**Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer**
Vol. 24, No. 2, March 2025, pp. 347∼358
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v24i2.4794


# Novel Application of K-Means Algorithm for Unique Sentiment Clustering in 2024 Korean Movie Reviews on TikTok Platform

**Baiq Rima Mozarita Erdiani, Ario Yudo Husodo, Ida Bagus Ketut Widiartha**
Universitas Mataram, Mataram, Indonesia

**Article Info**

***Article history:***

Received January 08, 2025
Revised February 05, 2025
Accepted March 15, 2025

***Keywords:***

*Clustering;*
*IndoBERT;*
*K-Means Algorithm;*
*Korean;*
*Sentiment Analysis;*
*Tiktok.*

***ABSTRACT***

In recent years, social media has become one of the main factors influencing public perception of films. As a rapidly growing video-sharing platform, TikTok plays a crucial role in shaping audience opinions through comments, short reviews, and user discussions. This phenomenon is increasingly relevant in the Korean film industry, attracting global attention with its diverse genres and engaging narratives. However, a deep understanding of how audiences respond to films based on genre remains limited, especially in the dynamic context of social media. Therefore, this study aims to analyze audience sentiment toward Korean films released in 2024 on TikTok, focusing on sentiment distribution across four main genres: comedy, romance, action, and fun stories. The research methodology includes data collection through web crawling on TikTok, followed by text preprocessing and feature extraction using IndoBERT. Sentiment classification uses SentimentIntensityAnalyzer to categorize comments into positive, negative, or neutral. Since the dataset consists of unlabeled text, K-Means clustering is employed to identify sentiment groupings, with validation using principal component analysis to ensure cluster quality. The findings indicate that the romance and comedy genres are predominantly associated with neutral sentiment, reaching 89.6% and 87.4%, respectively. In contrast, the action genre exhibits higher sentiment polarization, with 14.9% positive and 24.7% negative sentiment. The fun story genre shows a more evenly distributed sentiment pattern. The main challenges include determining the optimal number of clusters and addressing imbalanced sentiment distribution across genres. This study provides valuable insights for filmmakers and marketers to understand audience reactions on social media better, enabling more targeted promotional strategies. Additionally, it contributes to the literature on sentiment analysis in the film industry, emphasizing the importance of genre-specific audience reception patterns for future research.



***Corresponding Author:***

Baiq Rima Mozarita Erdiani, +6287713801758,
Department of Informatics Engineering, Faculty of Engineering,
Universitas Mataram, Mataram, Indonesia,
Email: baiqrimamozarita7@gmail.com.

How to Cite:
B. R. Erdiani, A. Husodo, and I. Ketut Widiartha, ”Novel Application of K-Means Algorithm for Unique Sentiment Clustering in 2024 Korean Movie Reviews on TikTok Platform”, *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol.24, No.2, pp. 347-358, March, 2025.

## 1. INTRODUCTION

In today's digital era, the Korean Wave (Hallyu) has become a global phenomenon affecting various cultural aspects, including the film industry. The success of South Korea's entertainment industry is inseparable from the government's support in reforming television and allocating resources to develop technology and popular culture [1]. Since then, Korean cinema has experienced rapid growth and has successfully penetrated the international market. Based on a report from the Korean Film Council (KOFIC), the Korean film industry has consistently produced films with more than 10 million viewers since 2012 [2]. Some films, such as Train to Busan (2016), Parasite (2019), and Exhuma (2024), have attracted global attention and sparked widespread discussion on various digital platforms.

TikTok's influence shapes public opinion on various subjects, including movies. The platform enables users to share reviews through short videos, enhanced by commenting and hashtagging features that promote discussions [3]. User comments typically express positive, negative, or neutral sentiments, making sentiment analysis a crucial tool for assessing public reception, particularly in the social media sphere. Several prior studies align with this research: Rahmadani and Chintya Tampubolon (2022) [4] conducted sentiment analysis of TikTok social media using the Naïve Bayes Classifier algorithm to identify and classify negative comments, aiming to foster more positive interactions on the platform. Their study achieved an accuracy of 80% in sentiment classification. Setiawan et al. (2023) [5] conducted sentiment analysis on Indonesian TikTok reviews using LSTM and IndoBERTweet algorithms to classify sentiments into negative, neutral, and positive categories. Their study found that IndoBERTweet outperformed LSTM, achieving an accuracy of 80%, while LSTM had an accuracy of 78%. This research highlights the effectiveness of deep learning-based NLP models in analyzing user sentiment on social media platforms. Apriani et al. (2024) [6] conducted a sentiment analysis study on using TikTok as a learning medium using the Naïve Bayes Classifier algorithm. The study aimed to evaluate user sentiment regarding TikTok as an educational tool by analyzing 176 collected data points. Their findings indicated that the model achieved an accuracy of 75.27%, with a precision of 80% and a recall value of 58.38%. These results provide insight into user perceptions of TikTok's role in education and offer recommendations for further educational content development on the platform. Finally, Jung et al. (2024) [7] conducted a study on the normalization of vaping on TikTok using a mixed-methods approach that combined computer vision, natural language processing (NLP), and qualitative thematic analysis. Their research identified five major themes in TikTok posts related to vaping: vape product marketing, TikTok influencers, general vaping, vape brands, and vaping cessation. Using the ResNet50 model, the study achieved an impressive $F_1$-score of 0.97 in classifying vaping-related content. The findings suggest that TikTok influencers subtly integrate vaping into popular culture and daily life, thereby normalizing vaping behaviors among youth.

Several previous studies have explored sentiment analysis in various contexts, including TikTok and movie reviews, using different methodologies. Rahmadani and Chintya Tampubolon (2022) [4] focused on sentiment analysis of TikTok comments using the Naïve Bayes Classifier to classify negative comments with 80% accuracy. Setiawan et al. (2023) [5] conducted sentiment analysis on Indonesian TikTok reviews, comparing LSTM and IndoBERTweet models, with IndoBERTweet outperforming LSTM with 80% accuracy. Apriani et al. (2024) [6] applied the Naïve Bayes Classifier to analyze user sentiment regarding TikTok as an educational tool, with 75.27% accuracy. Meanwhile, Jung et al. (2024) [7] examined vaping normalization on TikTok using a mixed-methods approach, integrating computer vision and natural language processing. Although these studies demonstrate the effectiveness of sentiment analysis in TikTok-related contexts, they are mostly limited to general social media sentiment, education, and specific topics such as vaping. In addition, previous studies on sentiment analysis in movies mostly use Twitter and IMDb data, utilizing LSTM for sentiment classification and BERT to enhance contextual understanding. Some studies have also integrated clustering techniques to identify audience reaction patterns. However, TikTok-based sentiment analysis for movies is still under-explored. This research will address the disruption by focusing on sentiment analysis of movie-related content on TikTok and utilizing the K-Means learning technique that will collaborate with IndoBERT to gain insight into broader audience perceptions that no one has ever studied. Recent research has demonstrated that transformer-based models, such as IndoBERT, exhibit superior contextual understanding in Indonesian sentiment analysis [8]. Studies have shown that IndoBERT enhances the accuracy of sentiment classification in Indonesian text, making it a valuable tool for natural language processing tasks. Additionally, the K-Means algorithm has proven effective in clustering unlabeled sentiment data. For instance, research applying K-Means to movie review sentiments on Weibo revealed its capability to distinguish dominant sentiment groups effectively. Comparative studies on clustering algorithms further indicate that K-Means delivers competitive performance in social media-based sentiment segmentation. However, despite these advancements, there remains a gap in understanding how IndoBERT and K-Means can be integrated to improve sentiment analysis in Indonesian-language datasets. Previous studies have primarily focused on their individual performance, yet limited research explores their combined effectiveness. This study aims to bridge this gap by evaluating the synergy between IndoBERT and K-Means in sentiment analysis. The findings contribute to the broader field of machine learning-based sentiment analysis by providing insights into model performance, potential enhancements, and implications for research and community applications, particularly in analyzing public opinion on social and political issues. By addressing the urgency of sentiment analysis in Indonesian, this research builds upon at least five state-of-the-art studies published between 2021 and 2025, emphasizing the need for more robust methodologies in this

field. The novelty of this study lies in its exploration of IndoBERT's deep contextual understanding in combination with K-Means clustering to improve sentiment classification, filling a critical research gap and offering practical applications for public discourse analysis and community decision-making.

## 2. RESEARCH METHOD

This study employs a quantitative research approach to examine audience sentiment analysis of the Korean movie "2024" on the TikTok platform, aiming to understand how audiences receive each movie genre. A quantitative approach is chosen for its ability to analyze large datasets and derive statistical insights, which is essential for capturing the diverse sentiments expressed in user comments. The study utilizes the K-means Clustering algorithm to group reviews based on sentiment patterns generated from audience comments. Additionally, the IndoBERT model extracts features from the review text, allowing for numerical representation that facilitates further processing and analysis. This combination of techniques enables a comprehensive understanding of audience sentiment across different movie genres, providing valuable insights into viewer preferences and reactions.

Sentiment analysis is performed using SentimentIntensityAnalyzer (VADER) which is a method that will be used as modeling in sentiment analysis and can determine the diversity that exists in the data through emotional intensity, according to the Lexicon data dictionary that is already available [9]. This process is useful for classifying reviews into three sentiment categories, namely positive, negative, and neutral. Through this approach, the research aims to identify audience preferences for the most desirable Korean movie genres in 2024 and understand the sentiment distribution within each genre. Figure 1 below is the flow of methodology used in this research.

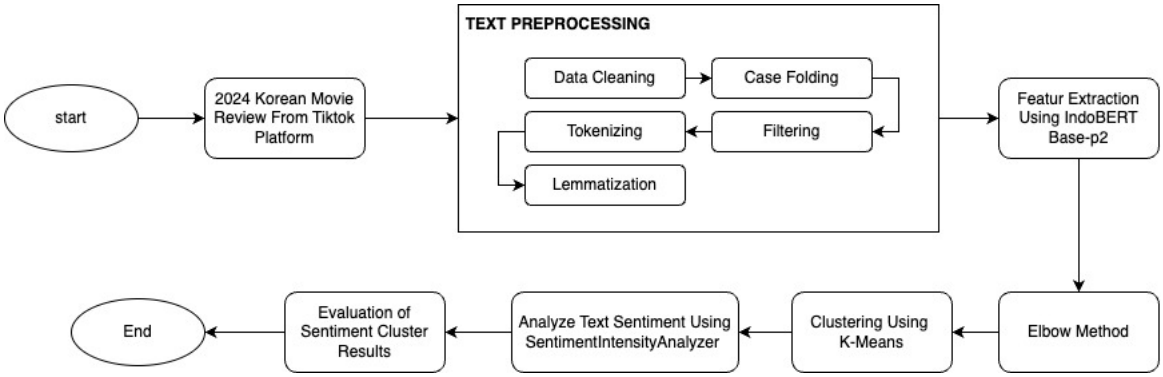


Figure 1. Project flow

In Figure 1, the methodology steps in text processing begin with data collection from Korean movie reviews in 2024 on the TikTok platform. After that, the data will be entered into the text preprocessing process. In the text preprocessing stage, the data goes through several stages: data cleaning, case folding, filtering, tokenizing, and finally lemmatization. Data that has gone through the preprocessing stage is represented in vector form (embedding) using the IndoBERT method. Furthermore, entering the elbow method stage is one of the methods used to determine the best number of clusters by looking at the percentage of each cluster that will form an elbow at a point [10]. The K-means algorithm is used to perform clustering at this stage. Clustering will divide or group unlabeled data into several clusters based on the dataset's analysis of similarity or dissimilarity to obtain a relationship with each other. This stage aims to provide unlabeled sentiment analysis based on clustering results using SentimentIntensityAnalyzer from the VADER (Valence Aware Dictionary and Sentiment Reasoner) library. At this stage, the results obtained will be visualized using a bar chart, and then after that, an analysis will be carried out.

### 2.1. Collect Data

The data collection method in this study employs a crawling technique to gather data from the TikTok platform, focusing on user-generated content related to the movie "2024." This approach allows for systematically collecting comments and reviews that meet specific research criteria. The research criteria are based on the movie genre, which is categorized into four distinct genres: comedy, romance, action, and thriller. The study aims to capture a diverse range of audience sentiments and reactions by targeting these genres. This comprehensive data collection process ensures that the analysis reflects the varied perspectives of viewers across different genres, providing a robust foundation for the subsequent sentiment analysis.

### 2.2. Data Preprocessing

Data preprocessing is the initial stage of data processing that aims to clean and prepare the text to make it easier to analyze and process. This stage is very important, as raw data often contains many irrelevant and unstructured elements that must be converted into a more organized format. The data cleaning process involves removing URLs, non-alphanumeric characters (except spaces), numbers, emojis, and initial or final spaces in the text, utilizing the Pandas Library. The removal of symbols and numbers is done because they do not have specific meanings that correlate with news topics [11]. After data cleaning, letter folding is performed to standardize the text by converting all characters to lowercase or uppercase using the Natural Language Toolkit (NLTK) library. Next, stopword removal removes irrelevant words from the text based on a predefined list of stopwords, which the NLTK library also facilitates. The tokenization process separates words in sentences into individual tokens, removing all symbols and characters, with separation based on spaces [12]. Finally, lemmatization is applied to parse inflected words into their base form, known as lemma, which improves the accuracy of text analysis by linking words with similar meanings [13]. This comprehensive preprocessing approach ensures the data is well structured and ready for further analysis.

### 2.3. Feature Extraction

IndoBERT is a word embedding model that has gone through a previous training process called pre-trained word embedding. It is trained using large and generalized datasets so that it can better understand the meaning and structure of syntax [14]. IndoBERT is usually implemented using the Hugging Face Transformer library, which is a popular library for transformer-based models, including BERT.

### 2.4. Clustering Using K-Means

K-Means is one of the non-hierarchical clustering procedures. K-means is a popular clustering method widely used in various fields because it is simple and easy to implement [15]. The K-means method is an unsupervised learning clustering method that aims to predict or classify objects so that the distance between objects to the data center in one data is minimal without having to do training first [16]. The steps involved in the K-Means algorithm are as follows [8]: First, the dataset is prepared. Next, the number of clusters is determined, and a random centroid point is selected for each cluster. Subsequently, the distance between each data point and the centroids is calculated to measure the proximity of the data to the centroids. Finally, data points are grouped into clusters based on their proximity to the centroids. The formula used to calculate the distance between two data points, xxx and yyy, is shown in Equation 1:

$$d(x, y) = \sum_{n+1}^{1} (x_i - y_i)^2 \tag{1}$$

Here, $d(x, y)$ represents the distance between the testing data $x$ and the training data $y$, where $x_i$ is the $i$-th testing data points and $y_i$ is the i-th training data point. This calculation is critical for determining the similarity of objects and plays a central role in grouping data during the clustering process. The iterative nature of the K-Means algorithm ensures efficient convergence towards well-defined clusters, making it a powerful tool for data analysis and segmentation.

### 2.5. Analyzed Text Sentiment Using SentimentIntensityAnalyzer

Sentiment analysis is a process that involves extracting, processing, and understanding data presented in unstructured text automatically [17]. This technique is essential for retrieving valuable information and sentiment from opinionated sentences, allowing researchers to gauge public sentiment on various topics. By analyzing the emotional tone behind the text, sentiment analysis can provide insights into how individuals feel about a particular subject, product, or service. The process typically employs natural language processing (NLP) techniques to identify and categorize sentiments as positive, negative, or neutral. As a result, sentiment analysis has become a vital tool in fields such as marketing, social media monitoring, and customer feedback analysis, enabling organizations to make data-driven decisions based on audience perceptions.

## 3. RESULT AND ANALYSIS

This section will discuss the results of the above stages, such as data preprocessing, feature extraction using IndoBERT, clustering using the K-Means algorithm, and analyzing sentiment that has undergone the clustering process. Each stage will be

written to see its analysis and effectiveness in processing, clustering, and understanding text data patterns to produce the desired sentiment.

### 3.1. The Result of Data Preprocessing

Data preprocessing is a process of preparing data before further processing. This is done before the classification process, which is needed to clean, remove, and change data sources in the form of non-alphabetic characters and other words that are not needed [18]. The data sources in this study were taken from reviews and reviewer comments on the TikTok platform regarding the Korean movie 2024. The data obtained includes various user reviews that provide their opinions in short texts. The amount of data obtained is 10,431 text data.

The application of data cleaning, such as removing URLs in the text, all non-alphanumeric characters except spaces, punctuation marks, and symbols, all numbers, emojis, and spaces at the beginning and end of the text, results in changes in the amount of each genre's data. Details of the data used, along with the amount and results of data cleaning, can be seen in Table 1.

Table 1. Data Cleaning Result

| Genre | Before Data Cleaning | After Data Cleaning |
|---|---|---|
| Comedy | 2.680 data | 2.306 data |
| Romance | 2.400 data | 2.399 data |
| Action | 2.863 data | 2.543 data |
| Thriller | 2.488 data | 1.844 data |
| Total | 10.431 data | 9.092 data |

After data cleaning, case folding is implemented to ensure the processed data has a uniform word form. This step involves converting all words into lowercase letters, which helps to eliminate inconsistencies caused by variations in capitalization. By standardizing the text in this manner, the analysis can focus on the actual content of the words rather than their formatting. This process is crucial for improving the accuracy of subsequent analyses, such as sentiment evaluation. The results of this case folding process can be seen in Table 2.

Table 2. Case Folding Result

| Case Folding Result | |
|---|---|
| Before Case Folding | After Case Folding |
| ”drama dan film song jong ki Bagus semua” | ”drama dan film song jong ki bagus semua” |
| ”Dia jadi mirip kim sejeong deh pas jadi cewe” | ”dia jadi mirip kim sejeong deh pas jadi cewe” |

In this research, the stopwords removal process utilizes the function ”stopwords.words('indonesian')," which provides a comprehensive list of common Indonesian words frequently used but with little informational value in text analysis. By eliminating these stopwords, the analysis can focus on the more meaningful words that contribute to the overall sentiment and context of the reviews. This step is essential for enhancing the quality of the data, as it reduces noise and allows for a clearer understanding of the audience's opinions. Removing stopwords helps streamline the text, making it easier to identify key themes and sentiments expressed by users. The details of this stopwords removal process, including the specific words eliminated, can be seen in Table 3.

Table 3. Stopwords Removal Result

| Stopwords Removal Result | |
|---|---|
| Before Stopwords Removal | After Stopwords Removal |
| ”siapapun namanya tetep dipanggil ikjun” | ”namanya tetep dipanggil ikjun” |
| ”dia jadi mirip kim sejeong deh pas jadi cewe” | ”kim sejeong deh pas cewe” |
| ”judulnya apa” | ”judulnya” |

The tokenizing process is the initial stage in text processing that aims to break the text into small units called tokens. These tokens are usually words, phrases, or symbols that have meaning in a particular context. The implementation of this process results in text data that has been converted into tokens, as shown in Figure 2. The tokenizing results provide a more organized structure, facilitating further analysis of the text data. This step also helps identify syntactic patterns or structures in the text. With tokenizing, data can be processed more efficiently in subsequent stages, such as lemmatization or sentiment analysis.

```
Padded Sequence 2389: [  86 4223  673 4224    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0]
Padded Sequence 2390: [ 613  233    6 4225 4226 4227  108    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0]
Padded Sequence 2391: [4228  787 4229  274   12  108    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0]
Padded Sequence 2392: [4230  296   86 4231    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0]
Padded Sequence 2393: [ 404 4232  107 4233    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0]
```

Figure 2. Tokenizing result

Lemmatization converts words in a text to their base form or lexical form. Unlike stemming, lemmatization considers the context and meaning of the words to make the results more accurate. For example, the word ”running” will be changed to ”run” because this lemma form reflects the word’s basic meaning. In this research, the lemmatization process is performed using the NLTK library, as shown in Figure 3. This process simplifies the data and reduces word variation without losing meaning. The results of lemmatization are essential to support further data analysis, including classification and theme clustering.

```
Teks setelah Lemmatization:
                                                FilteredText  \
0                          drama film song jong ki bagus
1  bagus sedih perjuanganny hidup dnegara orangtp...
2                                    bilang jelek bagus
3  nonton gatega paksain nonton endingnya puas ba...
4        kalo sengsara kelam berharap perlawanan gitu

                                               LemmatizedText
0                          drama film song jong ki bagus
1  bagus sedih perjuanganny hidup dnegara orangtp...
2                                    bilang jelek bagus
3  nonton gatega paksain nonton endingnya puas ba...
4        kalo sengsara kelam berharap perlawanan gitu
```

Figure 3. Lemmatization result

### 3.2. Feature Extraction using IndoBERT

This research uses the IndoBERT model to generate numerical representations (embeddings) of Indonesian text. The model used is ”indobenchmark/indobert-base-p2” which is used specifically for Indonesian data. Tokenization is done using AutoTokenizer by ensuring the text is processed using the tensor format with padding and truncation to maintain consistency up to 512 tokens. This feature extraction represents the dataset we have used. Below is an example of representing the word ”movie” after using through’s feature extraction stage using IndoBERT. Below is a representation of feature extraction using IndoBERT, as shown in Figure 4.

```
# Ambil index dari token 'film'
try:
    index = tokenized_sentence.index("film")
    vector_virus = hidden_states[0, index, :].numpy()  # Representasi vektor kata "film"

    # Tampilkan hasil vektor
    print(f"Representasi vektor untuk kata 'film':\n{vector_film}")
except ValueError:
    print("Kata 'film' tidak ditemukan dalam kalimat.")
```

```
-3.18702847e-01 -5.46175003e-01 -1.99399471e-01 -2.14635515e+00
-4.65782702e-01 -8.14047039e-01 -7.64952302e-01 -1.59581125e+00
 1.28801465e-01  2.08123946e+00  1.84933722e-01 -2.54699445e+00
-1.55581963e+00 -1.83135796e+00  5.29838979e-01 -3.32468910e-01
 1.27604163e+00 -1.25947583e+00 -8.89016390e-01 -3.66425186e-01
 6.03191316e-01  4.76940395e-03 -2.26167589e-01 -5.78757226e-02
 1.01957011e+00 -8.98676038e-01  2.24383974e+00  1.59585953e+00
-2.81430960e-01 -7.55633354e-01  3.61208946e-01 -1.19915545e-01
 1.02475178e+00 -1.27802482e-02  2.45484339e+00  1.53829920e+00
 7.93264568e-01 -1.83236405e-01  7.38887191e-01 -5.96539140e-01
-4.46351796e-01  4.93344575e-01 -4.68009740e-01 -6.86087847e-01
-5.19704819e-01 -2.15302873e-02  7.20675826e-01 -9.71922934e-01
 1.44500232e+00  1.59831911e-01  9.22320783e-02  1.08997369e+00
-1.91500664e+00  1.24939489e+00 -2.42600545e-01  3.98796916e-01
-1.81042209e-01 -8.19222152e-01  1.42772865e+00  5.57447910e-01
-1.60561979e+00  4.70930904e-01  1.84470403e+00  1.15034483e-01
 1.31995833e+00 -9.56762373e-01 -8.61102700e-01 -3.19002807e-01
 2.61742566e-02  2.90823847e-01 -1.78298557e+00  1.04938619e-01
-1.09380087e+00 -6.84143245e-01 -9.29608107e-01  1.89558196e+00
 3.39115858e-01  1.60285676e+00  7.19933271e-01 -9.18792546e-01
 4.05918568e-01 -5.10469139e-01 -2.12971520e+00 -3.84566873e-01
```

Figure 4. Vector representation of the word ”film”

Each word that has gone through the modeling stage using IndoBERT also has a similar weight to other words. If the similarity value is closer to 1, then the word has a high level of semantic similarity. As shown in Figure 5, the similarity representation results show the words most semantically similar to ”film.” This analysis is important for understanding the semantic relationships between words in the data corpus, especially in the context of natural language processing (NLP). Thus, the IndoBERT model cannot only understand sentence structure but also capture the meaning and context of words with better precision.

```
print("Kata-kata yang paling mirip dengan 'film':")
for word, similarity in most_similar:
    print(f"{word}: {similarity:.6f}")
```

```
Kata-kata yang paling mirip dengan 'film':
drama: 0.649925
song: 0.522231
[SEP]: 0.461903
bagus: 0.457911
[CLS]: 0.414321
```

Figure 5. Similarity representation of the word ”film”

### 3.3. Clustering Using K-Means

After performing feature extraction, the next step is clustering using the K-means algorithm. In clustering, the data entered is the feature extraction result data, and the K value is the cluster value. Before determining the K value, the elbow method is performed first to calculate the optimal K value used for creating the cluster. The following is the result of the elbow method, which can be seen in Figure 5.

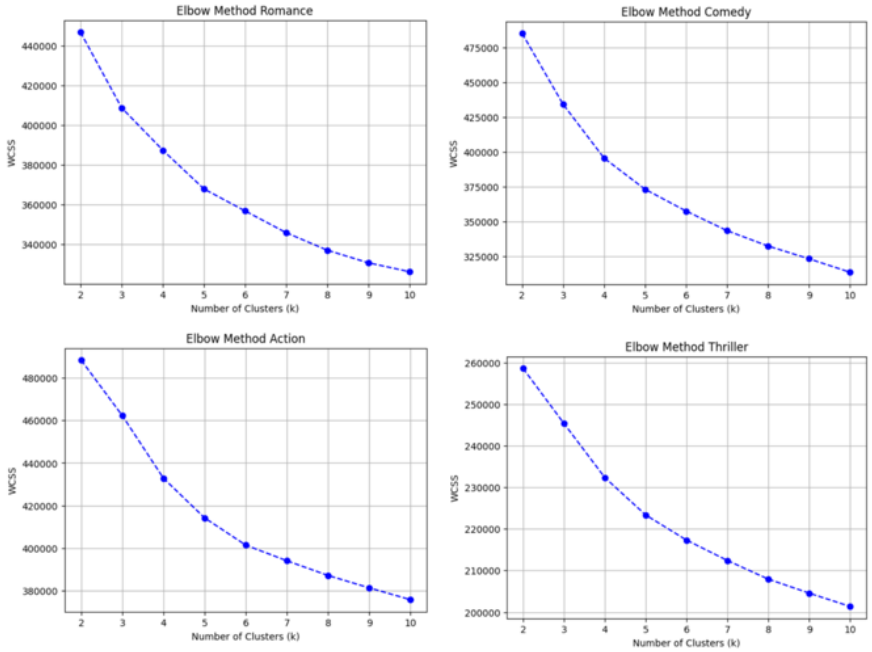


Figure 6. Determining the number of clusters using the elbow method

Based on Figure 6, the graph shows that the significant decrease in WCSS starts from the beginning and begins to slope at a certain value, which indicates an elbow point. For each genre, the elbow point is identified around a certain number of clusters, as seen in the graph, which is k=3 for each genre. After the value of k is obtained, it will enter the clusterization process.

The clustering process uses the K-means method, which groups the text based on the embedding representation generated during feature extraction. This embedding reflects the semantic meaning of the text. In the implementation, the embedding for all texts will be calculated using ”get_bert_embedding”, then converted into a numeric array. The K-Means algorithm is applied with the number of ”n_cluster = 3” according to the previous elbow method. Using the random parameter ”random_state=42”, this process will generate cluster labels on each text, which can be further analyzed to understand the patterns appearing in each group. Based on the clustering results using the K-Means algorithm below, the cluster distribution is visualized using Principal Component Analysis to reduce the embedding dimension generated by IndoBERT. Each cluster looks well separated, meaning the IndoBERT embedding successfully identified the semantic patterns. Figure 7 below is the clustering result using K-Means.

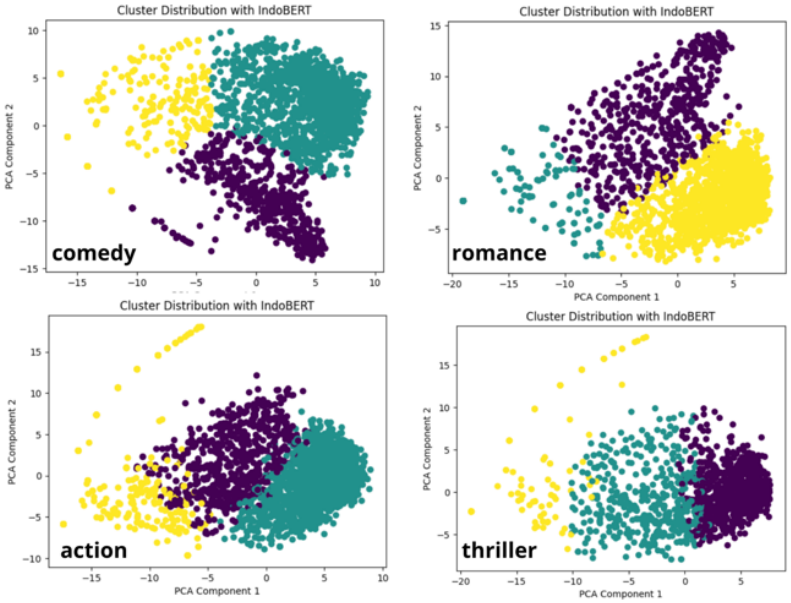


Figure 7. Clusters result using the k-means algorithm

### 3.4. Analyzed Text Sentiment Using SentimentIntensityAnalyzer

Based on sentiment analysis conducted using SentimentIntensityAnalyzer, each text that has gone through the lemmatization process will be classified into three categories: positive, negative, and neutral. The analysis results show that the sentiment distribution in each cluster has been generated by the K-Means algorithm. The sentiment percentage is calculated based on the number of sentiments in each cluster against the total sentiments in all clusters. The sentiment distribution across genres is shown in Figure 8, which highlights the sentiment analysis result for comedy, romance, action, and thriller.

**comedy**

| Sentiment Cluster | negative | neutral | positive |
|---|---|---|---|
| 0 | 39 | 391 | 129 |
| 1 | 31 | 1206 | 68 |
| 2 | 3 | 419 | 20 |

| Sentiment Cluster | negative | neutral | positive |
|---|---|---|---|
| 0 | 38 | 505 | 86 |
| 1 | 47 | 1395 | 94 |
| 2 | 1 | 363 | 14 |

**action**

**romance**

| Sentiment Cluster | negative | neutral | positive |
|---|---|---|---|
| 0 | 42 | 455 | 81 |
| 1 | 1 | 285 | 3 |
| 2 | 38 | 1410 | 84 |

| Sentiment Cluster | negative | neutral | positive |
|---|---|---|---|
| 0 | 23 | 746 | 50 |
| 1 | 10 | 407 | 25 |
| 2 | 4 | 145 | 2 |

**thriller**

Figure 8. Sentiment analysis results

### 3.5. Evaluation of Sentimen Clusters Result

In this section, each cluster will be analyzed and observed to determine which genres are most interesting to Korean film audiences throughout 2024 based on data from the TikTok platform. Compared to previous studies, such as those conducted by Chen et al. (2022), which primarily use multiple platforms as data sources for sentiment analysis, this study provides a novel perspective by leveraging TikTok's unique engagement features, such as short video reviews, interactive comments, and hashtag-driven trends. These features allow sentiment trends to be observed in a more dynamic and real-time manner.

Based on the results of the sentiment clustering distribution of each genre, the romance genre has the highest dominance of neutral sentiment with 89.6%, followed by the comedy genre at 87.4% and action at 60.4%. This finding differs from previous studies that analyzed Twitter and YouTube data, where romance and comedy genres typically had higher positive sentiments due to the written nature of reviews that often contain direct expressions of preference. On the other hand, TikTok users tend to engage in more neutral discussions, likely influenced by the short video format that emphasizes reactions over detailed written critiques.

Regarding positive sentiment, the action genre recorded the highest percentage at 14.9%, surpassing romance (7.0%) and comedy (9.4%), with the thriller genre having the lowest at 5.4%. This trend aligns with past research on YouTube reviews, where

action films often receive high engagement due to their visual appeal and excitement factor. However, the difference in sentiment distribution suggests that TikTok audiences may respond differently to action films than users on other platforms.

The highest negative sentiment was also observed in the action genre at 24.7%, significantly higher than romance (3.4%), comedy (3.2%), and thriller (2.6%). This contrasts with findings from studies that analyzed Twitter reviews, where romance films often had a higher proportion of negative sentiment due to critical discussions about clichés and predictable storylines. The lower negative sentiment in TikTok's romance and comedy genre reviews may indicate a more casual and entertainment-driven approach by users on this platform.

Figure 9 provides a combined visualization of the clustering and sentiment results discussed in this section, highlighting the distribution of neutral, positive, and negative sentiments across genres. These differences highlight the impact of platform choice on sentiment analysis results. While previous studies using Twitter and YouTube data focused on text-based opinions, this study demonstrates that TikTok provides a more interactive and engagement-driven sentiment distribution. The findings contribute to a broader understanding of audience preferences in the Korean film industry by showcasing how sentiment varies across social media platforms.

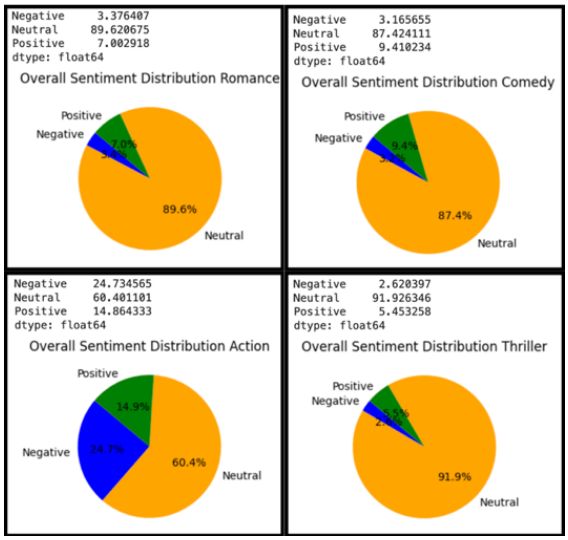


Figure 9. Combined visualization of clustering and sentiment results

## 4. CONCLUSION

Based on the research results of the 2024 Korean movie audience review data on the Tiktok platform, which only contains reviews in Indonesian and uses the K-means algorithm and SentimentIntensityAnalyzer, the romance genre is the most popular genre for viewers on the Tiktok platform in 2024 with a neutral sentiment dominance of 89.6%, followed by the comedy genre of 87.4%. On the other hand, the action genre stands out in positive sentiment at 14.9% but also has a negative sentiment range of 24.7%. The thriller genre shows relatively low numbers compared to the other three genres across all sentiment categories. This result shows that the audience prefers romance and comedy movies on the TikTok platform. Therefore, it is recommended that the Korean film industry should focus more on developing these two genres, while the action genre requires improvement in aspects that trigger negative sentiment.

In terms of algorithm usage, the combination of the IndoBERT feature extraction for feature extraction and the K-Means algorithm for clustering shows great potential in conducting unlabeled analysis. The K-Means algorithm can cluster the data into three main clusters, namely positive, negative, and neutral, using only PCA visualization, showing clear and separate cluster results. However, some limitations exist, such as difficulty in handling unbalanced data and data distribution, dependence on the optimal k value, and sensitivity to centroid initialization. Nevertheless, the advantages of the K-Means algorithm can be seen from its simplicity and efficiency in processing large data. For future research, it is recommended that other algorithms, such as DBSCAN or Gaussian Mixture Models, be used to overcome uneven data distribution. In addition, future research can also rely on reviews from other platforms to enrich the trend analysis of audience preferences. Overall, this approach has shown excellent results in understanding the sentiment patterns of Korean movie viewers through the TikTok platform.

## 5. ACKNOWLEDGEMENTS

The author would like to express sincere gratitude to everyone who has contributed to completing this research. Special thanks go to the research funders, particularly the author's family and Kaimas, for their unwavering support and encouragement. The author also appreciates all scientific contributors, especially the research supervisors and academic advisors, for their valuable guidance and insightful feedback throughout the research process. Further gratitude is directed to those who assisted with the research, including colleagues and friends who have offered support in various forms. Their help and motivation have been instrumental in overcoming challenges along the way.

## 6. DECLARATIONS

### AUTHOR CONTIBUTION

Baiq Rima Mozarita Erdiani: coding, testing, writing, and editing. Ario Yudo Husodo: conceptualization, methodology, writing review, and validation. Ida Bagus Ketut Widiartha: writing review and validation.

### FUNDING STATEMENT

This research was funded by the parents of the lead author in support of academic development. This funding is completely independent and does not involve third parties. Funds have been allocated to the primary needs of the research, including software and data collection. This approach ensures the research is conducted without pressure or conflict of interest from external sponsors. As such, the research results reflect the authors' objectivity and integrity.

### COMPETING INTEREST

The authors expressly declare that they have no conflict of interest in this research. No external parties influenced the study or its published results. All decisions in the study were made entirely based on objective data analysis and interpretation. The authors are committed to maintaining transparency throughout the research and publication process, which demonstrates their dedication to high research ethics.

## REFERENCES

[1] V. Glodev, G. Wijaya, and R. Ida, "The Korean Wave as the Globalization of South Korean Culture," vol. 22, no. 1, pp. 108–120, https://doi.org/10.32509/wacana.v22i1.2671.

[2] F. T. Laily and A. P. Purbantina, "Digitalisasi Industri Perfilman Korea Selatan melalui Netflix sebagai Alternatif Pasar Ekspor Film," vol. 4, no. 2, p. 141, https://doi.org/10.33021/exp.v4i2.1494.

[3] S. V. Mahardhika, I. Nurjannah, I. I. Ma'una, and Z. Islamiyah, "Faktor-Faktor Penyebab Tingginya Minat Generasi Post-Millenial di Indonesia terhadap Penggunaan Aplikasi TikTok," vol. 2, no. 1, pp. 40–53, https://doi.org/10.26740/sosearch.v2n1.p40-53.

[4] P. S. Rahmadani, F. C. Tampubolon, A. N. Jannah, N. L. H. Hutabarat, and A. M. Simarmata, "Tiktok Social Media Sentiment Analysis Using the Nave Bayes Classifier Algorithm," *Sinkron: jurnal dan penelitian teknik informatika*, vol. 6, no. 3, pp. 995–999, 2022, https://doi.org/10.33395/sinkron.v7i3.11579.

[5] J. C. Setiawan, K. M. Lhaksmana, and B. Bunyamin, "Sentiment Analysis of Indonesian TikTok Review Using LSTM and IndoBERTweet Algorithm," vol. 8, no. 3, pp. 774–780, https://doi.org/10.29100/jipi.v8i3.3911.

[6] E. Apriani, F. Oktavianalisti, L. D. H. Monasari, I. Winarni, and I. F. Hanif, "Analisis Sentimen Penggunaan TikTok Sebagai Media Pembelajaran Menggunakan Algoritma Naïve Bayes Classifier: Sentiment Analysis of Using TikTok as a Learning Media Using the Naïve Bayes Classifiers Algorithm," vol. 4, no. 3, pp. 1160–1168, https://doi.org/10.57152/malcom.v4i3.1482.

[7] S. Jung, D. Murthy, B. S. Bateineh, A. Loukas, and A. V. Wilkinson, "The Normalization of Vaping on TikTok Using Computer Vision, Natural Language Processing, and Qualitative Thematic Analysis: Mixed Methods Study," vol. 26,December, p. e55591, https://doi.org/10.2196/55591.

[8] C. Chen, B. Xu, J.-H. Yang, and M. Liu, "Sentiment Analysis of Animated Film Reviews Using Intelligent Machine Learning," vol. 2022, July, pp. 1–8, https://doi.org/10.1155/2022/8517205.

[9] R. Merdiansah, S. Siska, and A. A. Ridha, "Analisis Sentimen Pengguna X Indonesia Terkait Kendaraan Listrik Menggunakan IndoBERT," vol. 7, no. 1, pp. 221–228, https://doi.org/10.55338/jikomsi.v7i1.2895.

[10] D. Abimanyu, E. Budianita, E. P. Cynthia, F. Yanto, and Y. Yusra, "Analisis Sentimen Akun Twitter Apex Legends Menggunakan VADER," vol. 5, no. 3, pp. 423–431, https://doi.org/10.32672/jnkti.v5i3.4382.

[11] N. A. Maori and E. Evanita, "Metode Elbow dalam Optimasi Jumlah Cluster pada K-Means Clustering," vol. 14, no. 2, pp. 277–288, https://doi.org/10.24176/simet.v14i2.9630.

[12] L. Efrizoni, S. Defit, and M. Tajuddin, "Hybrid Modeling to Classify and Detect Outliers on Multilabel Dataset based on Content and Context," vol. 13, no. 12, pp. 550–559, 2022/34/30, https://doi.org/10.14569/IJACSA.2022.0131267.

[13] S. Armand, M. Hafid T, and M. Rafi Muttaqin, "Analisis Sentimen Sistem E-tilang pada Platform Twitter Menggunakan Metode Naïve Bayes," vol. 7, no. 3, pp. 1989–1994, https://doi.org/10.36040/jati.v7i3.7023.

[14] D. Khyani, B. S. Siddhartha, N. M. Niveditha, B. M. Divya, and Y. M. Manu, "An Interpretation of Lemmatization and Stemming in Natural Language Processing," vol. 22, no. 10, pp. 350–357, https://www.researchgate.net/publication/348306833.

[15] R. Rinandyaswara, Y. A. Sari, and M. T. Furqon, "Pembentukan Daftar Stopword Menggunakan Term Based Random Sampling Pada Analisis Sentimen Dengan Metode Naïve Bayes (Studi Kasus: Kuliah Daring Di Masa Pandemi)," vol. 9, no. 4, p. 717, https://doi.org/10.25126/jtiik.2022934707.

[16] Febriyanto A, D. S. S. Anggie, and I. Mulyadi, "Penerapan Algoritma K-Means terhadap Evaluasi Website E-commerce," vol. 3, no. 12, pp. 12–20, https://doi.org/10.59003/nhj.v3i12.1124.

[17] A. B. Saputra, P. W. Cahyo, M. Habibi, and A. Priadana, "Analysis and Visualization of BPJS on Twitter Using K-Means Clustering," vol. 3, no. 3, pp. 109–117, https://doi.org/10.31101/ijhst.v3i3.2466.

[18] D. Puspita and R. Syahri, "Penerapan Metode K-Means Clustering Untuk Pengelompokan Potensi Padi di Kota Pagar Alam," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 2, pp. 2187–2193, 2024, https://doi.org/10.36040/jati.v8i2.9432.

**[This page intentionally left blank.]**