Mathematical Modelling of Engineering Problems
Vol. 11, No. 12, December, 2024, pp. 3391-3401
Journal homepage: http://iieta.org/journals/mmep

Integrating Decision Tree and BIRCH Clustering Algorithms of BERTopic for Analyzing Public Sentiment on *Dirtyvote* Movie

Muhammad Muhajir[1,2], Dedi Rosadi[1*], Danardono[1]

[1] Department of Mathematics, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia
[2] Department of Statistics, Universitas Islam Indonesia, Yogyakarta 55584, Indonesia

Corresponding Author Email: dedirosadi@ugm.ac.id

https://doi.org/10.18280/mmep.111217

Received: 26 September 2024
Revised: 25 November 2024
Accepted: 4 December 2024
Available online: 31 December 2024

Keywords:
sentiment, topic modeling, Decision Tree, Gini Index, BERTopic, BIRCH clustering, Dirtyvote

ABSTRACT

This study analyzes public sentiment and topic modeling of YouTube comments on the politically charged film *Dirtyvote* during Indonesia's election period. Addressing the lack of robust methods for unstructured Indonesian-language social media data, the research proposes an integrative framework. This framework combines a Decision Tree algorithm with Gini Index for interpretable sentiment classification and BERTopic modified with BIRCH clustering to enhance stability and efficiency for large-scale topic modeling. The dataset comprises 76,502 YouTube comments, which were preprocessed to handle noise, informal language, and linguistic variations. Sentiment analysis results demonstrate the superior performance of the Decision Tree with Gini Index, achieving an accuracy of 98.72% and an F1-score of 96%, outperforming other methods such as SVM and Naïve Bayes. Meanwhile, BERTopic with BIRCH clustering achieved higher coherence metrics (e.g., CV, U_Mass, and NPMI) compared to standard BERTopic and K-Means clustering, showcasing its robustness in topic generation. This research contributes methodologically by introducing a scalable and interpretable framework for analyzing unstructured text data in Indonesian. Practically, it offers insights into public opinion dynamics on socio-political issues, highlighting the role of media in shaping perceptions. The findings underline the framework's potential for broader applications in sentiment analysis and topic modeling within diverse socio-political contexts.

1. INTRODUCTION

Elections in Indonesia are significant events that influence public discussions on political issues, election ethics, and the role of media in shaping public opinion [1]. The film *Dirtyvote* released during the election period, not only highlights controversial political issues but also successfully evokes emotional responses from the public, especially on social media platforms like YouTube. The public comments on this film provide an opportunity to explore public perceptions and sentiments regarding the socio-political dynamics reflected in the film [2, 3]. However, analyzing this large, complex, and often unstructured comment data requires an appropriate methodological approach to ensure accurate and interpretable results [4].

In this study, we introduce an innovative approach by integrating the Decision Tree algorithm optimized using the Gini Index for sentiment analysis and the BERTopic method with clustering modifications using BIRCH for topic modeling [5, 6]. This combination is expected to address the challenges of analyzing Indonesian-language text on social media, which is typically unstructured and linguistically diverse [7, 8].

Sentiment analysis and topic modeling are primary approaches to understanding public opinion as reflected in unstructured social media text data. Text data, especially in Indonesian, presents unique challenges, including the use of informal language, slang, and frequent spelling errors commonly found on platforms like YouTube. In this context, selecting the Decision Tree method with Gini Index optimization is highly relevant for sentiment analysis. The Decision Tree is known as an interpretable algorithm because classification results can be traced through the Decision Tree structure. Additionally, the Gini Index aids in selecting the most relevant features by minimizing impurity at each data split, thereby increasing classification accuracy [9-11].

Previous research shows that the Decision Tree model outperforms other methods like SVM, Naïve Bayes, or k-Nearest Neighbors in terms of interpretability, especially when applied to unstructured social media text [12, 13]. For instance, Apriliani et al. [10] demonstrated that a Decision Tree optimized with the Gini Index is more accurate in sentiment analysis than SVM on hotel comment data, which is similar to YouTube comments in its unstructured nature. Additionally, Ramasamy and Meena Kowshalya [14] highlighted that Gini Index-based feature selection improves Decision Tree performance for sentiment classification, particularly in heterogeneous text contexts.

For topic modeling, the BERTopic method was chosen for its ability to generate more contextual and accurate document embeddings using the BERT (Bidirectional Encoder

Representations from Transformers) model. BERTopic uses BERT-based sentence embeddings to capture semantic relationships between words in documents, enabling the identification of more meaningful and contextually appropriate topics [15]. In this study, we use IndoBERT as the embedding model, which is specifically designed for the Indonesian language, making it more effective in handling the local nuances of YouTube data in Indonesian. However, standard clustering methods in BERTopic, such as HDBSCAN, have limitations in clustering large unstructured data. Therefore, we modified BERTopic with BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies), which is more efficient and accurate for clustering high-dimensional data like text [16, 17].

The main novelty of this research lies in the integration of the Decision Tree with the Gini Index and the modification of BERTopic using BIRCH to address the challenges of complex social media text data. This combination offers several advantages, including:

Adaptability to unstructured data: This approach adapts well to unstructured Indonesian-language data, which is often filled with slang, spelling errors, and language variations. IndoBERT is used in BERTopic to generate embeddings relevant to Indonesian text. These embeddings enable BERTopic to generate more cohesive topics, which are then enhanced by BIRCH for more efficient and suitable clustering of large-scale data [16, 18, 19].

Higher clustering quality on large data: BIRCH was chosen as the clustering algorithm in BERTopic to improve cluster stability and quality. With its hierarchical structure, BIRCH can cluster diverse comment data more stably and meaningfully than other clustering methods, such as K-means, which has limitations in handling high-dimensional data. Ramadhani et al. [20] and subsequent research have shown that BIRCH can produce high-quality clusters with better efficiency on large datasets, making it suitable for large-scale social media topic modeling [17, 20].

High interpretability in sentiment analysis: The Decision Tree optimized with the Gini Index offers an advantage in terms of model interpretability, which is crucial in public opinion analysis. The Gini Index enables feature selection that prioritizes impurity minimization, assisting sentiment classification with higher accuracy. This distinguishes Decision Tree from other algorithms like SVM or Naïve Bayes, which are often less interpretable in text data sentiment analysis contexts [10, 11, 14].

Previous studies in sentiment analysis and social media topic modeling often used algorithms like SVM, Naïve Bayes, or k-Nearest Neighbor. However, this research demonstrates that the Decision Tree algorithm with the Gini Index outperforms others in interpretability and accuracy when applied to unstructured comment data. For example, Naïve Bayes has limitations in handling data imbalance and often performs suboptimally on heterogeneous social media data [9, 21]. Meanwhile, SVM, although highly accurate, is less interpretable due to its vector-based approach and does not provide clear information on feature importance in classification [12]. In topic modeling, BERTopic modified with BIRCH offers better cluster quality than methods like K-means or HDBSCAN. K-means, for instance, is less efficient in handling high-dimensional data, while HDBSCAN often produces unstable clusters on large datasets [16, 22].

This integrative approach not only contributes methodologically to the analysis of large and unstructured social media data but also enables a deeper understanding of public opinion dynamics on sensitive issues in a political context. By utilizing a combination of the Decision Tree optimized with the Gini Index and BERTopic integrated with BIRCH, this study presents an analytical framework that is precise, efficient, and interpretable for analyzing unstructured text data such as YouTube comments. The findings of this research reinforce the role of media in shaping public perception and offer new perspectives on how public comment data can be analyzed to depict the sentiments and topics dominating socio-political discourse.

Based on the structure of this paper, Section 2 reviews related work, focusing on existing approaches in sentiment analysis and topic modeling. Section 3 describes the data collection, preprocessing, and methodological choices that shape our framework. Section 4 presents the results and discusses their implications for public opinion analysis. Finally, Section 5 concludes by summarizing the findings, addressing limitations, and suggesting directions for future research.

2. RELATED WORKS

2.1 Decision Tree with Gini Index

Decision Tree has become one of the most widely used methods for text classification due to its high interpretability and ability to handle unstructured data [9, 23]. This method divides the data into smaller subsets based on selected features, with each split aiming to maximize class homogeneity. Two commonly used criteria for feature selection are Information Gain and Gini Index. While Information Gain is frequently used due to its capability to reduce uncertainty at each split [10], research has shown that this approach is prone to overfitting, especially when attributes have numerous unique values [24].

On the other hand, the Gini Index calculates data impurity by assessing class distribution at each split. This method is more stable than Information Gain, particularly for datasets with imbalanced class distributions [25]. The Gini Index also tends to produce simpler Decision Trees, which is crucial for model interpretability and generalizability [11]. However, literature focusing on the effectiveness of the Gini Index in sentiment analysis on unstructured data remains limited, particularly for Indonesian data with its rich linguistic variations. This study addresses this gap by applying Decision Tree with the Gini Index for sentiment analysis on social media comments.

Previous studies indicate that Support Vector Machine (SVM) is often employed for sentiment analysis due to its high accuracy [12]. However, as a "black-box" model, SVM lacks the interpretability provided by Decision Tree. Additionally, Naïve Bayes, known for its speed and simplicity, often struggles to handle the semantic complexity of unstructured data such as YouTube comments [21]. Thus, the adoption of Decision Tree with the Gini Index in this study not only offers better interpretability but also achieves high accuracy in capturing sentiment patterns in imbalanced datasets [6].

2.2 BERTopic with BIRCH clustering

Topic modeling has evolved from traditional probabilistic methods like Latent Dirichlet Allocation (LDA) to

embedding-based approaches such as BERTopic. LDA, while popular, has limitations in capturing semantic relationships between words, especially in short and unstructured texts [25]. In this context, BERTopic provides an advantage by leveraging semantic representations based on BERT [17]. By using sentence-level BERT embeddings, BERTopic captures semantic nuances, producing more cohesive and contextually relevant topics [25].

However, the default clustering methods in BERTopic, such as HDBSCAN, often lack stability for large datasets [26]. This study introduces the integration of BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies) within BERTopic to address this limitation. BIRCH adopts a hierarchical approach for clustering, enabling the formation of more stable and efficient clusters compared to methods like K-Means, which often produce spherical clusters that are less representative of textual data [18]. This research is among the first to integrate IndoBERT with BIRCH in BERTopic for topic modeling of Indonesian texts, offering significant contributions to understanding social media topic dynamics.

2.3 Research gap and justification of the proposed approach

While prior literature has explored various sentiment analysis and topic modeling methods, several research gaps remain:

Sentiment analysis: Studies on Decision Tree with Gini Index for unstructured data, particularly in the Indonesian language, are scarce. Most research only compares Gini Index with Information Gain without critically evaluating its effectiveness on social media data [10, 12].

Topic modeling: Alternative clustering integrations within BERTopic are rarely discussed in the literature, despite HDBSCAN's significant limitations on large datasets [26]. This study addresses this gap by proposing BIRCH as a more stable and efficient alternative.

The proposed approach not only addresses the shortcomings of conventional methods but also offers practical solutions to improve stability and efficiency in both topic modeling and sentiment analysis.

3. METHODS

3.1 Data collection

The data used in this study were obtained from public comments on YouTube related to the film *Dirtyvote*. Data were collected using a web scraping technique with the Comment Picker tool, which automatically extracted comments based on relevant keywords. After the extraction process, the data were filtered to remove duplicate and spam comments, resulting in a total of 76,502 comments. These comments contained various sentiment types (positive, negative, and neutral), providing insights into public perceptions of the socio-political issues highlighted in the film.

Dataset characteristics:

Dataset size: 76,502 comments.

Source: YouTube comments on the film *Dirtyvote*.

Language variation: Indonesian, with informal slang and regional language mixes [27].

Sentiment distribution: The data exhibited a heterogeneous sentiment distribution, including positive, negative, and neutral categories [28].

3.2 Data preprocessing

To ensure data quality, preprocessing was performed using Python libraries such as NLTK and Sastrawi [29]. The preprocessing steps included [30-32]:

Case folding: Converting all text to lowercase to ensure data consistency.

Removing usernames and hashtags: Eliminating irrelevant elements such as usernames and hashtags, which do not contribute to the analysis.

Removing URLs: Deleting links that are often irrelevant to the sentiment contained in the comments.

Filtering and removing punctuation: Eliminating symbols and punctuation to reduce noise in the data.

Normalization: Converting words with non-standard spellings into their standard forms to improve analysis accuracy [29].

Removing stop words: Using Sastrawi to remove common words in Indonesian that lack contextual significance, such as "dan," "di," "yang", etc.

Removing specific words: Removing certain words deemed irrelevant to the context of sentiment analysis or topic modeling.

These steps resulted in cleaner text data, optimized for further analysis with a focus on relevant sentiment and topic features.

3.3 Design of proposed methods

This study addresses the challenges of analyzing unstructured text data, particularly in public opinion and sentiment on social media platforms like YouTube, by proposing an integrated methodology that combines Decision Tree with Gini Index for sentiment analysis and BERTopic with BIRCH clustering for topic modeling [9, 16, 28]. The proposed framework ensures a robust, scalable, and interpretable approach to processing large-scale and complex datasets [24, 33].

The Decision Tree algorithm, enhanced by the Gini Index, minimizes data impurity to prioritize relevant features, ensuring accurate classification of sentiments (positive, negative, or neutral) [11, 25]. The Gini Index measures the homogeneity of data splits, selecting features that result in the least impurity [10]. This makes the Decision Tree highly effective for handling unstructured and noisy datasets like public comments on YouTube [12]. Compared to methods like Support Vector Machines (SVM) or Naïve Bayes, this approach offers superior interpretability, enabling researchers to trace how each sentiment is classified [13].

For topic modeling, BERTopic employs contextual embeddings generated by IndoBERT, capturing nuanced semantic relationships in Indonesian texts [16, 18, 34]. This is further refined by integrating BIRCH clustering, which replaces conventional methods like K-Means and HDBSCAN [17, 34]. Unlike these methods, BIRCH offers a hierarchical approach, ensuring stable and efficient clustering even in high-dimensional and large-scale datasets [17, 19]. This enhancement significantly improves the accuracy and interpretability of identified topics, particularly in datasets with complex linguistic variations [30, 35].

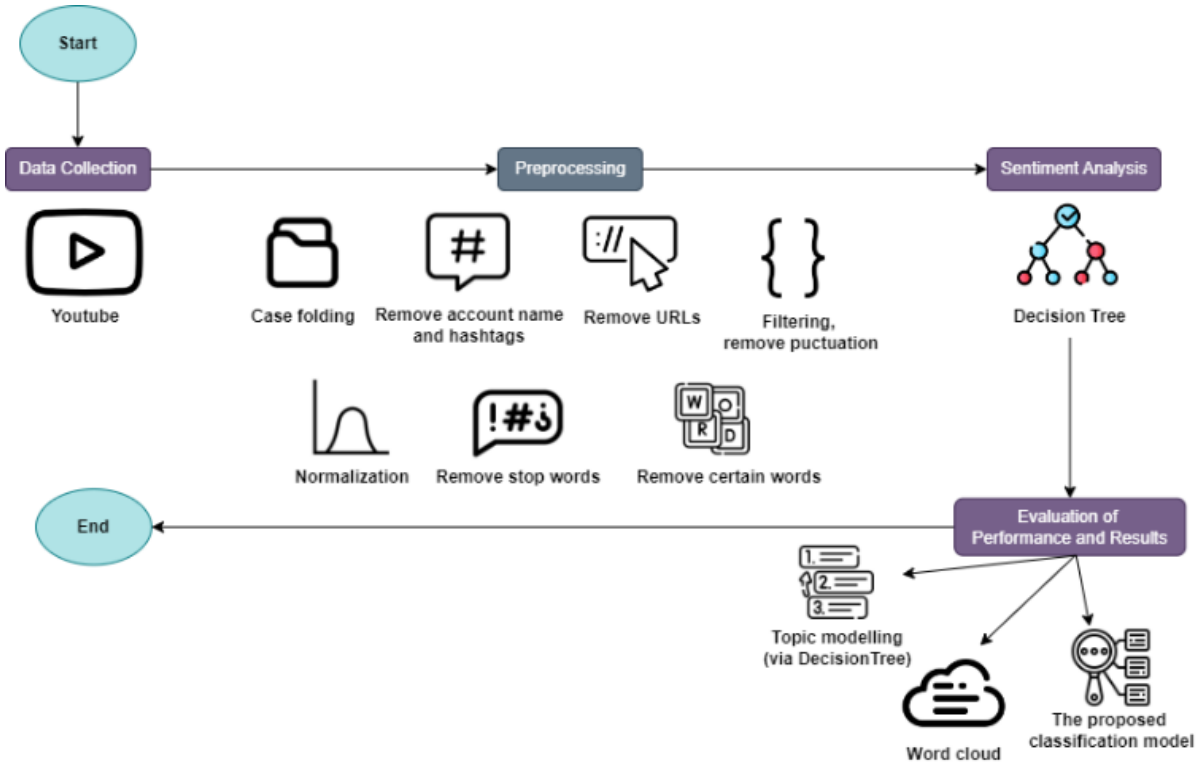

Figure 1. The framework of the proposed model

The proposed model framework, illustrated in Figure 1, outlines the steps in this study to analyze public sentiment and topics derived from YouTube comments on the film *Dirtyvote*. Through this approach, the study aims to gain deeper insights into public responses and socio-political dynamics reflected in the data.

This research was conducted through several main stages, as illustrated in the process diagram, to perform sentiment analysis and topic modeling on YouTube comments related to the film *Dirtyvote*. Each stage is described in detail below:

Stage 1: Data collection

Stage 2: Data preprocessing

Stage 3: Sentiment analysis

After preprocessing, sentiment analysis was performed using the Decision Tree model optimized with Gini Index as the feature selection criterion. This model classified sentiments in the comments into positive, negative, and neutral categories based on features derived from the text.

Stage 4: Evaluation of performance and results

This stage involved evaluating the model's performance using metrics such as accuracy, precision, recall, and F1-score. The evaluation aimed to assess the superiority of the Decision Tree with Gini Index model compared to other methods such as Decision Tree with Information Gain, SVM, and Naïve Bayes [36].

The model training process was conducted using an 80:20 data split for training and testing. Specifically, 80% of the data was used to train the model, while the remaining 20% was utilized to test the model's performance on unseen data.

Additionally, the document clustering stage was modified by implementing the BIRCH algorithm, chosen for its ability to efficiently handle large datasets and produce stable clusters [20]. This process was further enhanced by utilizing the BM25 weighting scheme [37], which assigns greater importance to contextually relevant words in the analysis, and IndoBERT-based text embeddings, designed to better capture the semantic nuances of the Indonesian language.

Supplementary visualizations, such as Word Clouds, were used to display the distribution of keywords within the dataset, providing insights into the frequency and relevance of specific terms in the comments [38].

End stage:

The final results of the research include the proposed classification model and the evaluation of both sentiment analysis and topic modeling. This model is expected to contribute significantly to understanding public opinion, particularly within the socio-political context highlighted by the film *Dirtyvote*.

3.4 Algorithm for integrating Decision Trees with Gini Index and BERTopic using BIRCH

The integration of Decision Tree with Gini Index for sentiment analysis and BERTopic with BIRCH for topic modeling follows these key steps:

3.4.1 Sentiment analysis using Decision Tree with Gini Index

Let D represent the dataset containing n instances, where each instance has features $X = \{x_1, x_2, \dots, x_m\}$ and a target class $y \in \{c_1, c_2, \dots, c_k\}$ representing sentiment categories (e.g., positive, negative, and neutral) [9].

Gini Index calculation: For a dataset D, the Gini Index $G(D)$ is calculated as:

$$G(D) = 1 - \sum_{i=1}^{k} \left(\frac{|D_i|}{|D|}\right) \quad (1)$$

In Eq. (1), $|D_i|$ is the number of instances belonging to class c_i, and $|D|$ is the total number of instances [11, 25].

Feature split: For a feature x_j, calculate the weighted Gini Index for a split S into subset D_L and D_R:

$$GS(x_j) = \frac{|D_L|}{|D|}.G(D_L) + \frac{|D_R|}{|D|}.G(D_R) \quad (2)$$

In Eq. (2), the feature x_j that maximizes $GS(x_j)$ is selected for the split [24].

Tree construction:

•Begin with root node containing all data D.

•Recursively split D into subset D_L and D_R using selected feature x_j.

•Stop splitting when $G(D) = 0$ (pure node) or a maximum depth d_{max} is reached [37].

Prediction: For a test instance t, traverse the tree to classify t based on the majority class in the leaf node [16].

3.4.2 Topic modeling using BERTopic with BIRCH clustering

Let $T = \{t_1, t_2, \dots, t_n\}$ represent the set of n text comments. The BERTopic algorithm consists of the following steps:

Text embedding: Generate embeddings for each text t_i using IndoBERT:

$$E(t_i) = IndoBERT(t_i), E = \{E(t_1), \dots, E(t_n)\} \quad (3)$$

In Eq. (3), $E(t_i) \in \mathbb{R}^d$ is a d dimensional vector representation [18].

Dimensionality reduction: Apply Uniform Manifold Approximation and Projection (UMAP) to reduce the dimensionality of the embeddings in Eq. (4) [39]:

$$U(E) = UMAP(E), U(E) \in \mathbb{R}^p, p \ll d \quad (4)$$

Clustering with BIRCH:

•Represent the reduced embeddings as input $U(E)$.

•Construct a clustering tree hierarchically, where each node summarizes a cluster of embeddings.

•The clustering process is mathematically defined in Eq. (5):

$$CF = (N, LS, SS) \quad (5)$$

where, N is the number of points in the cluster; LS is the linear sum of data points; SS is the squared sum of data points.

In Eq. (6), assign each new embedding $u \in U(E)$ to the closest cluster CF if:

$$\|LS - u\| \leq threshold \quad (6)$$

Otherwise, create a new cluster [17].

Topic representation: For each cluster CF_k, extract dominant keywords $\{w_1, w_2, \dots, w_q\}$ using BM25 which is defined mathematically in Eq. (7) [40]:

$$BM25(w, CF_k) = \frac{TF(w, CF_k).IDF(w).(k_1 + 1)}{TF(w, CF_k) + k_1(1 - b + b.\frac{|CF_k|}{avgdl})} \quad (7)$$

where, $TF(w, CF_k)$ is the frequency of term w in cluster document CF_k, $|CF_k|$ is the size of cluster, *avgdl* is the average document length in CF_k, k_1 and b are adjusted parameters and $IDF(w) = \log(\frac{N - n(w) + 0.5}{n(w) + 0.5})$.

3.4.3 Integration of Decision Tree and BERTopic

Let D_s and D_t represent the datasets used for sentiment analysis and topic modeling, respectively:

D_s: Input for Decision Tree to classify sentiment y.

D_t : Input for BERTopic to generate topic clusters $= \{T_1, T_2, \dots, T_m\}$.

The integration process is as follows:

•Perform sentiment classification on D_s using the Decision Tree in Eq. (8):

$$y_i = DecisionTree(x_i), \forall x_i \in D_s \quad (8)$$

•Generate topic clusters for D_t using BERTopic with BIRCH in Eq. (9):

$$T_j = BERTopic_{BIRCH(t_j)}, \forall\ t_j \in D_t \quad (9)$$

•Combine results to analyze the relationship between sentiment y and topics T in Eq. (10):

$$Analysis = \{(y_i, T_j)\}, i = 1, \dots, n; j = 1, \dots, m \quad (10)$$

4. RESULTS AND DISCUSSION

This section analyzes the topic modeling and sentiment analysis steps, as well as the performance of the proposed IndoBERT classification model. In this research, the Word cloud technique is used for the analysis of "usage of common words". Word cloud is one of the most popular techniques to represent textual data in data visualization. In Figure 2, it can be seen that the comment is dominated by the words "film, dirty, vote, election, video, aja, rakyat, jokowi". For this reason, the most significant words do not appear in the visualized word set in the sentiment analysis and topic modeling processes. Thus, the dominant words are ignored in the word cloud representation.

4.1 Sentiment analysis and evaluation

4.1.1 Sentiment distribution of comment

The findings of this study show that people's reactions to the film *Dirtyvote* are often positive, while the dominant topics vary greatly in each sentiment analysis. Therefore, topic modeling and sentiment analysis play an important role in revealing topic variations in this study.

In this case, after the 76502 comments in the initial dataset had been pre-processed, they were analyzed using the IndoBERT method, a transform-based sentiment analysis tool to classify them into three different sentiment polarities: 39,714 positive comments (51.91%), 7,182 negative comments (9.39%), and 29,606 neutral comments (38.70%). Figure 3 shows the percentage of sentiment polarity in the netizen's comment dataset. It can be observed that half of the reactions to the *Dirtyvote* film are positive, while the percentage of negative reactions is very small. There is no doubt that the difference is influenced by the rigging of the 2024 election.

Figure 2. Word cloud of *Dirtyvote* comment

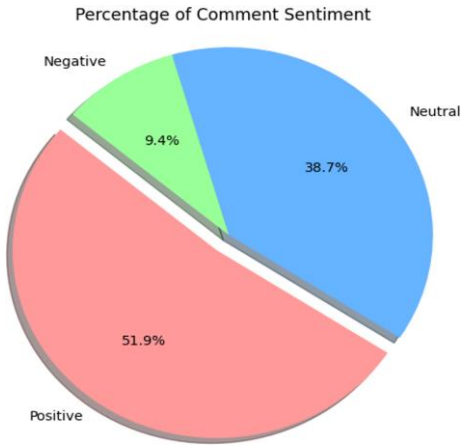

Figure 3. The comment dataset's sentiment polarity percentages

4.1.2 Sentiment model evaluation

This study uses the Decision Tree method with Gini Index feature selection for sentiment analysis on the film *Dirtyvote*. The Gini Index enables the model to separate data more effectively at each node, resulting in superior accuracy and balanced metrics by focusing on key sentiment-indicative features like emotionally charged words, enhancing the model's ability to accurately classify positive, negative, and neutral sentiments. To ensure objective results, the dataset is split into 80% for training and 20% for testing. Table 1 shows the model's testing performance, comparing the Gini Index-based Decision Tree with other methods, including Decision Tree with Information Gain, SVM, Naïve Bayes, and k-Nearest Neighbors (k-NN). Evaluation metrics accuracy, precision, recall, and F1-score help analyze each method's strengths and weaknesses to identify the most effective approach for unstructured comment data.

Table 1. Evaluation metrics of sentiment model using classification method

Evaluation Metrics (%)	Method				
	Decision Tree with Gini Index	Decision Tree with Information Gain	SVM	Naïve Bayes	k-NN
Accuracy	98.72	97.37	97.90	94.43	93.99
Precision	96.00	97.00	97.00	95.00	94.00
Recall	95.00	85.00	88.00	65.00	63.00
F1-Score	96.00	90.00	92.00	72.00	69.00

Table 1 presents the evaluation results of classification methods used in sentiment analysis of comments on the *Dirtyvote* film. The Decision Tree with Gini Index feature selection and SVM stood out with high performance. SVM achieved an accuracy of 97.90%, precision of 97.00%, recall of 88.00%, and an F1-score of 92.00%. Meanwhile, the Decision Tree with Gini Index feature selection slightly outperformed in terms of accuracy (98.72%), recall (95.00%), and F1-score (96.00%).

Table 1 shows the evaluation results of various sentiment classification methods using four main metrics: accuracy, precision, recall, and F1-score. These results highlight the strength of the Decision Tree with Gini Index as a superior classification method for sentiment analysis. With an accuracy of 98.72%, this model not only demonstrates exceptional ability in overall sentiment classification but also achieves high precision and recall values, at 96% and 95%, respectively. The high F1-score of 96% indicates an optimal balance between precision and recall, which is crucial in sentiment analysis applications on unstructured data like social media comments. This means that the model is not only accurate in identifying correct sentiments but also effective in capturing a diverse range of sentiments, making it an ideal choice for this study.

As a comparison, Decision Tree with Information Gain achieved slightly lower accuracy (97.37%), with higher precision (97%) but lower recall (85%). The resulting F1-score of 90% suggests that this model is less optimal in balancing precision and recall compared to the Decision Tree using Gini Index. With lower recall, this model may miss certain true sentiment examples, making it slightly less reliable in handling the range of sentiments in the comment data.

Support Vector Machine (SVM) also demonstrated solid performance with an accuracy of 97.90% and high precision of 97%. However, the recall value of 88% indicates that SVM is not as strong as Decision Tree with Gini Index in capturing the full spectrum of sentiment in imbalanced data. The F1-score of 92% suggests that while SVM is effective, it is slightly less optimal than the Gini Index model in balancing precision and recall, particularly in highly variable data contexts.

On the other hand, Naïve Bayes and k-Nearest Neighbors (k-NN) show lower performance in terms of both accuracy and the balance between precision and recall. Naïve Bayes recorded an accuracy of 94.43% with decent precision (95%), but a much lower recall (65%) indicates that this model is less effective at recognizing all relevant sentiment examples. With an F1-score of 72%, this model shows significant limitations in handling complex sentiment variations. k-NN has the lowest performance, with accuracy of 93.99%, precision of 94%, and recall of 63%, resulting in an F1-score of only 69%. These weaknesses indicate that both Naïve Bayes and k-NN are not ideal for sentiment analysis in this context, as they often fail to capture more subtle or complex sentiments in unstructured comment data.

Overall, the Decision Tree with Gini Index emerges as the most superior model in this study. Not only does it offer very high accuracy, but it also maintains an ideal balance between precision and recall, making it the top choice for sentiment analysis in large, unstructured social media data. With solid performance, SVM can be considered a good alternative, although slightly less optimal in recall. Conversely, the limitations shown by Naïve Bayes and k-NN emphasize the importance of selecting the right model in sentiment analysis, especially for highly variable data like public comments on social media platforms. These findings reinforce the relevance of Decision Tree with Gini Index as a reliable tool for detecting sentiment patterns in large datasets, while also contributing significantly to the development of text analysis methods for socio-political applications in Indonesia and similar contexts.

4.2 Integration of BIRCH for improved clustering in BERTopic

Figures 4-6 present the results of topic modeling comparisons based on coherence metrics CV, U_mass, and NPMI. These results were obtained from experiments involving the partitioning of the *Dirtyvote* film data, which included 1,000 to 10,000 comments. The experiments aimed to assess the effectiveness of BERTopic integrated with BIRCH clustering compared to standard BERTopic and BERTopic integrated with K-Means clustering. The findings demonstrate that BERTopic with BIRCH achieves higher topic coherence and stability, showcasing its advantages over alternative clustering methods.

In terms of CV coherence displayed in Figure 4, BERTopic with K-Means demonstrates the highest initial performance, starting at approximately 0.82 for N = 100. This suggests its effectiveness in generating highly coherent topics from smaller datasets. However, as the dataset size increases, there is a marked decline in CV values, which drop to around 0.53 by N = 1000. This trend indicates that K-Means struggles to maintain topic coherence in larger, unstructured datasets. In contrast, BERTopic with BIRCH clustering exhibits more stable CV coherence across all data sizes, beginning at 0.5 when N = 100 and remaining between 0.35 and 0.45 as the dataset expands. Standard BERTopic, lacking specialized clustering, shows greater fluctuations and achieves generally lower coherence values, starting at 0.5 for N = 100 but dipping as low as 0.3 around N = 200, before stabilizing near 0.4.

The U_Mass metric depicted in Figure 5 further emphasizes the advantages of BIRCH clustering. BERTopic with BIRCH begins with a U_Mass value close to 0.0 for N = 100 and maintains a similar value as dataset sizes increase, indicating high semantic coherence and stability. Conversely, BERTopic with K-Means experiences a steep decline in U_Mass, starting at roughly -0.2 when N = 100 and dropping to about -0.5 by N = 1000, highlighting its limitations in preserving topic coherence in larger datasets. Although standard BERTopic performs slightly better than K-Means in the middle range (N = 400 to 600), it does not achieve the stability of BIRCH, fluctuating between -0.1 and -0.3 across dataset sizes.

In the NPMI metric illustrated in Figure 6, BERTopic with BIRCH clustering again showcases its robustness. It achieves a high initial NPMI score of around 0.25 at N = 100, peaks near 0.3 at N = 400, and maintains stable values above 0.1 as the dataset increases. In contrast, BERTopic with K-Means starts with a lower NPMI of approximately -0.15 and gradually improves to 0.1 by N = 1000, although it remains lower than BIRCH. Standard BERTopic displays greater variability, starting below 0.0 and rarely exceeding 0.2 across all dataset sizes, indicating less consistent topic coherence.

In conclusion, the detailed performance analysis shows that BERTopic with BIRCH clustering is superior in maintaining stable and high coherence across various metrics and dataset sizes. BIRCH clustering demonstrates resilience, particularly with larger datasets, maintaining CV between 0.35 and 0.45, U_Mass near 0.0, and NPMI consistently above 0.1. In contrast, while BERTopic with K-Means shows strong initial performance in smaller datasets, it faces significant challenges as data size increases, with notable declines in both CV and U_Mass. Standard BERTopic exhibits greater variability and generally lower coherence across metrics. Therefore, the BIRCH clustering approach within BERTopic is highly recommended for large-scale, unstructured text data, especially in social media contexts where data volume and variability are substantial. This analysis underscores BIRCH's potential as a robust clustering technique to enhance topic coherence and interpretability in complex data environments.

4.3 Topic modeling of sentiment analysis

In this study, the sentiment analysis conducted on public reactions to the movie *Dirtyvote* involved only positive and negative sentiment data, as shown in Table 2. This decision was based on the aim of focusing the interpretation on the most obvious expressions of acceptance or rejection of the film, thus allowing for a sharper and more targeted analysis of the polarization of public opinion. Neutral sentiment data, which does not indicate a clear pro or con stance, was excluded from this analysis to avoid diluting the intensity of reactions that could obscure a deeper understanding of the true sentiment of the public.

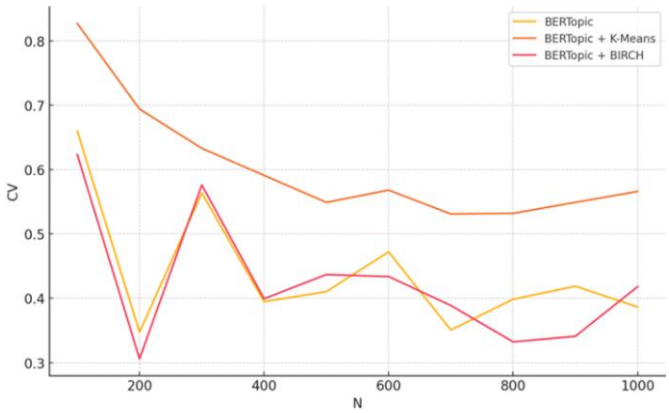

Figure 4. Comparison of CV for different methods

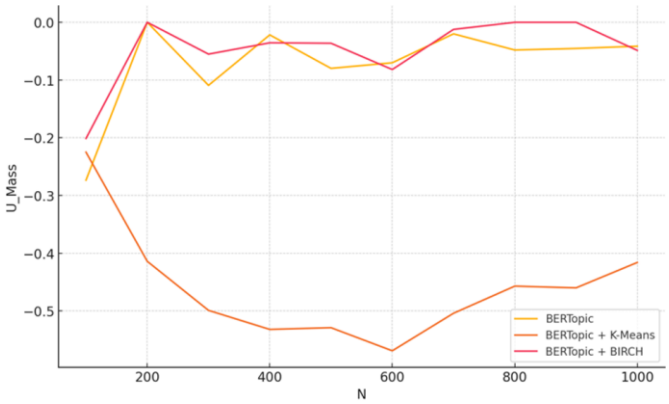

Figure 5. Comparison of U_Mass for different methods

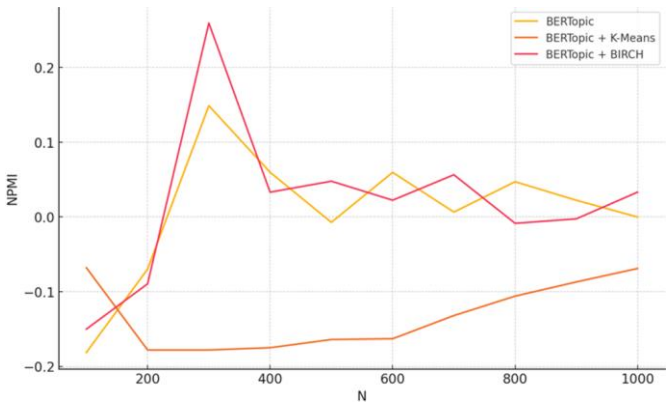

Figure 6. Comparison of NPMI for different methods

Table 2. BERTopic sentiment polarity regarding the film *Dirtyvote*

Sentiment Polarity	Topic Word	Explanation
Positive	"kpu", "capres", "bawaslu", "partai", dan "pilpres"	Positive discussions focused on the role of electoral institutions such as the KPU, Bawaslu, and political parties in the context of the 2024 elections. This topic shows that the film sparked conversations about the important role of these institutions in maintaining election integrity.
	"bikin", "horor", dan "capek"	*Dirtyvote* is considered a work that exposes the dark side of elections in an in-depth and perhaps laborious yet informative way. This reflects the appreciation for the effort put in by the production team in making this film.
	"jokowi", "presiden", dan "pemilu"	*Dirtyvote* sparked discussions related to President Jokowi. This discussion most likely relates to the President's role in the context of the election and the issues raised in the movie.
	"menjatuhkan", "menyudutkan", "tenang", "kampanye", dan "tujuan"	*Dirtyvote* films are considered effective tools to criticize and highlight the various strategies and goals of political campaigns that may involve cheating, but are done in a calm and structured manner.
	"mahfud", "beliau", "gubernur", "gibran", dan "anak"	Public figures such as Mahfud MD and Gibran Rakabuming Raka gained attention in the context of the discussion triggered by the film. This shows how the movie evokes discussions about the integrity and political engagement of these figures.
	"negeri", "selamatkan", dan "tercinta"	Patriotic sentiments and a drive to save democracy and electoral integrity in Indonesia. This shows that the movie succeeded in generating awareness and a spirit of nationalism among its audience.
	"golput", "tps", "pemilih", "paslon", dan "quick"	Voter participation, polling stations, and the overall election process are also important aspects of the positive response to this film.
	"tim", "vote", "dirty", dan "lindungi"	Voter participation, polling stations, and the electoral process in general were also an important part of the positive response to the movie.
Negative	"hadir", "pejabat", "kelompok", "faedah", dan "janji"	Dissatisfaction with the presence or actions of officials and associated groups in the electoral context.
	"golput", "memilih", "kampanye", dan "daerah"	Negative discussion on the phenomenon of abstention and election campaigns in various regions.
	"konstitusi", "etika", "identitas", "busuk", dan "melek"	Reflect criticism of constitutional and ethical violations in the electoral process.
	"pahlawan", "pejuang", "adu", "rezim", dan "jabatan"	Signaling dissatisfaction with certain figures and power struggles that are perceived as not neutral.
	"krik", "tai", "tas", "mempan", dan "bocil"	Indicates the presence of negative comments and insults that may be directed at certain individuals or groups.
	"panik", "ngeri", "membosankan", dan "menyedihkan"	Describes feelings of panic, fear and disappointment with the election situation.
	"selamatkan", "negeri", "indonesiaku", "lekas", dan "negeriku"	Concerns about the country's future and the urge to save the country from electoral fraud.
	"opo", "parah", "iki", dan "kabeh"	Expressions that show dissatisfaction or disappointment in the local language.

Table 2 displays our analysis of representative opinions based on the topic analysis of positive sentiment towards the *Dirtyvote* film. The main discussion focused on the important role of electoral institutions such as the KPU and Bawaslu, as well as political parties in maintaining the integrity of the 2024 elections. The public's appreciation of the film can be seen from the discussion about revealing the dark side of elections, which is considered informative and in-depth. The film also sparked conversations related to President Jokowi and the role of public figures such as Mahfud MD and Gibran Rakabuming Raka, raising issues of integrity and their political involvement.

In addition, the film succeeded in generating patriotic awareness and a drive to save democracy in Indonesia. Discussions on voter participation, polling stations, and the electoral process in general were also an important part of the positive response. Appreciation for the teamwork in making the film and the importance of maintaining clean and fraud-free elections stood out in the responses. Overall, *Dirtyvote* succeeded in raising awareness and motivating positive action for the future of Indonesian democracy.

Based on the topic analysis of negative sentiments towards the film, several key themes emerged from the community discussions. Topics such as dissatisfaction with the actions of officials and related groups, and criticism of constitutional and ethical violations in elections were prominent. Discussions about the abstention phenomenon and election campaigns in various regions also showed deep disappointment.

Dissatisfaction with certain figures and the perceived unhealthy power struggle added to the negative sentiment towards the political situation.

In addition, feelings of panic, fear, and disappointment with the electoral situation reflected people's concerns about the future of the country. Common expressions of discontent expressed in local languages or slang show the breadth of negative reactions. Overall, this negative sentiment reflects people's deep concerns about the integrity of the election and the fraudulent practices undermining Indonesian democracy, which the film *Dirtyvote* effectively exposed.

5. CONCLUSIONS

This study successfully proposes an integrative framework for sentiment analysis and topic modeling of YouTube comments related to the film *Dirtyvote*. By combining the Decision Tree with the Gini Index for sentiment classification and BERTopic with BIRCH clustering for topic modeling, the research addresses critical gaps in managing unstructured, large-scale, and linguistically diverse data, particularly in the Indonesian context. The framework demonstrates significant improvements in classification accuracy and topic coherence compared to conventional methods, offering a robust solution for analyzing public discourse in the digital age.

However, several limitations must be acknowledged. First, the dataset is confined to YouTube comments, which may not fully capture broader public sentiment across other social media platforms. Second, while the preprocessing steps taken to manage linguistic variations in the Indonesian language are effective, they may overlook deeper contextual nuances. Third, the static nature of the BERTopic model restricts its ability to analyze temporal changes in public opinion over time.

The practical implications of this study are substantial. The findings offer valuable insights into public opinion dynamics, particularly in politically sensitive contexts, which can guide policymakers, media analysts, and campaign strategists. Furthermore, the proposed framework provides a scalable solution for real-world applications, such as monitoring public sentiment during elections or understanding the societal impact of media narratives.

Future research should consider expanding the dataset to include comments from multiple social media platforms to enhance the generalizability of the findings. Additionally, integrating dynamic topic modeling techniques could enable the tracking of changes in public opinion over time, yielding richer insights. Finally, incorporating more advanced natural language processing techniques tailored for low-resource languages like Indonesian could further refine the framework and address existing limitations. These efforts will enhance the applicability of the framework in broader socio-political contexts and beyond.

ACKNOWLEDGMENT

The authors express their gratitude to the Center for Higher Education Funding (BPPT) and the Indonesia Endowment Funds for Education (LPDP) for their financial support of this research.

REFERENCES

[1] Alamsyah, A., Rochmah, W.Y., Nurnafia, A.N. (2021). Deciphering social opinion polarization towards political event based on content and structural analysis. arXiv preprint arXiv:2102.08249. https://doi.org/10.48550/arXiv.2102.08249

[2] Perdana, D.D. (2020). Reception analysis of related audience by watching “sexy killers” the documentary film. In 2nd International Media Conference 2019 (IMC 2019), pp. 86-98. https://doi.org/10.2991/assehr.k.200325.009

[3] Iedwan, A.S., Mauliza, N., Pristyanto, Y., Hartanto, A.D., Rohman, A.N. (2024). Comparative performance of SVM and multinomial Naïve Bayes in sentiment analysis of the Film'Dirty Vote'. Scientific Journal of Informatics, 11(3): 839-848. https://doi.org/10.15294/sji.v11i3.10290

[4] Hidayatullah, A.F., Apong, R.A., Lai, D.T., Qazi, A. (2023). Corpus creation and language identification for code-mixed Indonesian-Javanese-English Tweets. PeerJ Computer Science, 9: e1312. https://doi.org/10.7717/PEERJ-CS.1312

[5] Zhu, E., Yen, J. (2024). BERTopic-driven stock market predictions: Unraveling sentiment insights. arXiv preprint arXiv:2404.02053. https://doi.org/10.48550/arXiv.2404.02053

[6] Shang, S., Shi, M., Shang, W., Hong, Z. (2016). Improved feature weight algorithm and its application to text classification. Mathematical Problems in Engineering, 2016(1): 7819626. https://doi.org/10.1155/2016/7819626

[7] Aji, A.F., Winata, G.I., Koto, F., Cahyawijaya, S., et al. (2022). One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. arXiv preprint arXiv:2203.13357. https://doi.org/10.48550/arXiv.2203.13357

[8] Koto, F., Lau, J.H., Baldwin, T. (2021). IndoBERTWEET: A pretrained language model for Indonesian twitter with effective domain-specific vocabulary initialization. arXiv preprint arXiv:2109.04607. https://doi.org/10.48550/arXiv.2109.04607

[9] Quinlan, J.R. (1986). Induction of Decision Trees. Machine Learning, 1: 81-106. https://doi.org/10.1007/bf00116251

[10] Apriliani, D., Abidin, T., Sutanta, E., Hamzah, A., Somantri, O. (2020). Sentiment analysis for assessment of hotel services review using feature selection approach based-on Decision Tree. International Journal of Advanced Computer Science and Applications, 11(4): 240-245. https://doi.org/10.14569/IJACSA.2020.0110432

[11] Singh, J., Tripathi, P. (2021). Sentiment analysis of Twitter data by making use of SVM, Random Forest and Decision Tree algorithm. In 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, pp. 193-198. https://doi.org/10.1109/CSNT51715.2021.9509679

[12] Bayhaqy, A., Sfenrianto, S., Nainggolan, K., Kaburuan, E.R. (2018). Sentiment analysis about E-commerce from tweets using Decision Tree, K-nearest neighbor, and Naïve Bayes. In 2018 International Conference on

Orange Technologies (ICOT): Nusa Dua, Bali, Indonesia, pp. 1-6. https://doi.org/10.1109/ICOT.2018.8705796
[13] Zerrouki, K., Hamou, R.M., Rahmoun, A. (2022). Sentiment analysis of tweets using Naïve Bayes, KNN, and Decision Tree. International Journal of Organizational and Collective Intelligence, 10(4): 35-49. https://doi.org/10.4018/ijoci.2020100103
[14] Ramasamy, M., Meena Kowshalya, A. (2022). Information gain based feature selection for improved textual sentiment analysis. Wireless Personal Communications, 125(2): 1203-1219. https://doi.org/10.1007/s11277-022-09597-y
[15] Alotaibi, A., Rahman, A., Alhaza, R., Alkhalifa, W., Alhajjaj, N., Alharthi, A., Abushoumi, D., Alqahtani, M., Alkhulaifi, D. (2022). Spam and sentiment detection in Arabic tweets using MARBERT model. Mathematical Modelling of Engineering Problems, 9(6): 1574-1582. https://doi.org/10.18280/mmep.090617
[16] Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805
[17] Zhang, T., Ramakrishnan, R., Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. ACM Sigmod Record, 25(2): 103-114. https://doi.org/10.1145/235968.233324
[18] Koto, F., Rahimi, A., Lau, J.H., Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. arXiv preprint arXiv:2011.00677. https://doi.org/10.48550/arXiv.2011.00677
[19] Sawant, S., Yu, J., Pandya, K., Ngan, C.K., Bardeli, R. (2022). An enhanced BERTopic framework and algorithm for improving topic coherence and diversity. In 2022 IEEE 24th International Conference on High Performance Computing & Communications; 8th International Conference on Data Science & Systems; 20th International Conference on Smart City; 8th International Conference on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys), Hainan, China, pp. 2251-2257. https://doi.org/10.1109/HPCC-DSS-SmartCity-DependSys57074.2022.00332
[20] Ramadhani, F., Zarlis, M., Suwilo, S. (2020). Improve BIRCH algorithm for big data clustering. IOP Conference Series: Materials Science and Engineering, 725(1): 012090. https://doi.org/10.1088/1757-899X/725/1/012090
[21] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers 2012.
[22] Lorbeer, B., Kosareva, A., Deva, B., Softić, D., Ruppel, P., Küpper, A. (2018). Variations on the clustering algorithm BIRCH. Big Data Research, 11: 44-53. https://doi.org/10.1016/j.bdr.2017.09.002
[23] Pandiangan, N., Buono, M.L.C., Loppies, S.H.D. (2020). Implementation of Decision Tree and Naïve Bayes classification method for predicting study period. Journal of Physics: Conference Series, 1569(2): 022022. https://doi.org/10.1088/1742-6596/1569/2/022022
[24] Jain, V., Phophalia, A., Bhatt, J.S. (2018). Investigation of a joint splitting criteria for Decision Tree classifier use of information gain and Gini Index. In TENCON 2018 - 2018 IEEE Region 10 Conference, Jeju, Korea (South), pp. 2187-2192. https://doi.org/10.1109/TENCON.2018.8650485
[25] Raileanu, L.E., Stoffel, K. (2004). Theoretical comparison between the Gini Index and information gain criteria. Annals of Mathematics and Artificial Intelligence, 41: 77-93. https://doi.org/10.1023/B:AMAI.0000018580.96245.c6
[26] Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3: 993-1022.
[27] Rianto, Mutiara, A.B., Wibowo, E.P., Santosa, P.I. (2021). Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation. Journal of Big Data, 8(1): 1-16. https://doi.org/10.1186/s40537-021-00413-1
[28] Naradhipa, A.R., Purwarianti, A. (2012). Sentiment classification for Indonesian message in social media. In 2012 International Conference on Cloud Computing and Social Networking (ICCCSN), Bandung, Indonesia, pp. 1-5. https://doi.org/10.1109/ICCCSN.2012.6215730
[29] Satapathy, R., Guerreiro, C., Chaturvedi, I., Cambria, E. (2017). Phonetic-based microtext normalization for twitter sentiment analysis. In 2017 IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, USA, pp. 407-413. https://doi.org/10.1109/ICDMW.2017.59
[30] Achsan, H.T.Y., Suhartanto, H., Wibowo, W.C., Dewi, D.A., Ismed, K. (2023). Automatic extraction of Indonesian stopwords. International Journal of Advanced Computer Science and Applications, 14(2): 166-171. https://doi.org/10.14569/IJACSA.2023.0140221
[31] Saputra, F.T., Wijaya, S.H., Nurhadryani, Y. (2020). Lexicon addition effect on lexicon-based of Indonesian sentiment analysis on twitter. In 2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia, pp. 136-141. https://doi.org/10.1109/ICIMCIS51567.2020.9354269
[32] Puvvula, D., Rodda, S. (2024). Enhancing decision making through aspect based sentiment analysis using deep learning models. Mathematical Modelling of Engineering Problems, 11(10): 2849-2858. https://doi.org/10.18280/mmep.111028
[33] McInnes, L., Healy, J., Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426. https://doi.org/10.48550/arXiv.1802.03426
[34] Geni, L., Yulianti, E., Sensuse, D.I. (2023). Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia using IndoBERT language models. Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika, 9(3): 746-757. https://doi.org/10.26555/jiteki.v9i3.26490
[35] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794. https://doi.org/10.48550/arXiv.2203.05794
[36] Abdi, A., Shamsuddin, S.M., Hasan, S., Piran, J. (2019). Deep learning-based sentiment classification of evaluative text based on multi-feature fusion. Information Processing & Management, 56(4): 1245-1259. https://doi.org/10.1016/j.ipm.2019.02.018
[37] Hsu, B.M. (2020). Comparison of supervised classification models on textual data. Mathematics, 8(5):

851. https://doi.org/10.3390/MATH8050851
[38] Felix, C., Franconeri, S., Bertini, E. (2017). Taking word clouds apart: An empirical investigation of the design space for keyword summaries. IEEE Transactions on Visualization and Computer Graphics, 24(1): 657-666. https://doi.org/10.1109/TVCG.2017.2746018
[39] McInnes, L., Healy, J., Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. Preprint arXiv:1802.03426. https://doi.org/10.48550/arXiv.1802.03426
[40] Yang, C.Z., Du, H.H., Wu, S.S., Chen, X. (2012). Duplication detection for software bug reports based on bm25 term weighting. In 2012 Conference on Technologies and Applications of Artificial Intelligence, Tainan, Taiwan, pp. 33-38. https://doi.org/10.1109/TAAI.2012.20