P-ISSN: 2656-1743 | E-ISSN: 2656-1735 **JURNAL RISET INFORMATIKA**
DOI: https://doi.org/10.34288/jri.v8i1.470 Vol. 8, No. 1. December 2025
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2022

COMPARATIVE MACHINE LEARNING ALGORITHMS FOR YOUTUBE SENTIMENT ANALYSIS ON DPR DEMONSTRATION 2025 USING LEXICON

Syafri Samsudin[-1], **Ahmad Abdul Chamid**[-2], **Ahmad Jazuli**[-3]

Teknik Informatika
Universitas Muria Kudus
syafrisamm@gmail.com[-1], abdul.chamid@umk.ac.id[-2], ahmad.jazuli@umk.ac.id[-3]

Abstract

The high volume of public comments on YouTube regarding the DPR Demonstrasion August 2025, which reached 43,910 raw data, presents a significant challenge in conducting efficient sentiment analysis. Time and cost limitations in manual labeling for large-scale datasets are a major obstacle in the development of predictive models. This study aims to address this problem by proposing a hybrid approach that integrates Lexicon-Based auto-labeling with a comparative evaluation of five Machine Learning algorithms. The research methodology included a text preprocessing stage that generated 40,097 unique comments, feature extraction using TF-IDF, and data sharing with an 80:20 ratio. The performance of the Support Vector Machine algorithm was comprehensively compared to Random Forest, Decision Tree, K-Nearest Neighbors, and Naive Bayes. The results of the experiment showed that the SVM model recorded the most superior performance with an accuracy of 96.5% and a weighted F1-Score of 0.966. This score significantly outperformed other benchmarking algorithms, where Random Forest came in second place with 89.2% accuracy, followed by Decision Tree at 85.6%, KNN at 84.6%, and Naive Bayes at the lowest with 84.0%. These findings validate that the integration of Lexicon-Based labeling with SVM classification is a highly accurate, robust, and efficient solution for handling sentiment analysis on large-scale social media data in Indonesia.

Keywords: Sentiment Analysis; Machine Learning; Lexicon-Based; YouTube Comments; DPR Demonstration

Abstrak

Tingginya volume komentar publik di YouTube mengenai Demonstrasi DPR Agustus 2025, yang mencapai 43.910 data mentah, menghadirkan tantangan signifikan dalam melakukan analisis sentimen yang efisien. Keterbatasan waktu dan biaya dalam pelabelan manual untuk dataset berskala besar menjadi hambatan utama dalam pengembangan model prediksi. Penelitian ini bertujuan untuk mengatasi masalah tersebut dengan mengusulkan pendekatan hibrida yang mengintegrasikan pelabelan otomatis Lexicon-Based dengan evaluasi komparatif lima algoritma Machine Learning. Metodologi penelitian mencakup tahap text preprocessing yang menghasilkan 40.097 komentar unik, ekstraksi fitur menggunakan TF-IDF, dan pembagian data dengan rasio 80:20. Kinerja algoritma Support Vector Machine diperbandingkan secara komprehensif terhadap Random Forest, Decision Tree, K-Nearest Neighbors, dan Naive Bayes. Hasil eksperimen menunjukkan bahwa model SVM mencatatkan kinerja paling unggul dengan akurasi mencapai 96,5% dan weighted F1-Score 0,966. Skor ini secara signifikan mengungguli algoritma pembanding lainnya, di mana Random Forest menempati posisi kedua dengan akurasi 89,2%, diikuti oleh Decision Tree sebesar 85,6%, KNN sebesar 84,6%, dan Naive Bayes di posisi terendah dengan 84,0%. Temuan ini memvalidasi bahwa integrasi pelabelan Lexicon-Based dengan klasifikasi SVM merupakan solusi yang sangat akurat, tangguh, dan efisien untuk menangani analisis sentimen pada data media sosial berskala besar di Indonesia.

Keywords: Analisis Sentimen; Machine Learning; Lexicon-Based; Komentar YouTube; Demo DPR

INTRODUCTION

The development of information technology has fundamentally changed the landscape of public communication, with social media evolving into a major space for expression and discussion (Umrona, Anwar, & Soelistijadi, 2025). Platforms like YouTube, which initially focused on video, have now become one of the largest discussion forums in the world through its

comment section, making it a very rich and unfiltered source of textual data to gauge public sentiment (Muhayat, Fauzi, & Indra, 2023). The opinion data contained in it can be extracted and classified into sentiment categories (such as positive or negative), thus providing an important data foundation for conducting sentiment analysis of public responses (Ardiansyah, Agustina, Maryani, & Pribadi, 2025).

Socio-political events, such as the "August 2025 DPR Demonstration", consistently trigger strong public reactions and generate massive volumes of opinion data on these platforms (Adriana, Suarjaya, & Githa, 2023). This phenomenon is interesting when viewed through the lens of the digital public sphere. Here, social media not only serves as a conveyor of information, but has evolved into a space for free discussion where people can form their own opinions independently (Utomo, 2022). In situations of political conflict, commotion in cyberspace is often a direct reflection of friction that occurs in the real world. This accumulation of "digital emotions" can even be the initial signal of mass movements. Therefore, the sentiment analysis of the DPR Demo is very important. This study is not just a technical matter of text processing, but a valid way to measure support for parliament and public trust in the government in the midst of a crisis (Ningsih & Fatah, 2025). YouTube plays a central role in this because it is able to turn previously passive viewers into active participants through the comment section, creating an honest and unfiltered set of public political preferences

Analysis of public opinion on this platform is very important to understand people's preferences and expectations on political issues, as is the analysis of regional head election debates which shows that YouTube comments are rich in insights into public sentiment (Chamid, Nindyasari, Azizah, & Hariyadi, 2025). However, the extreme volume of data makes manual analysis to understand public sentiment impossible, requiring a computational approach through Natural Language Processing (NLP) to automatically classify sentiment (Atinna & Akbar, 2025).

Previous research has shown that the performance of Machine Learning algorithms varies greatly depending on the characteristics of the dataset and the context of the issue being discussed. For example, (Mola, Lete, Triyanto, Ajilo, & Widiastuti, 2024) found that the Support Vector Machine (SVM) excelled with 91% accuracy compared to Naive Bayes (80%) in the case of the inauguration of DPR artists. On the other hand, (Umrona, Anwar, & Soelistijadi, 2025) succeeded in implementing K-Nearest Neighbor (KNN) to analyze the sentiment of the Pagar Laut issue, despite facing challenges for minority classes. On the other hand, (Adriana, Suarjaya, & Githa, 2023) compared SVM and Random Forest on the demonstration issue, where SVM again showed the dominance of accuracy. Meanwhile, the study by (Ningsih & Fatah, 2025) relies only on a purely Lexicon-Based approach without machine learning algorithms, which although efficient, often has limitations in the accuracy of complex sentence context classification.

Although previous research has shown the strength of SVM, a significant methodological gap remains. The pure Machine Learning approach, as used (Mola, Lete, Triyanto, Ajilo, & Widiastuti, 2024), has a major drawback, namely its reliance on the availability of large amounts of manually-labeled training data. This manual labeling process is labor-intensive and costly because it demands high consistency and expert involvement to ensure data quality (Chamid, Widowati, & Kusumaningrum, 2024). In addition, unstructured textual data often contains complex semantic relationships and ambiguities that are difficult to process without robust pre-processing or advanced modeling (Jazuli, Widowati, Chamid, & Kusumaningrum, 2025). On the other hand, the purely Lexicon-Based approach, as used (Ningsih & Fatah, 2025), while efficient because it does not require training data, often proves to be less accurate and difficult in understanding the complex context, irony, or sarcasm in natural language. This study proposes a comparative study to evaluate the performance of various Machine Learning algorithms, namely Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naive Bayes, Random Forest, and Decision Tree, which are trained using the data from the automatic labeling of the Lexicon-Based method (InSet dictionary). This approach aims to find the best model that is able to classify public sentiment in the case of the August 2025 DPR Demonstration with the highest accuracy.

Based on this background, this study aims to compare and analyze the performance of these algorithms in the case study of the August 2025 DPR Demonstration. The main contribution of this research includes methodological and practical aspects. Methodologically, the study offers a comprehensive evaluation of which machine learning algorithms are the most resilient in handling large-scale political data labeled automatically. Practically, this research is expected to provide a computational emotional picture of the

community related to the demonstration, as well as strengthen the use of social media as a valid and accurate source of social data for policy makers..

RESEARCH METHODS

Research methodology is the steps taken when conducting research so that it can be organized and structured according to the desired flow. This research was conducted with the aim of analyzing the sentiment of youtube comments on the August 2025 DPR Demonstration. This stage of research is a proposed framework specifically designed to integrate lexicon-based autolabeling with the evaluation of various classification models, as presented in Figure 1.

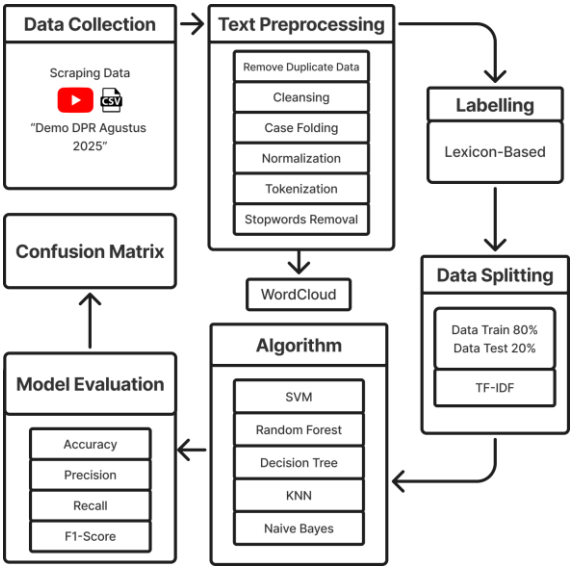

Figure 1. Research Stages

Based on Figure 1. The stages of research in analyzing data can be described as follows:

Data Collection

The first stage is the collection of primary data. The data was obtained from the youtube platform, by linking several youtube links about the August 2025 DPR demonstration. The youtube comment data period is August to October 2025, with a total of 43910 comment data. Data in the form of text comments was collected from the YouTube platform using the *scraping* using YouTube Data API V3 by targeting videos relevant to the keyword "August 2025 DPR Demo". The collected raw data is then stored in CSV file format for further processing.

Text Preprocessing

Text Preprocessing is a fundamental stage in the Natural Language Processing flow where the original text document is converted into a clean and structured data format (Syofiani, Alam, & Sulistyo, 2023). The existence of noise includes symbols, punctuation, and non-standard words. This is very common with social media raw data. So, this noise needs to be cleaned so as not to interfere with the process of optimizing the accuracy of model classification (Utami, Jazuli, & Khotimah, 2021).

The technical implementation of this research is carried out using the Python 3 programming language in the Google Colaboratory environment (Google Colab). A series of important libraries are leveraged to support the entire analysis flow: googleapiclient is used for data acquisition through the YouTube Data API v3, while Pandas and NumPy facilitate data manipulation and structural analysis. In the pre-processing stage, NLTK is applied for stopword removal (Ratnaswari, Wibowo, & Kartika, 2025), supported by re (Regular Expression) modules and strings for text cleanup. Data visualization was generated using Matplotlib, Seaborn, and WordCloud (Ningsih & Fatah, 2025). Furthermore, the Scikit-learn library serves as the primary framework for feature extraction and classification algorithm development, including SVM, KNN, Naive Bayes, Random Forest, and Decision Tree (Umrona, Anwar, & Soelistijadi, 2025).

Therefore, the text preprocessing pipeline is designed and executed to clean the text. These stages include:

1. Remove Duplicate Data (eliminates any identical comments that may appear).
2. Cleansing, which is the process of separating punctuation marks and symbols other than the alphabet (Ratnaswari, Wibowo, & Kartika, 2025).
3. Case Folding, a process to change uppercase characters (capitals) to lowercase letters (lowercase) (Undap, Rantung, & Rompas, 2021).
4. Normalization, a process that aims to change the list of slang words into more formal words according to KBBI or words that have not yet become root words (Undap, Rantung, & Rompas, 2021).
5. Tokenization, which divides sentences into tokens or word parts and is an important step in the text preprocessing process (Undap, Rantung, & Rompas, 2021).
6. Stopwords Removal, which is the process of eliminating words that are considered insignificant or excessive from a document (Chanda & Pal, 2023).

Lexicon-Based

To avoid subjective and time-consuming manual labeling, this study applied automatic labeling using the *Lexicon-Based*. The strategy of minimizing reliance on manually labeled data through an automated or semi-automated approach has been proven effective in improving the efficiency of the development of Indonesian-language sentiment analysis models (Chamid, Widowati, & Kusumaningrum, 2022). *Lexicon-Based* is one of the fundamental methods in sentiment analysis that works by utilizing a dictionary (lexicon) that has been prepared beforehand. In contrast to the approach *Machine Learning* which requires manual training data labeling, this method can perform data labeling automatically. This lexicon contains a list of words that have been given a marker of polarity or sentiment orientation (e.g., positive, negative, or neutral (Umrona, Anwar, & Soelistijadi, 2025).

This method identifies and analyzes sentiment in a text by matching the words in the text with the words in the lexicon dictionary (Ratnaswari, Wibowo, & Kartika, 2025). This approach has proven to be effective for automatic labeling, as shown in various studies in Indonesia using VADER or InSet dictionaries.

Data Splitting

Once the entire dataset has a label, the data is divided (*Data Splitting*) into two parts: 80% as training data (*Data Train*) and 20% as test data (*Data Test*).

TF-IDF

Allowing Machine Learning algorithms to process text data, it is necessary to convert clean text into numerical format, this procedure is known as feature extraction. The word weighting method used in this study is Term Frequency-Inverse Document Frequency (TF-IDF). This statistical method is used to measure the importance or relevance of the word level in a document relative to the overall set of documents. This technique is much more effective than sentiment classification, which reveals which words have an important role in defining the classification (Siddiq, Jayasri, Suhendi, Hidayat, & Rizky, 2025). There are two TF-IDF weight calculation metrics, namely, Term Frequency to measure how often the word *(t)* appears in a particular document *(d),* the higher the TF value, the higher the frequency, and then there is the Inverse Document Frequency to assess how unique a word is throughout the corpus. words that appear in many documents such as conjunctions will have a low IDF value. While words that rarely appear will get the highest score. The IDF formula is expressed in equation: (Wibowo, Witanti, & Susilawati, 2024) (1):

$$idf(t) = \log(N/df(t)\ ) \quad \text{..... (1)}$$

where (N) represents the total number of documents in the corpus, and *df(t)* is the number of documents that contain the word *(t).* The final value of the TF-IDF weight in a word is obtained by multiplying the two values, as in equation (2):

$$tf - idf(t,d) = tf(t,d) * idf(t) \quad \text{..... (2)}$$

With this mechanism, words that have a high frequency of occurrence in one specific document (high TF) but are rarely found in other documents in the collection (high IDF) will get a large TF-IDF weight score. This indicates that the word has high significance and is an important characteristic in the context of the sentiment analysis of the document.

Algorithm

This study conducted a comparative study of five Machine Learning algorithms to evaluate their performance in sentiment classification. The algorithms tested included Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, Random Forest, and Decision Tree. This thorough comparison is important to determine the most optimal and reliable algorithm for handling the characteristics of the YouTube comment dataset, given that algorithm performance can vary depending on the data distribution (Chamid, Nindyasari, & Ghozali, 2025).

a) Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the machine learning techniques that falls under the umbrella of supervised learning, whose foundation is built on statistical learning theory. This method has gained widespread recognition for its reliability in solving a wide range of classification challenges in the real world, from the recognition of handwriting patterns to the categorization of text documents. The main strength that sets SVM apart from other algorithms is its outstanding capabilities in managing high-dimensional data. This makes it an excellent choice for digital text analysis, where data is often represented in vectors with very large dimensions (such as the TF-IDF feature extraction results). In the sentiment analysis scenario, SVM is used to train the model to be

able to distinguish sentiment (for example, positive or negative) by studying the characteristics and patterns contained in the training data (Ratnaswari, Wibowo, & Kartika, 2025).

In this study, the SVM model was configured using Linear Kernel (kernel=*'linear'*). The selection of linear kernels is based on their effectiveness in separating sparse text data, where linear separation is often more efficient than non-linear kernels. In addition, the *random_state=42* parameter is applied to ensure consistency of results on each training execution. Conceptually, the way SVM works is centered on finding the most optimal hyperplane. This hyperplane serves as a decision boundary that separates two different classes of data. The main goal of this algorithm is to find the hyperplane that has the largest margin, i.e. the maximum distance between the separator plane and the nearest data point of each class. To achieve this, SVM performs a mathematical optimization process to determine the best parameters (weight vector w and bias b) that minimize the loss function. In the case of linear classification, the commonly used loss function is known as Hinge Loss, the formula of which can be seen in equation (3) (Uyun & Qoiriah, 2024).

$$L(w,b) = w^{T}w + C\sum \max\left(0, 1 - y^{(i)}\left(w^{T}F^{(i)} + b\right)\right) \quad \ldots\ldots (3)$$

b) K-Nearest Neighbors (KNN)

KNN is a supervised learning method that classifies new data based on the majority sentiment of its nearest 'K' neighbors (Hakim & Sugiyono, 2024). The determination of these neighbors is calculated based on the distance to the training data (Fasnuari, Andrian, Yuana, & Chulkamdi, 2022). In this experiment, the model is configured with parameters *n_neighbors=5* (*K=5*). This value is chosen as the standard to balance the trade-off between noise (if K is too small) and bias (if K is too large).

c) Naive Bayes

Naïve Bayes is a probabilistic classification algorithm that works by calculating the probability of the occurrence of a word in each sentiment class. The variant used is Multinomial Naïve Bayes (MultinomialNB) with default parameters, This variant was chosen because it is specifically designed for text data with word frequency or weight features, so it is very computationally efficient (Efraim & Ermatita, 2023).

d) Random Forest

Random Forest is an ensemble classification method consisting of a set of Decision Trees trained on random data and features. This algorithm is known to be accurate for large amounts of data, where the final outcome is determined by a majority vote of all trees. To ensure the stability of the model, this algorithm is configured with 100 Decision Trees (*n_estimators=100*). The random_state=42 parameter is also applied to control the randomness of the bootstrapping process, ensuring that the model produces consistent outputs (Syafia, Hidayattullah, & Suteddy, 2023).

e) Decision Tree

Decision Tree is an algorithm that forms a tree-shaped classification structure. This model divides data based on features that provide information gain or maximum Gini Index. Similar to other algorithms, the *random_state=42* parameter is used to lock randomness in tree formation so that performance evaluation remains objective (Adi, Bakkara, Zega, Vielita, & Rakhmawati, 2024).

Model Evaluation

Model Evaluation is an important step in the machine learning development cycle to determine how effective, reliable, and quality the model that has been built is. This stage serves as a benchmark to assess the model's ability to predict new data based on the patterns that have been learned during the training process (Merdiansah, Siska, & Ali Ridha, 2024).

In this study, the performance of the classification model was measured using the Confusion Matrix method. This method works by comparing the prediction labels generated by the model with the actual sentiment labels on the test data. This comparison allows researchers to see how accurately the model categorizes data into positive, negative, or neutral classes. The basic structure of the Confusion Matrix consists of four main elements: True Positive (TP), which represents correctly predicted positive data; True Negative (TN), for accurately predicted negative data; False Positive (FP), which indicates a prediction error where negative data is considered positive; and False Negative (FN), for errors where positive data is predicted to be negative.

These four components are then used as basic variables in the calculation of quantitative evaluation metrics to obtain a more comprehensive picture of performance (Umrona, Anwar, & Soelistijadi, 2025):

1) *Accuracy*: Measures how many predictions (TP and TN) are correct from the overall data.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{..... (4)}$$

2) *Precision*: Measures the accuracy of the model when predicting a comment as positive.

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{..... (5)}$$

3) *Recall*: Measures the model's ability to rediscover all comments that were supposed to be positive.

$$Recall = \frac{TP}{TP+FP} \quad \text{..... (6)}$$

4) *F1-Score*: The harmonic average of *Precision* and *Recall*, which provides a balanced picture of the model's performance.

$$F1 - Score = 2.\frac{Precision\ .\ Recall}{Precision+Recall} \quad \text{..... (7)}$$

RESULTS AND DISCUSSION

This study uses *scraping* techniques to obtain user review data from the YouTube platform. The following are the results of the research and discussion of the stages carried out in analyzing sentiment, starting from *data collection, text preprocessing*, *data labeling* to *compare algorithms* and *model evaluation*.

Data Collection

The data collection process (*scrap*) was carried out using YouTube Data API V3 to take public comments from videos with the keyword "August 2025 DPR Demo". The data taken were authors, comments, likes, and published_at (publication time), which covered the period from August to October 2025.

The total raw data collected was 43,910 comments. This data is then stored in a DataFrame Pandas for further processing. A snapshot of the raw data obtained is presented in Figure 2.

	author	comment	likes	published_at
0	@YusupAep-z2u	Ayo demo jilid 2 sampai DP bubar	0	2025-09-02T18:01:22Z
1	@Surahman-g8w	Kesenjangan sosial yg cukup jauh antara rakyat...	0	2025-08-31T05:58:46Z
2	@liuameng2720	Masih kalah sm thn 98.	0	2025-08-30T20:59:56Z
3	@Toms_Colonel	Media Jangan Jadi Kompor	1	2025-08-30T03:08:37Z
4	@allie7577	klo gedung DPR di jaga ketat,semoga para demo ...	0	2025-08-29T18:46:07Z

Figure 2. Data Collection Results

Text Preprocessing

The raw data in Figure 2 still contains a lot of *Noise* So it is necessary to go through the pre-processing stage. This process follows the flow that has been described in the research method, which includes *Remove Duplicate Data*, *Data Cleansing*, *Case Folding*, *Normalization*, *Tokenization*and *Stopwords Removal*.

a. *Remove Duplicate Data*

The first step is *Remove Duplicate Data*. Initial data amounted to 43,910 comments. After the deduplication process based on the comment field, the data is reduced to 40,409 unique entries. This shows that as many as 3,501 duplicate comments were successfully removed.

b. *Data Cleansing*

The data of 40,409 comments was then continued to the *Data Cleansing* process. This process aims to remove non-textual and irrelevant elements such as URLs, *usernames*, symbols, HTML tags, numbers, and emojis. The results of *this cleansing* stage can be seen in Table 1.

Table 1. Data Cleansing Results

Before	After
Kesenjangan sosial yg cukup jauh antara rakyat dan pejabatnya...	Kesenjangan sosial yg cukup jauh antara rakyat dan pejabatnya
Walaupun sudah punya gaji tinggi tetap aja korupsinya makin rakus.. Krn hukum di Indonesia buat koruptor sangat ringan.. Sementara rakyat kecil menjerit berjuang untuk menyambung hidup...	Walaupun sudah punya gaji tinggi tetap aja korupsinya makin rakus Krn hukum di Indonesia buat koruptor sangat ringan Sementara rakyat kecil menjerit berjuang untuk menyambung hidup
Ironis negara kaya sda alam hanya di nikmati segilintir orang.... Merdeka hy untuk mereka bukan untuk kita sebagai masyarakat yg terpinggirkan	Ironis negara kaya sda alam hanya di nikmati segilintir orang Merdeka hy untuk mereka bukan untuk kita

	sebagai masyarakat yg terpinggirkan
Bubarkan DPR cukup menteri aja karna DPR tidak tunduk dengan peresiden pak PERABOWO BERAVO❤❤❤❤❤❤❤❤❤❤❤❤❤❤❤❤❤❤	Bubarkan DPR cukup menteri aja karna DPR tidak tunduk dengan peresiden pak PERABOWO BERAVO

c. *Case Folding*

After *cleansing*, the process continues with *Case Folding*. This process aims to standardize the entire text into a lowercase format to avoid ambiguity. The results of *case folding* can be seen in Table 2.

Table 2. Case Folding Results

Cleansing	Case Folding
Kesenjangan sosial yg cukup jauh antara rakyat dan pejabatnya Walaupun sudah punya gaji tinggi tetap aja korupsinya makin rakus Krn hukum di Indonesia buat koruptor sangat ringan Sementara rakyat kecil menjerit berjuang untuk menyambung hidup Ironis negara kaya sda alam hanya di nikmati segilintir orang Merdeka hy untuk mereka bukan untuk kita sebagai masyarakat yg terpinggirkan	kesenjangan sosial yg cukup jauh antara rakyat dan pejabatnya walaupun sudah punya gaji tinggi tetap aja korupsinya makin rakus krn hukum di indonesia buat koruptor sangat ringan sementara rakyat kecil menjerit berjuang untuk menyambung hidup ironis negara kaya sda alam hanya di nikmati segilintir orang merdeka hy untuk mereka bukan untuk kita sebagai masyarakat yg terpinggirkan
Bubarkan DPR cukup menteri aja karna DPR tidak tunduk dengan peresiden pak PERABOWO BERAVO	bubarkan dpr cukup menteri aja karna dpr tidak tunduk dengan peresiden pak perabowo beravo

d. *Normalization*

The next step is *Normalization*. This process converts non-standard words (slang, abbreviations) into standard forms based on a standardized dictionary that has been prepared. This dictionary (*kamuskatabaku.xlsx*) is loaded from an external repository https://github.com/analysisdatasentiment/kamus_kata_baku/raw/main/kamuskatabaku.xlsx as a reference for word mapping. An example of normalization results is presented in Table 3.

Table 3. Normalization Results

Case Folding	Normalization
kesenjangan sosial yg cukup jauh antara rakyat dan pejabatnya walaupun sudah punya gaji tinggi tetap aja korupsinya makin rakus krn hukum di indonesia buat koruptor sangat ringan sementara rakyat kecil menjerit berjuang untuk menyambung hidup ironis negara kaya sda alam hanya di nikmati segilintir orang merdeka hy untuk mereka bukan untuk kita sebagai masyarakat yg terpinggirkan	kesenjangan sosial yang cukup jauh antara rakyat dan pejabatnya walaupun sudah punya gaji tinggi tetap saja korupsinya makin rakus karena hukum di indonesia buat koruptor sangat ringan sementara rakyat kecil menjerit berjuang untuk menyambung hidup ironis negara kayak sda alam hanya di nikmati segilintir orang merdeka hai untuk mereka bukan untuk kita sebagai masyarakat yang terpinggirkan
bubarkan dpr cukup menteri aja karna dpr tidak tunduk dengan peresiden pak perabowo beravo	bubarkan dpr cukup menteri saja karena dpr tidak tunduk dengan presiden pak perabowo beravo

e. *Tokenization*

Next, *the Tokenization* process is carried out to break down each sentence in a comment into individual word units (tokens). The result of this process is a list of words can be seen in Table 4.

Table 4. Tokenization Results

Normalization	Tokenization
kesenjangan sosial yang cukup jauh	['kesenjangan', 'sosial', 'yang', 'cukup', 'jauh', 'antara', 'rakyat', 'dan',

antara rakyat dan pejabatnya walaupun sudah punya gaji tinggi tetap saja korupsinya makin rakus karena hukum di indonesia buat koruptor sangat ringan sementara rakyat kecil menjerit berjuang untuk menyambung hidup ironis negara kayak sda alam hanya di nikmati segilintir orang merdeka hai untuk mereka bukan untuk kita sebagai masyarakat yang terpinggirkan	'pejabatnya', 'walaupun', 'sudah', 'punya', 'gaji', 'tinggi', 'tetap', 'saja', 'korupsinya', 'makin', 'rakus', 'karena', 'hukum', 'di', 'indonesia', 'buat', 'koruptor', 'sangat', 'ringan', 'sementara', 'rakyat', 'kecil', 'menjerit', 'berjuang', 'untuk', 'menyambung', 'hidup', 'ironis', 'negara', 'kayak', 'sda', 'alam', 'hanya', 'di', 'nikmati', 'segilintir', 'orang', 'merdeka', 'hai', 'untuk', 'mereka', 'bukan', 'untuk', 'kita', 'sebagai', 'masyarakat', 'yang', 'terpinggirkan']
bubarkan dpr cukup menteri saja karena dpr tidak tunduk dengan presiden pak perabowo beravo	['bubarkan', 'dpr', 'cukup', 'menteri', 'saja', 'karena', 'dpr', 'tidak', 'tunduk', 'dengan', 'presiden', 'pak', 'perabowo', 'beravo']

f. *Stopwards Removal*

The pre-processing stage ends with *Stopwords Removal*. This process of removing common words in Indonesian that do not have sentimental weight (e.g.: "yang", "di", "dan") can be seen in Table 5. This process utilizes *the Indonesian* stopwords corpus from *the NLTK* library.

Table 5. Stopward Removal Results

Tokenization	Stopwards Removal
['kesenjangan', 'sosial', 'yang', 'cukup', 'jauh', 'antara', 'rakyat', 'dan', 'pejabatnya', 'walaupun', 'sudah', 'punya', 'gaji', 'tinggi', 'tetap', 'saja', 'korupsinya', 'makin', 'rakus', 'karena', 'hukum', 'di', 'indonesia', 'buat', 'koruptor', 'sangat', 'ringan', 'sementara', 'rakyat', 'kecil', 'menjerit', 'berjuang', 'untuk', 'menyambung', 'hidup', 'ironis', 'negara', 'kayak', 'sda', 'alam', 'nikmati', 'segilintir', 'orang', 'merdeka', 'hai', 'untuk', 'mereka', 'bukan', 'untuk', 'kita', 'sebagai', 'masyarakat', 'yang', 'terpinggirkan']	kesenjangan sosial rakyat pejabatnya gaji korupsinya rakus hukum indonesia koruptor ringan rakyat menjerit berjuang menyambung hidup ironis negara kayak sda alam nikmati segilintir orang merdeka hai masyarakat terpinggirkan
['bubarkan', 'dpr', 'cukup', 'menteri', 'saja', 'karena', 'dpr', 'tidak', 'tunduk', 'dengan', 'presiden', 'pak', 'perabowo', 'beravo']	bubarkan dpr menteri dpr tunduk presiden perabowo beravo

WordCloud

To validate the effectiveness of all stages of preprocessing, a WordCloud visualization of the clean data is presented in Figure 3. This visualization highlights the terms that appear most frequently and meaningfully after the removal of noise and stopwords.

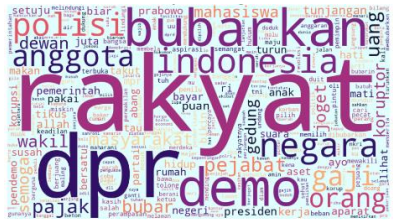

Figure 3. Wordcloud After Preprocessing

As illustrated in Figure 3, WordCloud displays words that are highly relevant to the topic "August 2025 DPR Demonstration". Dominant terms such as "rakyat", "DPR", "bubarkan", "polisi", and "demo" appear with significant frequency. The dominance of these keywords indicates that the preprocessing stage has managed to eliminate irrelevant noise, so that the core topics of public discourse in particular the demands of

accountability and criticism of institutions can be clearly seen for further sentiment analysis.

Labelling Data

After the data is clean, each comment is labeled sentiment. At this stage, a Lexicon-Based approach (*Lexicon-Based*). Each word in the comment is matched to an Indonesian sentiment dictionary (InSet) downloaded from the GitHub repository https://raw.githubusercontent.com/fajri91/InSet/master. The overall sentiment of a comment is determined based on the aggregation of the score of the words in it; where sentiment_score = positive_count - negative_count. Comments with a total score of ≤ 0 are labeled "Negative" and comments with a score of > 0 are labeled "Positive". The results of this labeling process are presented in Table 6.

Table 6. Labelling Data Results

Stopwards Removal	Score	Sentiment
bubarkan dpr gausah berdalih undang pokoknya bubarin	-2	Negative
ya allah lindungi saudara berjuang menegakkan keadilan diseluruh penjuru negeri allah maha pelindung engkau pelindung berlindung engkau tolong ya rabb amin ya rabbalalamiin	3	Positive

To provide a clearer understanding of the landscape of public opinion, the class distribution of sentiment labeling results is analyzed. As illustrated in Figure 4, the dataset shows a significant imbalance in sentiment polarity.

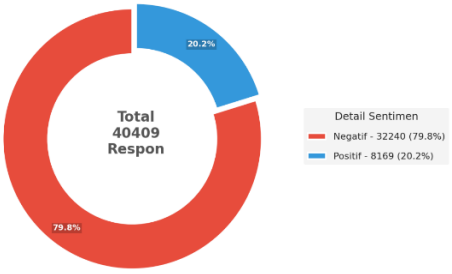

Figure 4. Distribution of Public Response Sentiment

Based on the data presented in Figure 4, out of a total of 40,409 responses, the majority of comments were dominated by negative sentiment. Specifically, 32,240 comments (79.8%) were classified as "Negative", while only 8,169 comments (20.2%) were classified as "Positive". This distribution indicates a strong tendency for public dissatisfaction with the topic.

Data Splitting

The dataset, which now has 40,097 unique data labeled ('Positive' or 'Negative') is then divided into two parts: training data and test data. This division uses an 80:20 ratio, resulting in 32,077 training data (80.0%) and 8,020 test data (20.0%), as shown in Figure 5.

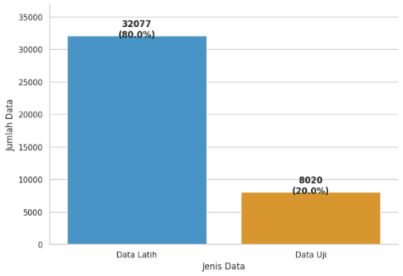

Figure 5. Data Splitting

Once divided, the text data is converted into a numeric vector using TfidfVectorizer. *Vectorizer* these are fit_transform on the training data (X_train) to learn vocabulary and convert it into a TF-IDF weight vector. Then *vectorizer* The same is used to transform test data (X_test) to be ready for evaluation.

Compare Algorithm

This section presents the results of the performance evaluation of the five machine learning algorithms tested. Comparisons were made based on the Accuracy, Precision, Recall, and F1-Score metrics to determine the most optimal algorithm for handling YouTube comment datasets that have been labeled using the lexicon method.

To ensure the validity and reproducibility of the comparative results, each algorithm is configured with specific tuning parameters as follows:

1. Support Vector Machine, uses a 'Linear' Kernel to handle TF-IDF high dimensions, with parameters random_state=42 for consistency.
2. K-Nearest Neighbors, configured with the number of closest neighbors K=5 (n_neighbors=5) as the standard of bias-variance equilibrium.
3. Naive Bayes, using the optimal MultinomialNB variant for word frequency feature data.

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i1.470
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2022

4. Random Forest, built with 100 Decision Trees (n_estimators=100) and random_state=42.
5. Decision Tree, uses a default setting of random_state=42 to control the randomness of tree formation.

A summary of the performance of all models resulting from such configurations is presented in Table 7.

Table 7. Performance Comparison of Machine Learning Algorithms

Algorithm	Accuracy	Precision	Recall	F1-Score
Support Vector Machine	96.5%	0,945	0,949	0,966
Random Forest	89.2%	0,834	0,833	0,892
Decission Tree	85.6%	0,777	0,799	0,859
K-Nearest Neighbors	84.6%	0,788	0,682	0,829
Naïve Bayes	84.0%	0,830	0,630	0,807

Based on Table 7, the comparative results show significant performance variations between algorithms. The Random Forest algorithm ranks second best with an accuracy of 89.2% and an F1-Score of 0.892, demonstrating the ensemble model's ability to handle this data quite well. On the other hand, the Decision Tree, KNN, and Naive Bayes algorithms show lower performance, with an accuracy in the range of 84-85%. In particular, significant weaknesses were seen in the Naive Bayes and KNN algorithms in the Recall metrics, which only reached 0.630 and 0.682, respectively. This low recall value indicates that both algorithms fail to identify most of the data in the minority class of Positive sentiment and tend to be biased towards the majority class.

In contrast, in this comparative study, SVM proved to be the most robust algorithm. Not only does SVM excel in accuracy (96.5%), but it also has the highest Recall value (0.949), which means that it is able to recognize Positive and Negative sentiment classes equally well, minimizing false negatives significantly compared to other models.

Confusion Matrix

The Confusion Matrix visualization in Figure 6 is used to further dissect the performance of the best model obtained, namely the Support Vector Machine (SVM) algorithm. The Confusion Matrix shows the number of true and false predictions (False Positive & False Negative) of each model.

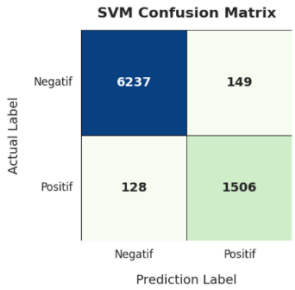

Figure 6. Confusion Matrix SVM

Performance Technically, Figure 6 confirms the robustness of SVM as the main classification model in this study. This model recorded a very minimal rate of prediction errors, with only 149 False Positives and 128 False Negative data. This low error rate validates that SVM is very precise in separating ambiguous public opinion, making it a reliable instrument for text-based social mapping.

However, behind these accuracy figures, this matrix reveals crucial social realities related to the 2025 DPR Demonstration. The data showed a sharp disparity between the negative class (6,237 comments) and the positive class (1,506 comments). The dominance of negative sentiment which reaches almost four times that of this positive sentiment indicates that the public perception of the DPR is very critical and skeptical. Sociologically, these findings confirm that YouTube has served as a space for articulating "digital emotions" where the public collectively voices distrust and rejection of legislative narratives, instead of providing support. This is in line with the general pattern of online public responses to controversial policy issues that tend to be dominated by negative comments as a form of social control.

CONCLUSIONS AND SUGGESTIONS

Conclusion

This study successfully conducted a comparative study of Machine Learning algorithms with lexicon-based labeling to classify public sentiment regarding the August 2025 DPR Demonstration, where the test results showed that the Support Vector Machine (SVM) significantly outperformed other comparator algorithms, including Random Forest, Decision Tree, KNN, and Naive Bayes, by achieving an accuracy of 96.5% and

a weighted F1-Score of 0.966. In addition to establishing SVM as the most robust model for this dataset, the study uncovered important social insights through sentiment distributions that showed overwhelming negative dominance of 32,240 comments (79.8%) and positive 8,169 comments (20.2%), an inequality that clearly reflects a deep "crisis of public trust" in the legislature and indicates that the digital public space on YouTube has evolved into a strategic forum for massive resistance and a means of criticism of the government, rather than providing support.

Suggestion

For future researchers, it is recommended to be able to try to apply this approach to datasets from different social media platforms such as X/Twitter or TikTok to test their generalizations, or combine deep learning and embedding techniques (such as Word2Vec or IndoBERT) as a substitute for TF-IDF to see the potential for improved accuracy. In addition, given the success of auto-labeling, it is also recommended to expand or fine-tune the InSet lexicon dictionary to include more specific political slang and terminology, so as to improve the quality of the automatically generated ground truths.

REFERENCES

Adi, S. I. R., Bakkara, B., Zega, K. A., Vielita, F. N., & Rakhmawati, N. A. (2024). Analisis Sentimen Masyarakat Terhadap Progress Ikn Menggunakan Model Decision Tree. *JIKA (Jurnal Informatika)*, *8*(1), 57. https://doi.org/10.31000/jika.v8i1.9803

Adriana, N. M. T. O., Suarjaya, I. M. A. D., & Githa, D. P. (2023). Analisis sentimen publik terhadap aksi demonstrasi di Indonesia menggunakan Support Vector Machine dan Random Forest. *DECODE: Jurnal Pendidikan Teknologi Informasi*, *3*(2), 257–267. https://doi.org/http://dx.doi.org/10.51454/decode.v3i2.187

Ardiansyah, A., Agustina, C., Maryani, I., & Pribadi, D. (2025). Analisis Sentimen pada Komentar YouTube terkait Pembahasan eSIM Menggunakan Metode Naive Bayes dan Random Forest. *Indonesian Journal on Software Engineering (IJSE)*, *11*(1 JUNI), 7–14. https://doi.org/10.31294/ijse.v11i1.26180

Atinna, A. N., & Akbar, M. (2025). Analisis sentimen masyarakat terhadap kebijakan Undang-Undang Tentara Nasional Indonesia (UU TNI) menggunakan Support Vector Machine. *Jurnal Komputer, Informasi Dan Teknologi*, *5*(1), 1–14. https://doi.org/https://doi.org/10.53697/jkomitek.v5i1.2603

Chamid, A. A., Nindyasari, R., Azizah, N., & Hariyadi, A. (2025). Analysis of Public Opinion on The Governor Candidate Debate Using LDA and IndoBERT. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*. https://doi.org/10.22219/kinetik.v10i3.2221

Chamid, A. A., Nindyasari, R., & Ghozali, M. I. (2025). Comparative Analysis of Machine Learning Algorithms for Predicting Patient Admission in Emergency Departments Using EHR Data. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *9*(2), 185–194. https://doi.org/10.29207/resti.v9i2.6188

Chamid, A. A., Widowati, & Kusumaningrum, R. (2022). Graph-Based Semi-Supervised Deep Learning for Indonesian Aspect-Based Sentiment Analysis. *Big Data and Cognitive Computing*, *7*(1), 5. https://doi.org/10.3390/bdcc7010005

Chamid, A. A., Widowati, & Kusumaningrum, R. (2024). Labeling Consistency Test of Multi-Label Data for Aspect and Sentiment Classification Using the Cohen Kappa Method. *Ingénierie Des Systèmes d Information*, *29*(1), 161–167. https://doi.org/10.18280/isi.290118

Chanda, S., & Pal, S. (2023). The Effect of Stopword Removal on Information Retrieval for Code-Mixed Data Obtained Via Social Media. *SN Computer Science*, *4*(5), 494. https://doi.org/10.1007/s42979-023-01942-7

Efraim, D. A., & Ermatita. (2023). Analisis Sentimen Pada Sosial Media Instagram Menggunakan Algoritma Naive Bayes (Studi Kasus : Timnas Futsal Indonesia). In *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)* (pp. 498–509). Retrieved from https://conference.upnvj.ac.id/index.php/senamika/article/view/2574

Fasnuari, D., Andrian, H., Yuana, H., & Chulkamdi, M. T. (2022). Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Penyakit Diabetes Melitus. *Antivirus : Jurnal Ilmiah Teknik Informatika*, *16*(2), 133–142. https://doi.org/10.35457/antivirus.v16i2.2445

Hakim, Z. R., & Sugiyono. (2024). Analisa Sentimen Terhadap Kereta Cepat Jakarta – Bandung

Menggunakan Algoritma Naïve Bayes Dan K-Nearest Neighbor. *Jurnal Sains Dan Teknologi*, *5*(3), 939–945. https://doi.org/10.55338/saintek.v5i3.1423

Jazuli, A., Widowati, Chamid, A. A., & Kusumaningrum, R. (2025). Transformer-based semantic indexing for aspect-based sentiment analysis using an enhanced index generation algorithm with BERT. *International Journal of Advanced Technology and Engineering Exploration*, *12*(127). https://doi.org/10.19101/IJATEE.2024.111102114

Merdiansah, R., Siska, S., & Ali Ridha, A. (2024). Analisis Sentimen Pengguna X Indonesia Terkait Kendaraan Listrik Menggunakan IndoBERT. *Jurnal Ilmu Komputer Dan Sistem Informasi (JIKOMSI)*, *7*(1), 221–228. https://doi.org/10.55338/jikomsi.v7i1.2895

Mola, S. A. S., Lete, P. R., Triyanto, T., Ajilo, B. J. A. J., & Widiastuti, T. (2024). Analisis sentimen menggunakan metode Naive Bayes dan Support Vector Machine pada kasus pelantikan artis sebagai anggota DPR RI tahun 2024. *HOAQ: Jurnal Teknologi Informasi*, *15*(1), 22–32. https://doi.org/https://doi.org/10.52972/hoaq.vol15no1.p22-32

Muhayat, T., Fauzi, A., & Indra, J. (2023). Analisis sentimen terhadap komentar video YouTube menggunakan Support Vector Machines. *Progresif: Jurnal Ilmiah Komputer*, *15*(2).

Ningsih, R. A., & Fatah, Z. (2025). Analisis sentimen komentar YouTube terhadap tragedi demo 25 Agustus menggunakan pendekatan lexicon-based. *JAMASTIKA: Jurnal Mahasiswa Teknik Informatika*, *4*(2).

Ratnaswari, S., Wibowo, N. C., & Kartika, D. S. Y. (2025). Analisis sentimen menggunakan metode lexicon-based dan support vector machine pada presiden dan wakil presiden Indonesia periode 2024–2029. *Jurnal Informatika Dan Teknik Elektro Terapan (JITET)*, *13*(1). https://doi.org/https://doi.org/10.23960/jitet.v13i1.5604

Siddiq, M. J., Jayasri, S., Suhendi, A., Hidayat, T., & Rizky, R. (2025). Analisis sentimen opini masyarakat terhadap Pilkada 2024 di media sosial Twitter menggunakan algoritma Naive Bayes. *Jurnal Informatika Dan Teknik Elektro Terapan (JITET)*, *13*(2). Retrieved from http://dx.doi.org/10.23960/jitet.v13i2.6280

Syafia, A. N., Hidayattullah, M. F., & Suteddy, W. (2023). Studi Komparasi Algoritma SVM Dan Random Forest Pada Analisis Sentimen Komentar Youtube BTS. *Jurnal Informatika: Jurnal Pengembangan IT*, *8*(3), 207–212. https://doi.org/10.30591/jpit.v8i3.5064

Syofiani, F., Alam, S., & Sulistyo, M. I. S. (2023). Analisis sentimen penilaian masyarakat terhadap childfree berdasarkan komentar di YouTube menggunakan algoritma Naive Bayes. *Jurnal Teknologi Informatika Dan Komputer MH. Thamrin*, *9*(2). https://doi.org/https://doi.org/10.37012/jtik.v9i2.1661

Umrona, R. D., Anwar, S. N., & Soelistijadi, R. (2025). Analisis sentimen komentar YouTube terkait kasus pagar laut menggunakan metode KNN (K-Nearest Neighbor). *JINTEKS: Jurnal Informatika Teknologi Dan Sains*, *7*(3), 1537–1544. https://doi.org/https://doi.org/10.51401/jinteks.v7i3.6251

Undap, M., Rantung, V. P., & Rompas, P. T. D. (2021). Analisis Sentimen Situs Pembajak Artikel Penelitian Menggunakan Metode Lexicon-Based. *Jointer - Journal of Informatics Engineering*, *2*(02), 39–46. https://doi.org/10.53682/jointer.v2i02.44

Utami, R. W., Jazuli, A., & Khotimah, T. (2021). Analisis Sentimen Terhadap Xiaomi Indonesia Menggunakan Metode Naïve Bayes. *Indonesian Journal of Technology, Informatics and Science (IJTIS)*, *3*(1), 21–30. https://doi.org/10.24176/ijtis.v3i1.7514

Utomo, W. P. (2022). Hoax and Paradox of Digital Public Sphere. *Jurnal Komunikasi Indonesia*, *11*(1). https://doi.org/10.7454/jkmi.v11i1.1024

Uyun, Q., & Qoiriah, A. (2024). Analisis sentimen opini publik terhadap program Merdeka Belajar Kampus Merdeka dengan algoritma Naive Bayes–Support Vector Machine (NBSVM). *JINACS: Journal of Informatics and Computer Science*, *6*(2).

Wibowo, I. S., Witanti, A., & Susilawati, I. (2024). Keyword Extraction Judul Berita Online Di Indonesia Menggunakan Metode TF-IDF. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *11*(1). https://doi.org/https://doi.org/https://doi.org/10.35957/jatisi.v11i1.6718