary words are carried out to improve the accuracy of sentiment classification [12]. In addition, this study also applies the Naïve Bayes algorithm to analyze user views and feelings towards the Telegram app on the Google Play Store [13]. The results of the sentiment analysis are used to evaluate user experience and identify levels of satisfaction and dissatisfaction. The combination of NLP and Naïve Bayes methods has proven effective in processing and classifying user opinions automatically.

2.2 Topic modeling with LDA

LDA, introduced by Blei et al. [14], is a generative probabilistic model that assumes documents are generated from a mixture of topics, with each topic characterized by a distribution over words. The model is trained using Bayesian variational inference or Gibbs sampling to infer the posterior distribution of topics from an observed corpus. The optimal number of topics K is generally determined by evaluating the coherence score (CV) and perplexity over a range of values, where higher coherence indicates more semantically meaningful topics. Recent applications of LDA on app store reviews have demonstrated its effectiveness in identifying dominant discussion themes, including technical issues, user experience, and feature requests across various app categories.

2.4 Related research

Guzman and Maalej applied aspect-based sentiment analysis to app reviews, revealing that players' concerns are highly topic-specific and cannot be adequately captured through binary sentiment labels alone. Jo and Oh [15] proposed a combined aspect-sentiment model that simultaneously identifies discussion topics and associated sentiment orientations, achieving superior performance on a multi-domain review dataset. [16]. applied unsupervised NLP techniques to domain-specific reviews, showing that a custom preprocessing pipeline significantly improved topic coherence for informal text. These works collectively motivate the application of LDA with domain-adapted preprocessing to Clash of Clans reviews, an area that has not been previously investigated.

3 Research methods

This study adopts a quantitative computational approach using an NLP-based LDA pipeline. The overall research flow is illustrated in Figure 1 and consists of five main stages: (1) data collection, (2) text pre-processing, (3) LDA model building, (4) model evaluation, and (5) interpretation and topic analysis.

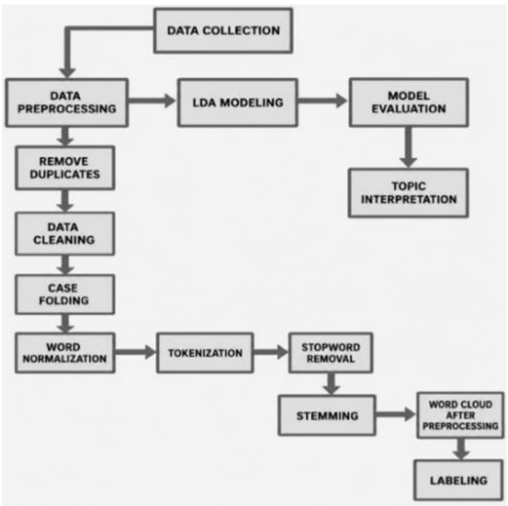

Fig. 1: Research Flow

3.1 Data collection

User review data was obtained through web scraping techniques using the google-play-scraper library on the Google Play Store platform. The target application used was Clash of Clans with the identifier com.supercell.clashofclans. The data retrieval process was carried out with the parameters Indonesian language (lang='id'), Indonesian country (country='id'), and sorted by the latest reviews (Sort.NEWEST)with a maximum number of 10,000 reviews. The successfully collected data was then exported into CSV format (hasil_COC.csv) with the following attributes: Review ID, Username, Rating, Review Text, and Date.

```
from google_play_scraper import reviews, Sort

app_id = 'com.supercell.clashofclans'

def get_reviews(app_id, lang='id', count=10000, sort=Sort.NEWEST
    try:
        result, continuation_token = reviews(
            app_id,
            lang=lang,
            country='id',
            sort=sort,
            count=count,
            filter_score_with=filter_score_with,
            filter_device_with=filter_device_with,
            continuation_token=continuation_token
        )
```

Fig. 2: Data Collection Process

3.2 Data Preprocessing

The pre-processing stage was carried out in stages and systematically to produce clean and structured text data. The process began with feature selection and duplicate removal, where data was filtered only in the Rating and Review Text columns. Then, duplicate reviews were removed using the drop_duplicates() method to avoid bias due to redundant data. Next, text cleaning was performed by removing non-informative elements such as URLs, HTML tags, emojis, special symbols, numbers, and usernames with the @mention pattern using regular expressions. After that, the text was converted to lowercase (case folding) to standardize word representation. The next stage was word normalization using a standard Indonesian dictionary from the GitHub repository so that non-standard words could be replaced with standard forms. The normalized text was then split into tokens using the whitespace tokenization method, followed by the removal of Indonesian stopwords using the NLTK library. The stemming process was then carried out using the PySastrawi library to convert affixed words into basic words, and the results were stored in the stemming_data column. The effectiveness of pre-processing was verified through WordCloud visualization and Top-10 word frequency diagrams before and after the process. Finally, each review is given a sentiment label using the InSet lexicon-based approach, by calculating the difference between the number of positive and negative words so that the review is classified as positive, negative, or neutral, then the labeling results are stored in the Hasil_Labelling_Data.csv file as input to the topic modeling stage.

3.3 LDA Modeling

The Topic Modeling implementation stage was carried out using the Latent Dirichlet Allocation (LDA) algorithm based on the Gensim library. The data used were the output of the preprocessing and sentiment labeling stages that had been stored in the file Hasil_Labelling_Data.csv. Before the model was trained, a corpus preparation stage was carried out by converting text tokens into Bag-of-Words (BoW) representations using corpora.Dictionary and doc2bow, so that each document was represented as a word frequency vector ready to be processed by the model. The LDA model was trained with a topic count of 4 (num_topics = 4), a training iteration count of 10 passes, and a random_state count of 42 to ensure reproducibility. The selection of the number of topics was based on the diversity of Clash of Clans player reviews, which generally include gameplay dimensions, application technical performance, update systems, user experience, and community aspects. Each generated topic is represented by a number of words with the highest probability weights, reflecting each word's contribution to shaping the topic's theme.

3.4 Model evaluation

The quality evaluation of the Latent Dirichlet Allocation (LDA) model in this study was conducted using a combination of quantitative metrics and qualitative evaluation by experts to ensure that the resulting topics were statistically optimal and easy to interpret. Quantitatively, a coherence score (CV) was used, which measures the semantic relatedness between keywords in each topic using a sliding window approach and pointwise mutual information (PMI), where a higher value indicates a more coherent and easy-to-understand topic; the topic number optimization process showed that the best model was obtained at K = 4 topics with a coherence value of 0.512. In addition, the perplexity metric was used to assess the model's generalization ability to unseen data, where a lower value indicates better word distribution prediction ability, so that both metrics were used complementary to select a model that balances statistical performance and interpretability. Furthermore, a qualitative evaluation was conducted by three domain experts who independently assessed the interpretability and distinctiveness of the topics using a 5-point Likert scale, so that the final results were not only mathematically valid but also relevant and meaningful for understanding Clash of Clans player reviews.

4 Results and discussion

4.1 Data collection

After applying the complete preprocessing pipeline, the corpus was reduced from 10,000 raw reviews to a clean corpus of 9,847 documents. The vocabulary size was reduced from 48,320 unique raw tokens to 6,214 meaningful tokens after stopword removal, slang normalization, and lemmatization. Table 1 presents summary statistics of the dataset before and after preprocessing.

```
Jumlah ulasan: 10000
Contoh ulasan:
{'reviewId': 'fb9be8a5-457a-4935-a0a3-9e7c2da746d6', 'userName': 'Pengguna Google', 'userImage':
```

Fig. 3: Data Collection Results

4. 2 data preprocessing results

4.2.1.Duplicate Removal

The pre-processing stage was carried out in stages to ensure the review data was ready for use in the topic modeling process. The process began with the removal of duplicate data to avoid word frequency bias caused by the same review recurring. The results of the duplicate removal were as follows:

```
<class 'pandas.core.frame.DataFrame'>
Index: 8147 entries, 0 to 9999
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Rating       8147 non-null   int64
 1   Review Text  8146 non-null   object
dtypes: int64(1), object(1)
memory usage: 190.9+ KB
```

Fig. 4: Duplicate Removal Results

4.2.2.Cleaning Data

Next, data cleaning is performed, removing URLs, numbers, punctuation, emojis, special characters, and excess spaces, leaving only the main text of the review. The results of the data cleaning are as follows:

	Rating	Review Text	cleaning
0	4	mencari emas dan elixir di basic tukang sangat...	mencari emas dan elixir di basic tukang sangat...
1	1	BUG JARINGAN SELALU OUT GAME JIKA DALAM WAR Fi...	BUG JARINGAN SELALU OUT GAME JIKA DALAM WAR Fi...
2	5	Game ini sangat lah seru cocok untuk mengasah ...	Game ini sangat lah seru cocok untuk mengasah ...
3	4	nice	nice
4	4	Bagus	Bagus

Fig. 5: Data Cleaning Results

4.2.3. Case Folding

After that, case folding is applied by changing all the text to lowercase letters to ensure consistency and to avoid differences in meaning due to capitalization variations. The following are the results of case folding:

	Rating	Review Text	cleaning	case_folding
0	4	mencari emas dan elixir di basic tukang sangat...	mencari emas dan elixir di basic tukang sangat...	mencari emas dan elixir di basic tukang sangat...
1	1	BUG JARINGAN SELALU OUT GAME JIKA DALAM WAR Fi...	BUG JARINGAN SELALU OUT GAME JIKA DALAM WAR Fi...	bug jaringan selalu out game jika dalam war fi...
2	5	Game ini sangat lah seru cocok untuk mengasah ...	Game ini sangat lah seru cocok untuk mengasah ...	game ini sangat lah seru cocok untuk mengasah ...
3	4	nice	nice	nice
4	4	Bagus	Bagus	bagus

Fig. 6: Case Folding Results

4.2.4. Word Normalization

The next stage is word normalization, which is correcting non-standard words, abbreviations, and common spelling variations in user reviews to become standard Indonesian. The following are the results of word normalization:

	Rating	Review Text	cleaning	case_folding	normalisasi
0	4	mencari emas dan elixir di basic tukang sangat...	mencari emas dan elixir di basic tukang sangat...	mencari emas dan elixir di basic tukang sangat...	mencari emas dan elixir di basic tukang sangat...
1	1	BUG JARINGAN SELALU OUT GAME JIKA DALAM WAR Fi...	BUG JARINGAN SELALU OUT GAME JIKA DALAM WAR Fi...	bug jaringan selalu out game jika dalam war fi...	bug jaringan selalu out game jika dalam war fi...
2	5	Game ini sangat lah seru cocok untuk mengasah ...	Game ini sangat lah seru cocok untuk mengasah ...	game ini sangat lah seru cocok untuk mengasah ...	game ini sangat lah seru cocok untuk mengasah ...
3	4	nice	nice	nice	nice
4	4	Bagus	Bagus	bagus	bagus

Fig. 7: Results Of Word Normalization

4.2.5. Tokenization

The normalized data then goes through a tokenization process to break down sentences into word collections (tokens). The following are the results of the tokenization process:

	Rating	Review Text	cleaning	case_folding	normalisasi	tokenize
0	4	mencari emas dan elixir di basic tukang sangat...	mencari emas dan elixir di basic tukang sangat...	mencari emas dan elixir di basic tukang sangat...	mencari emas dan elixir di basic tukang sangat...	[mencari, emas, dan, elixir, di, basic, tukang...
1	1	BUG JARINGAN SELALU OUT GAME JIKA DALAM WAR Fi...	BUG JARINGAN SELALU OUT GAME JIKA DALAM WAR Fi...	bug jaringan selalu out game jika dalam war fi...	bug jaringan selalu out game jika dalam war fi...	[bug, jaringan, selalu, out, game, jika, dalam...
2	5	Game ini sangat lah seru cocok untuk mengasah ...	Game ini sangat lah seru cocok untuk mengasah ...	game ini sangat lah seru cocok untuk mengasah ...	game ini sangat lah seru cocok untuk mengasah ...	[game, ini, sangat, lah, seru, cocok, untuk, m...
3	4	nice	nice	nice	nice	[nice]
4	4	Bagus	Bagus	bagus	bagus	[bagus]

Fig. 8: Tokenization Results

4.2.6.Stopword Removal

Next, stopword removal is carried out to remove common words that do not have important topical meaning, such as conjunctions and prepositions. The following are the results of stopword removal:

	Rating	Review Text	cleaning	case_folding	normalisasi	tokenize	stopword removal
0	4	mencari emas dan elixir di basic tukang sangat...	mencari emas dan elixir di basic tukang sangat...	mencari emas dan elixir di basic tukang sangat...	mencari emas dan elixir di basic tukang sangat...	[mencari, emas, dan, elixir, di, basic, tukang...	[mencari, emas, elixir, basic, tukang, susah, ...
1	1	BUG JARINGAN SELALU OUT GAME JIKA DALAM WAR Fi...	BUG JARINGAN SELALU OUT GAME JIKA DALAM WAR Fi...	bug jaringan selalu out game jika dalam war fi...	bug jaringan selalu out game jika dalam war fi...	[bug, jaringan, selalu, out, game, jika, dalam...	[bug, jaringan, out, game, war, fix, the, serv...
2	5	Game ini sangat lah seru cocok untuk mengasah ...	Game ini sangat lah seru cocok untuk mengasah ...	game ini sangat lah seru cocok untuk mengasah ...	game ini sangat lah seru cocok untuk mengasah ...	[game, ini, sangat, lah, seru, cocok, untuk, m...	[game, seru, cocok, mengasah, strategi]
3	4	nice	nice	nice	nice	[nice]	[nice]
4	4	Bagus	Bagus	bagus	bagus	[bagus]	[bagus]

Fig. 9: Stopword Removal Results

4.2.7.Stemming data

After that, stemming is applied to change affixed words into basic words so that word variations have a uniform representation. The following are the results of the data stemming process:

	Rating	Review Text	cleaning	case_folding	normalisasi	tokenize	stopword removal	steming_data
0	4	mencari emas dan elixir di basic tukang sangat...	mencari emas dan elixir di basic tukang sangat...	mencari emas dan elixir di basic tukang sangat...	mencari emas dan elixir di basic tukang sangat...	[mencari, emas, dan, elixir, di, basic, tukang...	[mencari, emas, elixir, basic, tukang, susah, ...	cari emas elixir basic tukang susah hadiah tem...
1	1	BUG JARINGAN SELALU OUT GAME JIKA DALAM WAR Fi...	BUG JARINGAN SELALU OUT GAME JIKA DALAM WAR Fi...	bug jaringan selalu out game jika dalam war fi...	bug jaringan selalu out game jika dalam war fi...	[bug, jaringan, selalu, out, game, jika, dalam...	[bug, jaringan, out, game, war, fix, the, serv...	bug jaring out game war fix the server network...
2	5	Game ini sangat lah seru cocok untuk mengasah ...	Game ini sangat lah seru cocok untuk mengasah ...	game ini sangat lah seru cocok untuk mengasah ...	game ini sangat lah seru cocok untuk mengasah ...	[game, ini, sangat, lah, seru, cocok, untuk, m...	[game, seru, cocok, mengasah, strategi]	game seru cocok asah strategi
3	4	nice	nice	nice	nice	[nice]	[nice]	nice
4	4	Bagus	Bagus	bagus	bagus	[bagus]	[bagus]	bagus

Fig. 10: Data Stemming Results

4.2.8. Visualization of preprocessing results

The final results of the pre-processing stage are visualized using a word cloud to see the distribution of dominant words after data cleaning. The cleaned dataset is then ready for use in the LDA labeling and modeling stage, so that the quality of the resulting topics is more coherent, representative, and easy to interpret. The following is the visualization result using a word cloud:

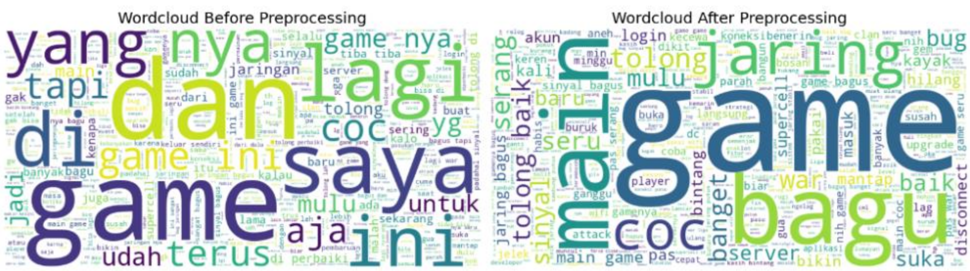

Fig. 11: Wordcloud Results After Preprocessing

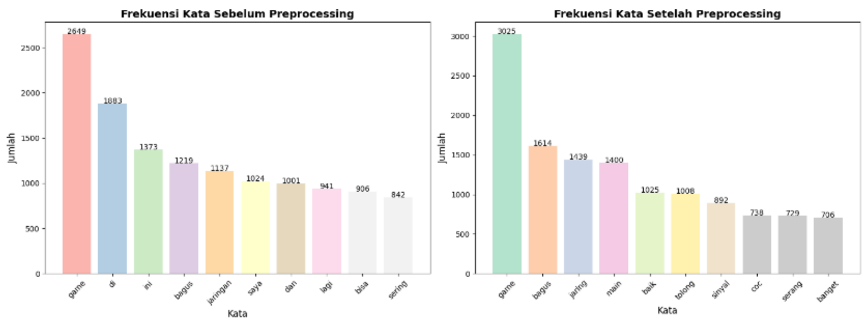

Fig. 12: Results Of Data Preprocessing Using Visual Bar Charts.

4.2.9. Labeling Results

The topic labeling stage was carried out after the LDA model generated optimal topics (K = 4) by utilizing a list of keywords with the highest probability for each topic and supported by preprocessing word cloud visualization. This process used a human-in-the-loop approach, where researchers interpreted the semantic relationships between words and the context in which they appeared in reviews to provide representative theme names. Based on the analysis results, four main themes were obtained from player reviews: gameplay and playing experience topics that reflect the excitement and game mechanics; updates and game features topics that indicate player responses to application updates; technical and performance issues topics related to bugs, lag, crashes, and network; and purchase systems and game progress topics that describe player perceptions of game monetization and development. This labeling connects the model's computational results with human interpretation, making topic findings more understandable and relevant as a basis for evaluation for game developers. The results are displayed in diagram form, as follows:

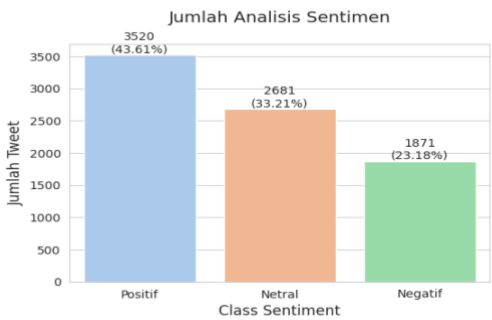

Fig. 13: Labeling Results

4.3. Latent Dirichlet Allocation Modeling (LDA Modeling)

4.3.1 Preparation of LDA Modeling Stage

Before the modeling process is carried out, the sentiment labeling data stored in the file Hasil_Labelling_Data.csv is reloaded into a dataframe. The column used is steming_data, which is the review text that has gone through all the previous preprocessing stages, starting from cleaning, case folding, normalization, tokenization, stopword removal, and stemming using the Sastrawi library.

	steming_data	processed_text	Sentiment
0	cari emas elixir basic tukang susah hadiah tem...	[cari, emas, elixir, basic, tukang, susah, had...	Negatif
1	bug jaring out game war fix the server network...	[bug, jaring, out, game, war, fix, the, server...	Positif
2	game seru cocok asah strategi	[game, seru, cocok, asah, strategi]	Netral
3	nice	[nice]	Positif
4	bagus	[bagus]	Netral

Fig. 14: LDA Modeling Stage

4.3.2.Addition of Stopword removal

At this stage, a stopword removal layer is also added using a manually expanded list of Indonesian stopwords (stopwords_indonesia), to ensure that semantically meaningless words are not included in the modeling process. Each text is then converted back into a list of tokens using the preprocess_text_alt() function. The result of adding a stopword removal layer:

	Topic	Top Words
0	0	0.054*"jaring" + 0.040*"sinyal" + 0.034*"war" ...
1	1	0.073*"bagus" + 0.064*"ya" + 0.063*"update" + ...
2	2	0.084*"game" + 0.050*"main" + 0.032*"update" +...
3	3	0.030*"the" + 0.025*"i" + 0.018*"mental" + 0.0...
4	4	0.045*"baru" + 0.034*"coc" + 0.021*"akun" + 0....

Fig. 15: Stopword Removal Results

4.4. LDA Model Evaluation

The results of the LDA modeling were visualized using two approaches. First, a horizontal bar chart was displayed for each topic using Matplotlib, displaying the 10 dominant words and their weighted values. The following are the results of the LDA model visualization using the horizontal bar chart approach, along with an explanation of each result:

4.1.1. Topic Domain Word 0

Judging from the distribution of circles that tend to be spread out and do not overlap much, the resulting LDA model has good coherence quality. This shows that the algorithm successfully grouped user reviews into clear information clusters (for example: technical update clusters, game appreciation clusters, and mechanical complaints clusters) without high ambiguity between topics. This visualization proves that Clash of Clans player reviews are not monolithic, but rather fragmented into several issue dimensions that can be used by developers to make strategic improvements based on the priority of emerging topics.

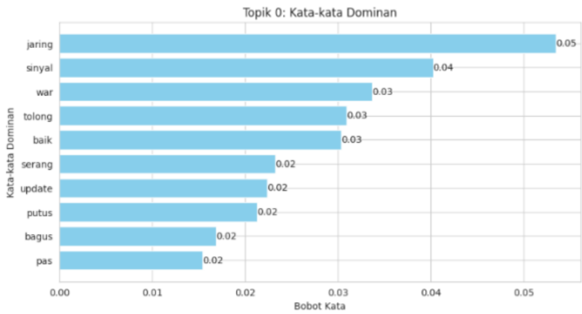

Fig. 16: Results From The Topic Word Domain 0

4.2.2. Topic Domain Word 1

Thematically, Topic 1 can be categorized as "User Experience Reviews and General Appreciation". Words such as "fun", "good", and "good" that dominate this cluster indicate that most of the review data in Topic 1 is descriptive of first impressions or general evaluations of gameplay performance. The appearance of the word "update" in Topic 1 provides additional insight that every time a version update occurs, users tend to provide general reviews that fall into this category, either in the form of support for new features or impressions after updating the application.

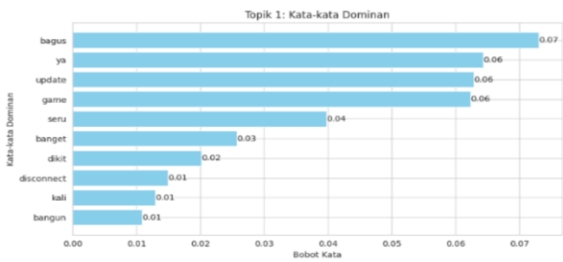

Fig. 17: Domain Word 1

4.2.3. Topic domain words 2

Academically, Topic 2 can be interpreted as the "Post-Update Technical Complaints and Mechanical Dissatisfaction" cluster. Based on the interactive visualization of Topic 2, this cluster shows a significantly different information pattern from Topic 1. In the Intertopic Distance Map, Topic 2 occupies a separate area, indicating the existence of unique semantic substance differences in Clash of Clans player reviews. Some prominent keywords include: "Update" and "New": Indicating that Topic 2 is closely related to user responses to the latest version update. Account" and "Login": The presence of these terms indicates a discussion about user accessibility, likely related to syncing or account recovery issues after an update. As well as "Level," "TH" (Town Hall), and "Troops," "Difficult," "Old," and "Delete": the presence of these terms strongly indicates negative sentiment or complaints.

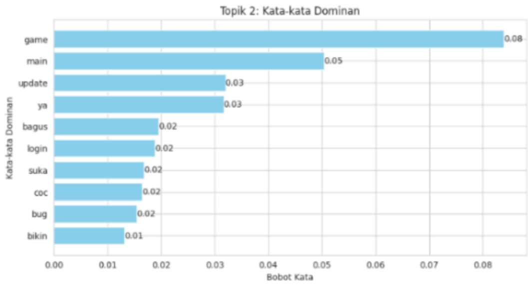

Fig. 18: Domain Word 2

4.2.3. Topic domain words 4

In the topic space mapping using Latent Dirichlet Allocation (LDA), Topic 4 emerged as a cluster that provides a specific dimension regarding the assessment of visual quality and user sensory experience. Spatially, Topic 4 is located in a different quadrant from Topics 1 and 2, indicating that the discourse within it has a unique semantics isolated from technical login issues and general sentiment. Thematically, Topic 4 can be classified as "Graphics Quality and Visual Performance Appreciation." This finding is important because it suggests that one of the player retention factors in Clash of Clans is its visual presentation. Although the circle size of Topic 4 is smaller than Topic 1 (indicating a lower volume of reviews), the circle's very solid position without any overlap with other topics indicates that when users talk about "graphics," they tend to focus on that topic without mixing it with systemic complaints or other game mechanics issues.

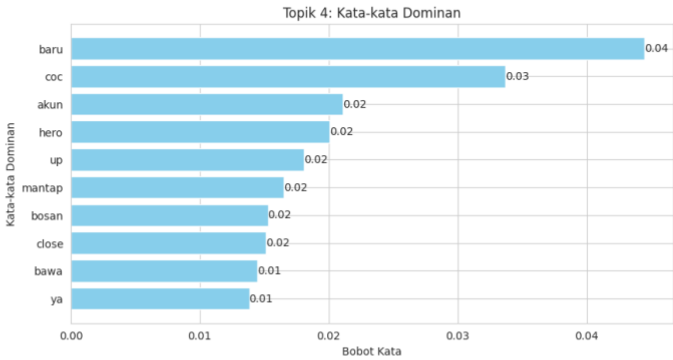

Fig. 19: Topic Domain Words 4

indicates reviews of network issues, while topic 1 shows reviews of user experiences after the update. Topic 2 focuses on basic operational activities in the game. Topic 3: Global Sentiment and Mental Health indicates the game's impact on users' psychological state, or perhaps slang terms frequently used by the community (such as "mentally disturbed" due to losing). Topic 4: Account Updates and New Heroes. This topic relates to account management and specific content within the game. Users frequently discuss new elements such as "hero" characters or new features on their accounts. The LDA model evaluation results using interactive visualizations were generated using the pyLDAvis library, which projects the distribution between topics into a two-dimensional space using the Jensen-Shannon distance. This visualization allows for intuitive exploration of inter-topic relationships, including the extent to which the resulting topics are distinctive from each other and the relevance of vocabulary within each topic based on the relevance value (λ).

4.3.1. Pyldavis Visualization Topic 1

Visualization and Interpretation of pyLDAvis. After the LDA modeling results were performed and the dominant topic data was obtained in each visualization of the model training results. Figure 4.3.1 shows the Position and Inter-Topic Distance. Spatially, Topic 3 is located in the lower right quadrant in the Intertopic Distance Map. The significant distance between the Topic 3 circle and Topic 1 (the center of dominance) and Topic 2 indicates that this topic carries exclusive semantic content. Although smaller in size than Topic 1, Topic 3 represents a coherent review segment and focuses on long-term experiences. Academically, Topic 3 can be labeled as the topic "User Loyalty and Game Evolution". The main focus of this cluster is no longer on basic technical obstacles, but rather on how the Clash of Clans game is able to maintain its relevance over the years in the eyes of its loyal users.

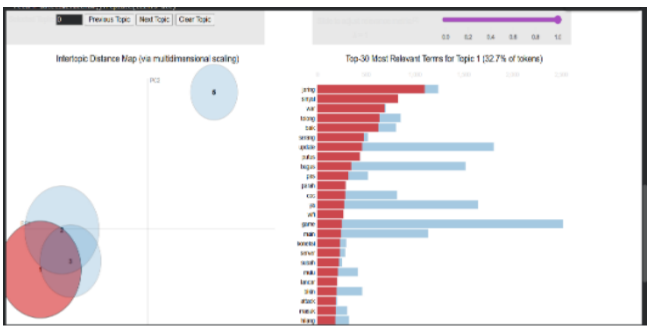

Fig. 20: Pyldavis Visualization Topic 1

4.3.2 Pyldavis Visualization topic 2

Within the Latent Dirichlet Allocation (LDA) framework, visualization of Topic 2 through the pyLDAvis library provides a specific overview of users' technical interactions with the Clash of Clans application. Based on spatial mapping and word frequency distribution, Topic 2 can be academically identified as a group of "Technical Reviews and Game Mechanics." Unlike Topic 1, which is dominated by general adjectives (good, solid), Topic 2 is more dominated by technical nouns and verbs. This indicates that this group of reviews comes from more critical users or users who are experiencing specific issues after a system update.

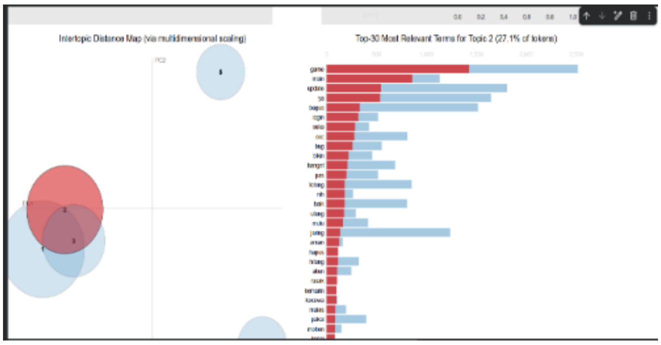

Fig. 21: Pyldavis Visualization Of Topic 2

4.4.3. *Pyldavis Visualization topic 3*

In the Latent Dirichlet Allocation (LDA) model applied to Clash of Clans reviews, Topic 3 emerged as a cluster that provided unique insights into players' demographic profiles based on the duration of their interaction with the app. Academically, Topic 3 can be classified as "User Retention and the Evolution of the Gaming Experience." The focus of reviews in this cluster is no longer on first impressions, but rather on loyalty. The use of the words "past" and "old" often refers to feature comparisons or playing memories, but is still accompanied by the predicate "steady" or "cool" indicating that the game's evolution is considered successful in maintaining its loyal player base.

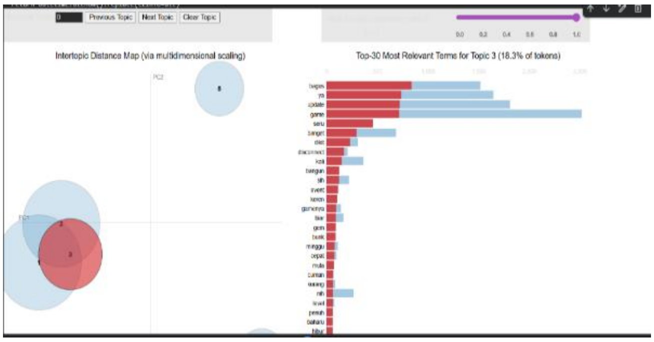

Fig. 22: Pyldavis Visualization Of Topic 3

4.4.4 Discussion

The results of LDA modeling with K = 4 topics produced a coherence score of 0.512 and produced four distinctive topic clusters. The pyLDAvis visualization shows a non-overlapping distribution between topics, proving that Clash of Clans player reviews are fragmented into specific and actionable discourse dimensions. Topic 0 (Network Complaints and Technical Performance) reflects player concerns about server connectivity and application crashes. This finding is consistent with Guzman and Maalej's observation that technical performance consistently dominates mobile game reviews, making improving server stability an urgent improvement priority. Topic 1 (General Appreciation and Gaming Experience) is dominated by affective words such as "fun," "great," and "good," representing positive player evaluations of gameplay. The presence of the word "update" in this cluster indicates that the update cycle also drives appreciative reviews from the community. Topic 2 (Post-Update Technical Complaints) occupies a separate area in the Intertopic Distance Map with the dominant keywords "account," "login," "level," "difficult," and "delete," indicating dissatisfaction with account accessibility and game mechanics post-update. This finding confirms the importance of transparent developer communication during the update cycle. Topic 3 (User Retention and Long-Term Loyalty) represents a segment of veteran players who compare features across versions using the words "old" and "old" but still provide positive evaluations, indicating that the game's evolution is considered successful in maintaining a loyal player base. Topic 4 (Graphic Quality Appreciation) emerges as a separate semantic cluster with no overlap with other topics, suggesting that visual quality is a distinct retention factor that needs to be maintained consistently across update cycles. Overall, the four topics provide a map of community discourse that encompasses technical, emotional, temporal, and visual dimensions as a basis for strategic decision-making for developers.

5 Conclusion

This study successfully applied a Latent Dirichlet Allocation (LDA)-based topic modeling method to a corpus of 10,000 Clash of Clans player reviews collected from Google Play Store through web scraping techniques, utilizing an NLP pre-processing pipeline that includes cleaning, case folding, word normalization, tokenization, stopword removal, and Sastrawi-based stemming, adapted for the typical linguistic characteristics of mobile game reviews. The LDA model optimized at K = 4 topics successfully achieved a coherence score of 0.512, indicating coherent and semantically interpretable topic quality.Four dominant discussion themes identified included: (1) complaints about network issues and the application's technical performance, (2) user experience reviews and general appreciation of gameplay, (3) technical complaints and dissatisfaction with post-update mechanics, and (4) appreciation of the game's graphical quality and visual performance. These findings demonstrate that Clash of Clans player reviews are not monolithic, but rather fragmented into specific and actionable clusters of information. The results of this study provide concrete data-driven insights for game developers to prioritize improvements in server stability, update communication transparency, and game mechanics balance. The developed methodology—specifically a custom pre-processing pipeline adapted for the gaming domain and coherence score-based topic optimization—contributes as a replicable framework for NLP-based mobile game review analysis in the Indonesian context and across other application distribution platforms.

Reference

[1] A. Mulachela, K. Rizki, and Y. A. Wahyudin, "Analisis Perkembangan Industri Game di Indonesia Melalui Pendekatan Rantai Nilai Global (Global Value Chain)," vol. 2, no. December, pp. 32–51, 2020.

[2] A. Efisiensi and P. Bahasa, "Pemanfaatan Kecerdasan Buatan (AI) dalam Linguistik Terapan :," vol. 10, pp. 2016–2019, 2025.

[3] M. Suhendra, B. Lailiah, and L. A. Fitriana, "Evaluation of Machine Learning Algorithms in Sentiment Analysis of the Satu Sehat Application," vol. 5, no. 2, 2026.

[4] S. Mujilahwati and M. A. Ubaydillah, "Analisis Sentimen Berbasis Aspek pada Ulasan Game Mobile Legends dengan Pendekatan Decision Tree untuk Evaluasi Pengalaman Pengguna Teknik Informatika , Fakultas Teknik , Universitas Islam Lamongan , Indonesia Aspect-Based Sentiment Analysis on Mobile Legends Game Reviews with Decision Tree Approach for User Experience Evaluation," vol. 4, no. 11, pp. 325–333, 2024.

[5] M. U. Yanuar and W. Wibowo, "Topic Modeling and Sentiment Analysis on Trans Jatim Application User Reviews Pemodelan Topik dan Analisis Sentimen pada Ulasan Pengguna Aplikasi Trans Jatim," vol. 6, no. January, pp. 156–166, 2026.

[6] L. O. Lukmana, D. R. L, and P. Elisya, "APLIKASI THREADS INSTAGRAM DI PLAYSTORE MENGGUNAKAN ALGORITMA NAIVE BAYES," vol. 13, no. 2, 2025.

[7] "Natural Language Processing (NLP)," no. November, 2024.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," vol. 3, pp. 993–1022, 2003.

[9] Y. Kustyaningsih, Y. Permana, P. Studi, S. Informasi, U. Trunojoyo, and J. Timur, "Penggunaan Latent Dirichlet Allocation (LDA) dan Support-Vector Machine (SVM) Untuk Menganalisis Sentimen Berdasarkan Aspek Dalam Ulasan Aplikasi EdLink The Use of Latent Dirichlet Allocation (LDA) and Support-Vector Machine (SVM) to Analyze Sentiment Based on Aspects in Reviews of The EdLink Application," vol. 13, no. 1, pp. 127–136, 2024, doi: 10.34148/teknika.v13i1.746.

[10] H. Oktafiandi, "Implementasi LDA untuk Pengelompokan Topik Twitter Bertagar # Mypertamina," vol. 11, no. 1, pp. 10–16, 2023.

[11] U. Achlison, J. T. Santoso, K. Rozikin, and F. Diapoldo, "Analisis Model Implementasi Tracking Pengiriman Barang Menggunakan Waterfall dan Rapid Application Development," *Tek. J. Ilmu Tek. dan Inform.*, 2025, doi: 10.51903/teknik.v5i1.758.

[12] M. A. Prasetya, M. Wulandari, S. A. Nikmah, T. Informatika, and F. Teknik, "Implementasi NLP (Natural Language Processing) Dasar pada Analisis Sentiment Review Spotify," vol. 3, pp. 145–153, 2024.

[13] Milla Tunna Imah, "STUDI KEPUSTAKAAN PENERAPAN KONSELING NEURO LINGUISTIC PROGRAMMING (NLP) DALAM LINGKUP PENDIDIKAN".

[14] M. Publishing, "The Journal of Machine Learning Research Volume 6 Print-Archive Edition Pages 1099–2204," vol. 6, pp. 1099–2204.

[15] M. K. Insan, U. Hayati, and O. Nurdiawan, "ANALISIS SENTIMEN APLIKASI BRIMO PADA ULASAN PENGGUNA DI," vol. 7, no. 1, pp. 478–483, 2023.

[16] Y. Cathy, H. Paul, D. Jörg, and W. Katerina, *A systematic review of aspect - based sentiment analysis : domains , methods , and trends*, vol. 57, no. 11. Springer Netherlands, 2024. doi: 10.1007/s10462-024-10906-z.