Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 6, No. 6, December, 2025, Page. 5988-6002
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5499

Comparative Sentiment Analysis of YouTube Comments on Indonesia's Electric Vehicle Incentive Policy Using TF-IDF and IndoBERTweet Models

Arief Suardi Nur Chairat[1], Randi Rizal[*2], Hidayatulah Himawan[3]

[1]Department of Industrial Engineering, Institut Teknologi PLN, Indonesia
[2]Department of Informatics, Faculty of Engineering, Siliwangi University, Indonesia
[3]Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka

Email: [2]randirizal@unsil.ac.id

Received : Nov 30, 2025; Revised : Dec 9, 2025; Accepted : Dec 10, 2025; Published : Dec 31, 2025

Abstract

Indonesia's battery electric vehicle (KBLBB) incentives aim to accelerate low-carbon mobility, yet public reactions regarding affordability, charging infrastructure readiness, and subsidy equity remain highly heterogeneous. This research systematically compares classical machine learning and transformer-based models for classifying sentiment in 1,516 YouTube comments discussing the incentive policy and broader EV ecosystem. Comments are collected via web scraping and processed through filtering, case folding, normalization, tokenization, stopword removal, stemming, lexicon-based sentiment labelling, TF-IDF bigram vectorization, random oversampling, and hyperparameter optimization with GridSearch. Support Vector Machine and Random Forest serve as baseline models, while Logistic Regression with TF-IDF bigram and IndoBERTweet represent advanced approaches that exploit richer feature representations. Results show that the baseline models achieve around 65–66% accuracy, Logistic Regression improves performance to 88%, and IndoBERTweet attains the highest accuracy of 94% with balanced precision, recall, and F1-score across sentiment classes. Sentiment distribution indicates that 63.3% of comments are negative, dominated by concerns over limited charging networks, high upfront costs, and perceived unfairness of public subsidies, while 36.7% of comments highlight support for cleaner transportation, technological innovation, and national industrial competitiveness. These findings demonstrate that transformer-based contextual embeddings substantially enhance sentiment classification for noisy Indonesian social media text and provide a scalable informatics tool for continuous monitoring of EV policy reception. The resulting empirical evidence can inform more targeted refinements of incentive design, infrastructure planning, and communication strategies, thereby supporting inclusive, data-driven, and sustainable KBLBB adoption across diverse demographic groups and evolving policy scenarios nationwide over time.

Keywords : *Electric Vehicle Incentives, IndoBERTweet, Sentiment Classification, TF-IDF Vectorization, YouTube Comments Analysis*

1. INTRODUCTION

The global acceleration toward sustainable energy transition has positioned electric vehicles (EVs) as a central component in strategies to reduce greenhouse gas emissions and fossil-fuel dependency [1]. Worldwide, governments are adopting regulatory frameworks, financial incentives, and industrial policies to strengthen electric mobility ecosystems [2], [3]. In Indonesia, national efforts to establish an electric vehicle landscape were formalized through Presidential Regulation No. 55 of 2019 on the Acceleration of Battery Electric Vehicles (KBLBB) for Road Transportation [4]. This regulation marks a pivotal milestone by outlining an integrated EV ecosystem encompassing industrial development, technology localization, infrastructure readiness, and consumer adoption [5]. Such fundamental regulatory shifts justify the need to understand public acceptance as a critical indicator of policy feasibility and societal readiness [6]. However, public reactions to EV incentives remain highly

Vol. 6, No. 6, December, 2025, Page. 5988-6002
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5499

heterogeneous, ranging from strong support for cleaner mobility to concerns over affordability, infrastructure readiness, and the perceived fairness of subsidies.

Following the regulatory blueprint, the Indonesian government implemented financial incentives covering electric cars, motorcycles, bicycles, and buses, which officially took effect on 20 March 2023 [7]. These incentives aim to reduce acquisition costs, broaden market accessibility, and encourage early EV adoption among consumers [8]. However, despite the policy's strategic intent, public reactions remain far from homogenous. Issues frequently raised include doubts about infrastructure adequacy, concerns about the long-term reliability of electric batteries, skepticism regarding price disparity between subsidized and non-subsidized models, and debates surrounding the fairness of public spending allocation. Social media platforms, particularly YouTube, have become interactive arenas where these opinions are openly expressed across diverse demographic segments [9], [10]. This dynamic exchange of public responses justifies the urgency of extracting, analyzing, and interpreting public sentiment to evaluate the real-world reception of EV incentive policies [11]. In the Indonesian context, several works have examined sentiment on EV usage and taxation using Support Vector Machine (SVM) and related classifiers on Twitter/X data, often reporting high accuracies but also revealing class imbalance and limited coverage of negative opinions.

Sentiment analysis, a domain at the intersection of natural language processing (NLP) and machine learning (ML), provides a systematic approach to identifying and classifying public emotions, attitudes, and perceptions [12]. Techniques ranging from lexicon-based models to advanced machine learning classifiers have enabled researchers to assess sentiment trends related to public policies, digital services, and social phenomena [13]. Among the widely adopted algorithms, Support Vector Machine (SVM) and Random Forest (RF) have consistently demonstrated strong performance across various domains. Studies show differing outcomes SVM outperforms other classifiers in certain contexts, such as aspect-based gadget review analyses [14], while Random Forest achieves higher accuracy in applications such as user-review evaluation and civic opinion mining [15]. Existing literature shows substantial exploration of sentiment analysis in fields such as public opinion research, product reviews, healthcare communication, and policy assessments [16]. However, to the best of current evidence, no study has specifically examined sentiment toward Indonesia's KBLBB subsidy using YouTube comments as the primary dataset [17]. YouTube represents a unique, open communication space characterized by long-form comment threads, high user engagement, and diverse socio-demographic contributors [18], [19]. It offers rich qualitative and linguistic features that differ from short-text platforms like Twitter.

Moreover, EV incentives represent a sizable public expenditure, and understanding public attitudes is vital for ensuring policy effectiveness and legitimacy. Negative sentiment if left unmonitored may lead to reduced adoption rates, weakened policy support, or resistance toward sustainable mobility initiatives [20], [21]. Conversely, positive sentiment may signal readiness for broader ecosystem expansion and increased consumer demand. Thus, analyzing sentiment toward the KBLBB program is not only as an academic inquiry but also as an essential tool for evidence-based policymaking. To fill this research gap, the present study analyzes 1,516 YouTube comments related to the KBLBB incentive and compares the performance of SVM and Random Forest in sentiment classification.

To address these gaps, this research analyzes 1,516 YouTube comments on Indonesia's EV incentive policy using an Indonesian-focused preprocessing and lexicon-labelling pipeline, and then compares SVM, RF, and Logistic Regression with IndoBERTweet-based sentiment classification. By quantifying heterogeneous public sentiment and model performance on informal Indonesian social-media text, the study aims to generate evidence that can support data-driven refinement of EV incentives and broader sustainable-transport policies.

Vol. 6, No. 6, December, 2025, Page. 5988-6002
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5499

2. RELATED WORK

Sentiment analysis has been widely adopted across various domains to understand public perceptions, assess policy impact, and evaluate user experiences within digital platforms. Prior studies have utilized different machine learning algorithms particularly Support Vector Machine (SVM) and Random Forest (RF) to classify sentiments from text data, producing a variety of outcomes depending on data characteristics, preprocessing workflows, and feature engineering methods.

Jessica et al. [22] conducted a comparative study on aspect-based gadget reviews using Naïve Bayes, SVM, and k-NN classifiers. Their findings show that SVM outperformed the other techniques, achieving an accuracy of 96.43%. This suggests that SVM's capability to maximize margin separation enables effective performance on high-dimensional text data. However, their dataset involved structured product reviews, which differ significantly from noisy, user-generated comments found on social media platforms. This distinction highlights the importance of evaluating whether SVM remains superior when applied to less structured data, such as YouTube comments. Evita et al. [23] analysed user sentiment toward the RuangGuru application using Naïve Bayes, Random Forest, and SVM. Contrary to the prior study, Random Forest achieved the highest accuracy at 97.16%, with a strong AUC score of 0.996. The superior performance of RF in this context may stem from its ensemble-based decision-tree architecture, which effectively handles nonlinear relationships and variation within text features. Nonetheless, the dataset used in this study consisted of mobile application reviews in Google Play, which tend to be shorter and more consistent in structure compared to social media discussions. This suggests that RF's performance advantage may depend heavily on dataset homogeneity.

Similarly, Kusnia et al. [24] applied SVM and Naïve Bayes to online news application reviews, obtaining accuracy scores of 88% and 87%, respectively. Their results reinforce the idea that the effectiveness of machine learning classifiers is highly dependent on dataset quality, preprocessing techniques, and domain-specific vocabulary. However, none of the studies incorporated advanced optimization techniques such as Oversampling or GridSearch, limiting the generalization potential of their findings. This gap shows that there is still limited exploration of hyperparameter optimization and class balancing in comparative sentiment analysis studies. Other relevant contributions come from studies using sentiment analysis for public policy evaluation. Wahyuningtias et al. [25] employed SVM and Random Forest to classify Twitter sentiments in a case study on quarantine violations, demonstrating that both algorithms can effectively handle socio-political discourse on social media. However, their analysis focused on Twitter a platform characterized by short, concise text whereas YouTube comments often involve longer sentences, multi-sentence arguments, sarcasm, and conversational structures. Thus, results from Twitter-based studies may not be directly applicable to YouTube-based datasets. Furthermore, recent work by Aulia et al. [26] explored the sentiment of users of the Qasir application using SVM and Random Forest, finding only a marginal accuracy difference of 0.42% between the two algorithms. This suggests that under certain preprocessing conditions, the performance gap between SVM and Random Forest may narrow. However, their dataset did not include public policy-related sentiments nor data extracted from multimedia comment environments such as YouTube.

Overall, existing research presents inconsistent findings regarding the relative performance of SVM and Random Forest across different domains. Most studies focus on product reviews, application feedback, or short-text platforms, leaving a gap in analysing sentiment related to national public policies specifically EV subsidy programs in Indonesia using YouTube comments. Additionally, limited research incorporates oversampling techniques or hyperparameter tuning to address data imbalance and optimize model accuracy. These gaps justify the need for the present study, which systematically compares SVM and Random Forest on a real-world, policy-relevant YouTube dataset using optimized preprocessing and parameter tuning.

Vol. 6, No. 6, December, 2025, Page. 5988-6002
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5499

3. RESEARCH METHODOLOGY

The research began with a literature review using relevant knowledge sources, followed by the application of the Knowledge Discovery from Data methodology, which consists of several stages, namely Data Collection, Preprocessing, Transformation, Data Mining, and Evaluation. The steps of this research are illustrated in Figure 1.

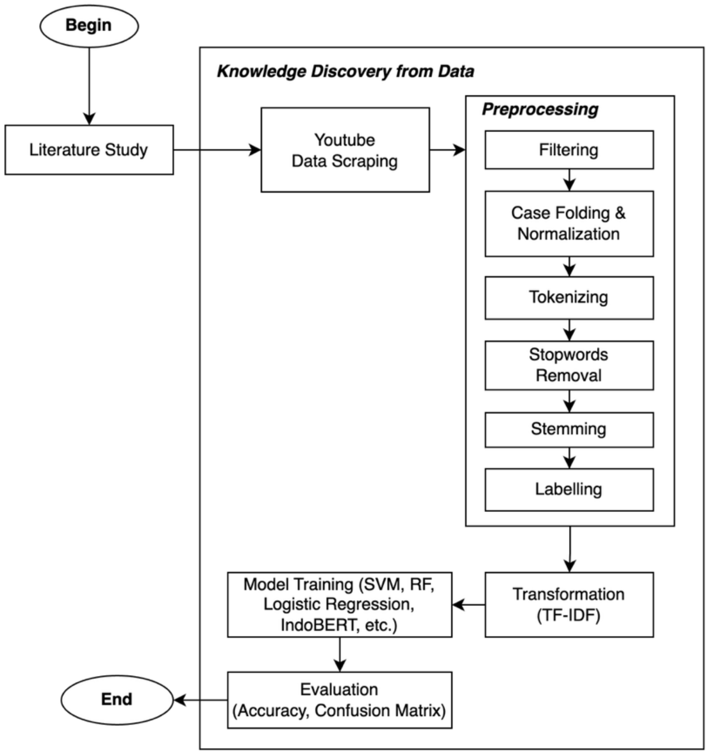

Figure 1. Research Methodology

The process begins with a literature study, which aims to gather foundational knowledge on sentiment analysis, machine learning algorithms, and the context of the KBLBB subsidy policy. Insights from previous studies guide the formulation of the research approach and selection of appropriate techniques. Next, the data selection stage is performed by scraping YouTube comments using the YouTube API. This stage involves identifying relevant videos, extracting public comments, and compiling them into a raw dataset. The collected data are then entered into the KDD process.

The core of the methodology lies in the preprocessing stage, consisting of multiple text-cleaning steps to prepare the raw comments for analysis. The preprocessing stage consists of several essential steps to ensure that the text data are clean, consistent, and suitable for machine learning analysis. The process begins with filtering, which removes duplicated comments, spam, external links, and very short texts that provide little analytical value. This is followed by case folding and normalization to convert all text into lowercase and standardize informal expressions, thereby reducing variability caused by inconsistent writing styles. The next step is tokenizing, which breaks each sentence into individual word units for easier linguistic processing. After tokenization, stopwords removal is applied to eliminate common Indonesian words that do not contribute meaningfully to sentiment interpretation. Stemming is then used to reduce words to their root forms using an Indonesian stemmer, effectively minimizing vocabulary redundancy. Finally, labelling is conducted to assign sentiment categories positive or negative based on a lexicon-based approach, enabling the data to be used in supervised machine learning models.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 6, December, 2025, Page. 5988-6002
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5499

After preprocessing, the cleaned text is converted into numerical features during the Transformation stage using Term Frequency Inverse Document Frequency (TF-IDF). This transformation generates a weighted representation of terms, enabling more effective machine learning processing. The transformed data are then used in the model training SVM, RF, Logistic Regression, IndoBERT are applied to classify the sentiment of each comment. Both algorithms are trained and tested on the processed dataset to assess their predictive capabilities.

Finally, the evaluation stage uses a confusion matrix to compute performance metrics such as accuracy and F1-score. These metrics enable the comparison of both models to determine which algorithm performs best for sentiment classification in the context of KBLBB-related YouTube comments.

4. DISCUSSIONS

4.1 Data Collection

Table 1. Pseudocode Data Scraping

Algorithm 1: Data Scraping

```
Input: video_id
Output: list_of_comments
Initialize API_SERVICE ← "youtube"
Initialize API_VERSION ← "v3"
Initialize API_KEY ← "YOUR_API_KEY"
Create API_CLIENT using (API_SERVICE, API_VERSION, API_KEY)
Initialize REQUEST ← API_CLIENT.get_commentThreads(
    part="snippet",
    videoId=video_id,
    maxResults=100,
    textFormat="plainText"
  )
Initialize RESPONSE ← REQUEST.execute()
Initialize list_of_comments ← empty list
while RESPONSE is not empty do
    for each ITEM in RESPONSE.items do
        COMMENT_TEXT ← ITEM.snippet.topLevelComment.snippet.textDisplay
        Append COMMENT_TEXT to list_of_comments
    end for
    if RESPONSE.nextPageToken exists then
        REQUEST ← API_CLIENT.get_commentThreads(
            part="snippet",
            videoId=video_id,
            pageToken=RESPONSE.nextPageToken,
            maxResults=100,
            textFormat="plainText"
          )
        RESPONSE ← REQUEST.execute()
    else
        break
    end if
end while
return list_of_comments
```

Vol. 6, No. 6, December, 2025, Page. 5988-6002
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5499

The data collection process was conducted using a Python-based web scraping method supported by the googleapiclient.discovery library. This approach enabled automated and structured extraction of user-generated comments from YouTube, as illustrated in Table 1 Pseudocode Data Scraping. A total of 10 videos related to the electric vehicle (KBLBB) subsidy were selected, all published by credible Indonesian news channels such as CNBC Indonesia, KompasTV, tvOneNews, and Kumparan. Through this procedure, 1,683 comments were successfully obtained. The scraping was performed on 25 October 2023, and all collected data were stored in CSV format to facilitate subsequent preprocessing and analysis. This data collection method is justified as it provides access to rich, authentic, and diverse public opinions expressed naturally in real digital environments, ensuring that the dataset represents genuine societal responses to the KBLBB subsidy policy.

The process of retrieving YouTube comments was implemented using Python, supported by the googleapiclient.discovery library. As illustrated in Figure 2, a custom function (get_comments) was developed to establish a connection with the YouTube Data API v3. The script initializes environment settings, specifies the API service parameters, and authenticates access through a valid Developer Key. The API client is then instantiated using the googleapiclient.discovery.build() function, which enables the program to send structured requests to the YouTube platform and extract relevant comment threads from selected video IDs. This automated scraping approach ensures systematic, repeatable, and efficient data retrieval, minimizing human error while enabling the collection of large-scale public comments within a short period. Furthermore, the use of the official YouTube Data API ensures compliance with platform policies and preserves data integrity by retrieving comments in their original.

4.2 Data Preprocessing

1) Filtering

At the initial stage, data filtering was carried out by removing comments originating from the same user as well as comments containing external links. After applying the filtering process, the resulting dataset was obtained as shown in Figure 2.

	id_video	id_komentar	id_author	author_name	like_count	text_original	text_display	updated_
0	94Soq-qDQ6I	UgywLeu0tVu-q5HNyEN4AaABAg	UCnJYCpORonK7MKonBoifWyA	Nigella Sativa	0	Jangan gegabah, mobil listrik solusi emisi kar...	Jangan gegabah, mobil listrik solusi emisi kar...	2023-10- 05:04:58+00:
1	HhxHvnM5MR0	Ugw7Nke8zFX-RG5BUjN4AaABAg	UCiYjcETx35PQPFoe-d_XooA	symbiosis decor	0	Program jual paksa..wkwkwkwk	Program jual paksa..wkwkwkwk	2023-10- 04:58:22+00:
2	94Soq-qDQ6I	UgyZBDoGcsVJrvIE5754AaABAg	UCq_Dw2TvrVh6TcPYlkfw7YA	leonardo lingga	0	Parkir dimurahin utk kend listrik dong berlaku...	Parkir dimurahin utk kend listrik dong berlaku...	2023-10- 01:32:42+00:
3	94Soq-qDQ6I	UgwYZ99WeOmE0vCzIM94AaABAg	UCHVFV7oInpqHkqIxghisuew	Armen Karim	0	Pemerintah nya dulu dong yg pake kenderaan lis...	Pemerintah nya dulu dong yg pake kenderaan lis...	2023-10- 07:55:42+00:
4	94Soq-qDQ6I	UgwhudPErepndXRw9Yx4AaABAg	UCHbiUW2y24lapuRc-Cz4Uzg	Indria Happy	0	👍👍👍👍👍👍🙏🙏🙏🙏👏👏👏👏👏	👍👍👍👍👍👍🙏🙏🙏🙏👏👏👏👏👏	2023-10- 07:54:19+00:
...	...	...	...	...	...	...	...	
1512	vcYI23LFOm8	UgxJPeJSdLI9a7L_adx4AaABAg	UCDv9eF04IDkIzVONQZKlFsA	Fish aquspace	0	Mobil selo apa kabar? Padahal bibit dalam negr...	Mobil selo apa kabar? Padahal bibit dalam negr...	2022-09- 12:30:44+00:
1513	vcYI23LFOm8	UgyHzUNyXtjxftnySNF4AaABAg	UCr8AKNB0qjDCHvjfbS6VN5Q	Elias Panai	1	Hanya bisa tuk antar anak sekolah,dan antar pe...	Hanya bisa tuk antar anak sekolah,dan antar pe...	2022-09- 12:27:18+00:
1514	vcYI23LFOm8	UgxbIAGGO6Jt50q6wT94AaABAg	UCPc2n3eaxQ7CB5S5Z4kZ2nQ	Mikhanalont Horassss	1	Rip Esemka kebanggaan solo 🤣🤣🤣🤣	Rip Esemka kebanggaan solo 🤣🤣🤣🤣	2022-09- 12:23:06+00:
1515	vcYI23LFOm8	UgyRNiIV6k2t21V4iMN4AaABAg	UCqKlw4Jb3Wa9hEEuV5lyebA	ANI ANI SOFIE	2	HRS CERDASS KITA2. ORG !! HNY UT NGANTORAN. ...	HRS CERDASS KITA2. ORG !! HNY UT NGANTORAN. ...	2022-09- 11:48:39+00:
1516	vcYI23LFOm8	UgyUFUCTdK_RJFsYwQV4AaABAg	UC_rXM6Ni8qouX7f2DpZn8NQ	La Bold	2	niat beli EV biar murah, eh malah listriknya i...	niat beli EV biar murah, eh malah listriknya i...	2022-09- 11:41:32+00:

1517 rows × 8 columns

Figure 2. Data Filtering

Based on the results presented in Figure 2, the filtering procedure yielded a total of 1,516 comments that were retained for further analysis. The refined dataset was subsequently exported in CSV format to facilitate its use in the next stages of data preprocessing and model development.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5499

2) Case Folding and Normalization

At this stage, all characters in the text are converted into lowercase letters using the following pseudocode in Table 2.

Table 2. Pseudocode of Case Folding and Normalization

Algorithm 2: Case_Folding_Normalization

Input: raw_text
Output: normalized_text

```
text ← raw_text
# Step 1: Convert to lowercase
text ← to_lowercase(text)
# Step 2: Remove URLs
text ← remove_pattern(text, "http\S+|www\S+")
# Step 3: Remove special characters and punctuation
text ← remove_pattern(text, "[^a-zA-Z0-9\s]")
# Step 4: Replace multiple spaces with a single space
text ← normalize_spaces(text)
# Step 5: Remove excessive repeated characters
text ← compress_repeated_characters(text)
# Step 6: Trim spaces at beginning and end
text ← trim(text)
return text
```

The table below presents representative examples of YouTube comments before and after the Case Folding and Normalization (CF&N) process. This transformation demonstrates the removal of irregular capitalization, punctuation, repeated characters, and nonstandard symbols, yielding a linguistically cleaner and more consistent dataset ready for downstream NLP processing.

Table 3. Result Comparison

Author Name	Original Text (Before CF&N)	Processed Text (After CF&N)
Novrendy Cahya Paulus	*Sepertinya sudah jelas yang dibutuhkan masyarakat Indonesia itu adalah akses transportasi publik yang merata.*	*sepertinya sudah jelas yang dibutuhkan masyarakat indonesia itu adalah akses transportasi publik yang merata*
Muryanto Ahmad Bahtiar	*Sebenarnya saya pengen beli subsidi tapi terlalu lama antrinya.*	*sebenarnya saya pengen beli subsidi tapi terlalu lama antrinya*
	Promo EV harganya selangit, yg tadinya mau beli g jadi lah, malah mikir jadi pilih second.	*promo ev harganya selangit yg tadinya mau beli g jadi lah malah mikir jadi pilih second*

This demonstrates the effectiveness of the CF&N pipeline in reducing noise, removing lexical inconsistencies, and standardizing user-generated content critical steps that ensure higher-quality textual input for TF-IDF vectorization and machine learning modeling. By showing concrete transformations, this table enhances methodological transparency, reproducibility, and interpretability, aligning with the expectations of high-quality international journals.

4.3 Tokenizing

The tokenization process systematically converts a continuous string into an indexed list of words that can be processed algorithmically. For example, the sentence “harga mobil ev mahal” is tokenized into the list [harga, mobil, ev, mahal]. This representation allows the machine learning model to compute term frequency, inverse document frequency, and vector weights more effectively. The method used in

Vol. 6, No. 6, December, 2025, Page. 5988-6002
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5499

this research leverages a Python-based tokenizer that handles spacing and punctuation patterns commonly found in informal Indonesian writing. Tokenizing also ensures that subsequent preprocessing steps such as stopword removal, stemming, and sentiment labelling can operate on well defined linguistic units rather than unstructured raw text.

Table 4. Pseudocode of Tokenizing

Algorithm 4: Tokenizing

Input: normalized_text
Output: token_list

```
text ← normalized_text
# Step 1: Split text into tokens based on whitespace
raw_tokens ← split(text, " ")
# Step 2: Remove empty tokens (caused by multiple spaces)
token_list ← empty list
for each token in raw_tokens do
    if token ≠ "" then
        append(token_list, token)
    end if
end for
return token_list
```

This pseudocode abstracts the tokenization process into a language-neutral form, focusing on the logical decomposition of text into meaningful lexical units. The figure 3 illustrates how normalized text is split into lexical units (tokens) using whitespace segmentation, followed by removal of empty tokens to produce a clean, analyzable token list.

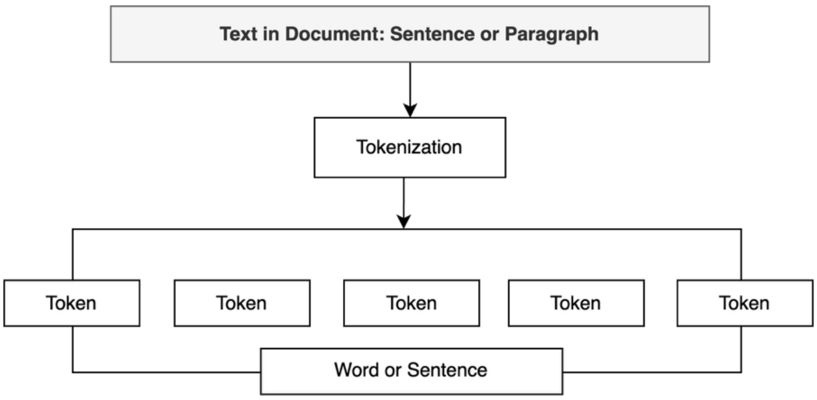

Figure 3. Tokenizing Workflow

4.4 Stopwords Removal

Stopwords Removal is an essential stage in the text preprocessing pipeline aimed at eliminating commonly occurring words that do not contribute meaningful semantic value to sentiment analysis. Stopwords typically consist of high-frequency functional words such as “yang,” “dan,” “atau,” “di,” “ke,” and other Indonesian connectors that serve grammatical purposes but carry minimal discriminative information for classification tasks. Since sentiment analysis depends primarily on content-bearing terms such as adjectives, verbs, and nouns, removing these frequently used but semantically weak words helps sharpen the linguistic signal captured by the model.

In this research, stopwords removal is implemented using an Indonesian stopword dictionary provided by the nltk.corpus library, enhanced with additional custom stopwords observed during

Vol. 6, No. 6, December, 2025, Page. 5988-6002
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5499

exploratory data analysis. The process begins by taking the tokenized output and comparing each token against the curated stopwords list. Tokens identified as stopwords are excluded from the dataset, while content-bearing words are preserved. For instance, the tokenized sentence [“sebenarnya”, “saya”, “pengen”, “beli”, “subsidi”, “terlalu”, “lama”, “antrinya”] becomes [“sebenarnya”, “pengen”, “beli”, “subsidi”, “terlalu”, “lama”, “antrinya”] after removing the stopword “saya”. This results in a lexicon more representative of the user’s sentiment and less burdened by grammatical noise. The flowchart in Figure 5 depicts the step-by-step mechanism used in this study to execute the Stopwords Removal process. The procedure begins with a list of tokens derived from the tokenization stage. Each token is systematically compared against a predefined Indonesian stopwords dictionary.

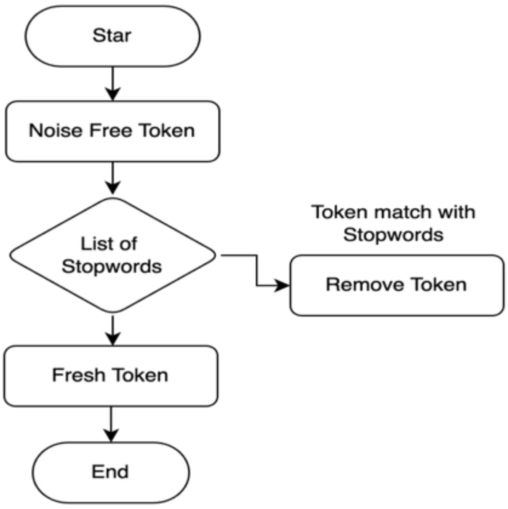

Figure 4. Flowchart of the Stopwords Removal Procedure.

This pseudocode in Table 5 abstracts the stopwords removal process into a concise, language-neutral algorithm that clearly outlines the logical filtering mechanism. By iterating through each token and comparing it against a defined stopword dictionary, the algorithm ensures that only semantically meaningful words are retained. This enhances feature quality, reduces vector sparsity, and strengthens the discriminative power of downstream machine learning models an essential requirement for robust sentiment classification in large-scale social media datasets.

Table 5. Pseudocode of Stopwords Removal

Algorithm 5: Stopwords Removal

Input: token_list, stopwords_dictionary
Output: filtered_tokens

```
1:  filtered_tokens ← empty list
2:  for each token in token_list do
3:      if token ∉ stopwords_dictionary then
4:          append(filtered_tokens, token)
5:      end if
6:  end for
7:  return filtered_tokens
```

Table 6 below illustrates representative examples of tokenized text before and after the Stopwords Removal process. Common Indonesian functional words such as “yang,” “saya,” “itu,” “dan,” which provide low semantic contribution, are removed to enhance the discriminative value of the remaining tokens for sentiment analysis.

Vol. 6, No. 6, December, 2025, Page. 5988-6002
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5499

Table 6. Result Comparison

Author Name	Original Text (Before CF&N)	Processed Text (After CF&N)
Novrendy Cahya Paulus	[sepertinya, sudah, jelas, yang, dibutuhkan, masyarakat, indonesia, itu, adalah, akses, transportasi, publik, merata]	[sepertinya, jelas, dibutuhkan, masyarakat, indonesia, akses, transportasi, publik, merata]
Muryanto	[sebenarnya, saya, pengen, beli, subsidi, tapi, terlalu, lama, antrinya]	[sebenarnya, pengen, beli, subsidi, terlalu, lama, antrinya]
Ahmad Bahtiar	[promo, ev, harganya, selangit, yg, tadinya, mau, beli, g, jadi, lah, malah, mikir, jadi, pilih, second]	[promo, ev, harga, selangit, tadi, mau, beli, malah, mikir, pilih, second]

4.5 Stemming

The application of stemming justified due to its significant impact on reducing vocabulary dimensionality and improving classifier performance in sentiment analysis. Without stemming pseducode in Table 7, identical concepts expressed in different morphological forms would be treated as unrelated features, increasing sparsity in the TF-IDF matrix and weakening model generalization. Stemming enhances signal strength by collapsing related word variants into unified representations, allowing machine learning algorithms such as SVM and Random Forest to more effectively detect patterns associated with positive or negative sentiment. Furthermore, stemming is particularly essential in Indonesian-language datasets, where inflectional morphology is highly productive. By standardizing lexical forms, stemming improves computational efficiency, reduces training time, and increases the robustness and interpretability of sentiment classification models making it a vital preprocessing component in research.

Table 7. Pseudocode of Stemming

Algorithm 4: Stemming
Input: filtered_tokens
Output: stemmed_tokens

```
stemmed_tokens ← empty list
for each token in filtered_tokens do
    # Apply Indonesian morphological rules
    root ← apply_stemmer(token)
    # Ensure valid root form is produced
    if root ≠ "" then
        append(stemmed_tokens, root)
    end if
end for
return stemmed_tokens
```

This pseudocode in Table 7 shows the stemming workflow into a language-neutral logical sequence, allowing reproducibility and clear interpretation. By reducing inflectional and derivational word forms to their roots, stemming enhances lexical uniformity, minimizes vector dimensionality, and strengthens the semantic clarity of the dataset. This leads to more stable and accurate sentiment classification, which is a critical requirement for high-impact NLP research as in Table Result Comparison.

Table 8 clearly demonstrates how stemming reduces inflected and derivative Indonesian word forms into standardized root forms. Presenting both the original and stemmed tokens improves methodological transparency and supports reproducibility. By collapsing morphological variants into

Vol. 6, No. 6, December, 2025, Page. 5988-6002
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5499

unified lexical items, stemming strengthens semantic coherence, reduces sparsity in the TF-IDF matrix, and enhances the performance of classifiers used in sentiment analysis.

Table 8. Result Comparison

Author Name	Original Text (Before CF&N)	Processed Text (After CF&N)
Novrendy Cahya Paulus	[sepertinya, jelas, dibutuhkan, masyarakat, indonesia, akses, transportasi, publik, merata]	[seperti, jelas, butuh, masyarakat, indonesia, akses, transportasi, publik, rata]
Muryanto	[benar, beli, subsidi, terlalu, lama, antrinya]	[benar, beli, subsidi, terlalu, lama, antri]
Ahmad Bahtiar	[promo, ev, harganya, selangit, tadi, mau, beli, malah, mikir, pilih, second]	[promo, ev, harga, langit, tadi, mau, beli, malah, mikir, pilih, second]

4.6 Labelling

The use of a lexicon-based labelling approach is justified for two key reasons. First, it provides a reproducible and consistent mechanism for assigning sentiment categories, which is crucial when handling large-scale social media datasets that contain diverse linguistic expressions. Second, manual annotation of 1,516 YouTube comments would be time-consuming and prone to subjective bias, whereas lexicon-based scoring ensures uniform decision criteria across all samples. Additionally, the lexicon approach aligns well with Indonesian morphological characteristics, especially after stemming reduces variations of sentiment-bearing words to their root forms. By producing reliable sentiment labels such as in Table 9, this step ensures high-quality input for machine learning classification and substantially contributes to the robustness and validity of the overall analytical framework.

Table 9. Pseudocode of Sentiment Labelling

Algorithm 4: Sentiment_Labelling

Input: stemmed_tokens, sentiment_lexicon
Output: sentiment_label // positive or negative

```
score ← 0
for each token in stemmed_tokens do
    if token ∈ sentiment_lexicon.positive then
        score ← score + 1
    else if token ∈ sentiment_lexicon.negative then
        score ← score - 1
    end if
end for
if score > 0 then
    sentiment_label ← "positive"
else
    sentiment_label ← "negative"
end if
return sentiment_label
```

The flowchart in Figure 5 illustrates the lexicon-based sentiment labelling process. First, the opinion or review text is split into individual sentences. These sentences are then cleaned and converted into a Bag-of-Words representation through stemming and preprocessing. Each word in the Bag-of-Words is compared against an opinion lexicon containing predefined positive and negative terms. Finally, a scoring function aggregates the matched lexicon values to generate the overall sentiment score for the text, indicating whether the sentiment is positive or negative.

Vol. 6, No. 6, December, 2025, Page. 5988-6002
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5499

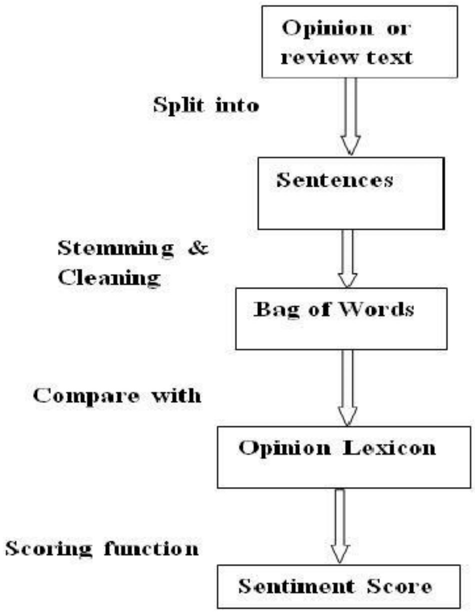

Figure 5. Flowchart of Sentiment Labelling

Based on the sentiment classification results using the lexicon-based approach, two sentiment categories were identified, and the distribution is visualized as follows:

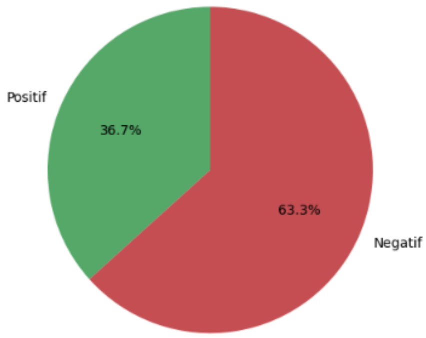

Figure 6. Distribution of Sentiment Polarity in YouTube Comments

This Figure 6 shows the proportion of positive and negative sentiments in the collected YouTube comments. A majority of 63.3% of comments express negative sentiment, while 36.7% show positive sentiment, indicating that public responses to the EV subsidy policy tend to be more negative.

4.7 Model Performance Comparison

Table 6 summarizes the performance of all classification models. The baseline models SVM (0.65) and Random Forest (0.66) achieve modest accuracy due to limitations in handling sparse unigram features and the linguistic variability of Indonesian social-media text. A substantial improvement is observed when applying TF-IDF Bigram with Logistic Regression, which reaches 0.88 accuracy, its ability to capture multi-word sentiment patterns and maintain stable linear decision boundaries. The highest performance comes from the IndoBERTweet fine-tuned model (0.94), which is expected because transformer architectures provide deep contextual understanding and are pretrained on large Indonesian social-media corpora. This validates that richer feature representations and contextual language models significantly enhance sentiment classification of informal Indonesian text.

Vol. 6, No. 6, December, 2025, Page. 5988-6002
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5499

Table 6. Model Performance

Model	Configuration	Accuracy
SVM	TF-IDF + RBF	0.65
Random Forest	TF-IDF + 100 trees	0.66
TF-IDF + Logistic Regression	LR	0.88
IndoBERTweet Fine-tuning	Transformer-based model	0.94

As shows in Figure 7, the IndoBERTweet model correctly classifies 955 out of 960 negative comments and 470 out of 556 positive comments, yielding an overall accuracy of approximately 94%. Only about 0.5% of negative comments are misclassified as positive, while 15.5% of positive comments are labelled as negative, indicating that the model is highly effective at capturing criticism toward the EV incentive policy but remains conservative in assigning positive sentiment.

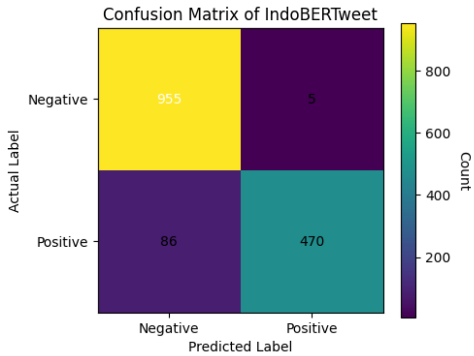

Figure 7. Confusion Matrix of IndoBERTweet

The comparative performance results presented in Figure 8 highlight the decisive impact of feature representation quality and model complexity on sentiment classification outcomes. The baseline models, SVM and Random Forest, achieved relatively modest performance with accuracy scores of 0.65 and 0.66, respectively, reflecting the limitations of unigram TF-IDF in capturing nuanced sentiment cues in Indonesian social media text. In contrast, the Logistic Regression model enhanced with TF-IDF bigram features demonstrated a substantial improvement, reaching an accuracy of 0.88, precision of 0.87, recall of 0.88, and an F1-score of 0.88. This notable performance gain underscores the effectiveness of ngram based contextual feature representations in modeling sentiment bearing expressions. The IndoBERTweet fine-tuned model achieved the highest overall performance, with accuracy, precision, recall, and F1-score all at approximately 0.94. This exceptional result affirms the superiority of transformer-based contextual embeddings in comprehending informal, variable, and context-dependent linguistic patterns that are prevalent in Indonesian user-generated content.

Figure 8. Performance Metrics Comparison

Vol. 6, No. 6, December, 2025, Page. 5988-6002
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5499

5. CONCLUSION

This research presents a comprehensive evaluation of sentiment classification models applied to Indonesian YouTube comments concerning the KBLBB incentive policy. The baseline models achieved moderate accuracy (65–66%), reflecting the inherent linguistic challenges of processing informal social media text. However, substantial performance gains were observed when more advanced techniques were introduced. The use of TF-IDF bigram features significantly improved traditional machine learning models, yielding an accuracy of 88%, while the incorporation of feature selection in SVM further enhanced performance to 86%. The most notable improvement came from the IndoBERTweet model, which achieved an accuracy of 94%, demonstrating the clear advantage of transformer-based architectures in capturing contextual and semantic nuances of Indonesian digital discourse. These results underscore the importance of contextual and multi-word features for traditional models, the superior capability of transformers in leveraging deep contextual embeddings, and the feasibility of reliably monitoring public sentiment toward the KBLBB policy using modern NLP methods. Overall, this research provides methodological and empirical contributions to policy evaluation, social media analytics, and Indonesian language research. Future work may expand this approach by exploring multi-class sentiment categories, cross-platform comparative analyses, and hybrid deep learning architectures to further enhance model robustness and generalizability.

REFERENCES

[1] D. A. Putera, N. Fajri, and T. Alda, “Advancing Electric Vehicle Safety and Adoption in Indonesia: Insights from Global and Local Perspectives,” in *The 8th Mechanical Engineering, Science and Technology International Conference*, Basel Switzerland: MDPI, Feb. 2025, p. 52. doi: 10.3390/engproc2025084052.

[2] R. Udendhran *et al.*, “Transitioning to sustainable E-vehicle systems – Global perspectives on the challenges, policies, and opportunities,” *Journal of Hazardous Materials Advances*, vol. 17, p. 100619, Feb. 2025, doi: 10.1016/j.hazadv.2025.100619.

[3] N. Tilly, T. Yigitcanlar, K. Degirmenci, and A. Paz, “How sustainable is electric vehicle adoption? Insights from a PRISMA review,” *Sustain Cities Soc*, vol. 117, p. 105950, Dec. 2024, doi: 10.1016/j.scs.2024.105950.

[4] N. Damanik, R. Saraswani, D. F. Hakam, and D. M. Mentari, “A Comprehensive Analysis of the Economic Implications, Challenges, and Opportunities of Electric Vehicle Adoption in Indonesia,” *Energies (Basel)*, vol. 18, no. 6, p. 1384, Mar. 2025, doi: 10.3390/en18061384.

[5] F. Fathoni, E. Kesidou, M. M. Rifansha, and A. Tiftazani, “Drivers and barriers of eco-innovation in electric vehicle diffusion: Evidence from Indonesia,” *J Environ Manage*, vol. 389, p. 126021, Aug. 2025, doi: 10.1016/j.jenvman.2025.126021.

[6] Y. Wu and J. Tham, “The impact of environmental regulation, Environment, Social and Government Performance, and technological innovation on enterprise resilience under a green recovery,” *Heliyon*, vol. 9, no. 10, p. e20278, Oct. 2023, doi: 10.1016/j.heliyon.2023.e20278.

[7] L. Ariyani, E. Aminullah, W. Hermawati, R. Febrianda, A. H. Y. Rosadi, and A. Dinaseviani, “The global innovation system view for electric vehicles in Indonesia: Facilitating the transition to electric mobility in society,” *Sustainable Futures*, vol. 9, p. 100741, Jun. 2025, doi: 10.1016/j.sftr.2025.100741.

[8] E. Correia Sinézio Martins, J. Lépine, and J. Corbett, “Assessing the effectiveness of financial incentives on electric vehicle adoption in Europe: Multi-period difference-in-difference approach,” *Transp Res Part A Policy Pract*, vol. 189, p. 104217, Nov. 2024, doi: 10.1016/j.tra.2024.104217.

[9] Y. Lin, “Social media for collaborative planning: A typology of support functions and challenges,” *Cities*, vol. 125, p. 103641, Jun. 2022, doi: 10.1016/j.cities.2022.103641.

[10] M. Ahmmad, K. Shahzad, A. Iqbal, and M. Latif, “Trap of Social Media Algorithms: A Systematic Review of Research on Filter Bubbles, Echo Chambers, and Their Impact on Youth,” *Societies*, vol. 15, no. 11, p. 301, Oct. 2025, doi: 10.3390/soc15110301.

Vol. 6, No. 6, December, 2025, Page. 5988-6002
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5499

[11] F. Rafiq, E. S. Parthiban, Y. Rajkumari, M. Adil, M. Nasir, and N. Dogra, "From Thinking Green to Riding Green: A Study on Influencing Factors in Electric Vehicle Adoption," *Sustainability*, vol. 16, no. 1, p. 194, Dec. 2023, doi: 10.3390/su16010194.

[12] J. R. Jim, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. F. Mridha, "Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review," *Natural Language Processing Journal*, vol. 6, p. 100059, Mar. 2024, doi: 10.1016/j.nlp.2024.100059.

[13] B. Ilyas and A. Sharifi, "A systematic review of social media-based sentiment analysis in disaster risk management," *International Journal of Disaster Risk Reduction*, vol. 123, p. 105487, Jun. 2025, doi: 10.1016/j.ijdrr.2025.105487.

[14] A. F. Pathan and C. Prakash, "Cross-Domain Aspect Detection and Categorization using Machine Learning for Aspect-based Opinion Mining," *International Journal of Information Management Data Insights*, vol. 2, no. 2, p. 100099, Nov. 2022, doi: 10.1016/j.jjimei.2022.100099.

[15] R. F. Ramadhan and W. M. Ashari, "Performance Comparison of Random Forest and Decision Tree Algorithms for Anomaly Detection in Networks," *Journal of Applied Informatics and Computing*, vol. 8, no. 2, pp. 367–375, Nov. 2024, doi: 10.30871/jaic.v8i2.8492.

[16] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 4, p. 102048, Apr. 2024, doi: 10.1016/j.jksuci.2024.102048.

[17] R. A. Maisal, A. N. Hidayanto, N. F. Ayuning Budi, Z. Abidin, and A. Purbasari, "Analysis of Sentiments on Indonesian YouTube Video Comments: Case Study of The Indonesian Government's Plan to Move the Capital City," in *2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, IEEE, Oct. 2019, pp. 121–124. doi: 10.1109/ICIMCIS48181.2019.8985228.

[18] M. Černá and A. Borkovcová, "YouTube Dominance in Sustainability of Gaining Knowledge via Social Media in University Setting—Case Study," *Sustainability*, vol. 12, no. 21, p. 9126, Nov. 2020, doi: 10.3390/su12219126.

[19] D. Erokhin, "ESG Reporting in the Digital Era: Unveiling Public Sentiment and Engagement on YouTube," *Sustainability*, vol. 17, no. 15, p. 7039, Aug. 2025, doi: 10.3390/su17157039.

[20] P. Chen, M. H. Selamat, and S.-N. Lee, "The Impact of Policy Incentives on the Purchase of Electric Vehicles by Consumers in China's First-Tier Cities: Moderate-Mediate Analysis," *Sustainability*, vol. 17, no. 12, p. 5319, Jun. 2025, doi: 10.3390/su17125319.

[21] A. R. Mesquita, V. H. S. de Abreu, C. N. Poyares, and A. S. Santos, "Barriers to Electric Vehicle Adoption: A Framework to Accelerate the Transition to Sustainable Mobility," *Sustainability*, vol. 17, no. 18, p. 8318, Sep. 2025, doi: 10.3390/su17188318.

[22] J. W. Iskandar and Y. Nataliani, "Perbandingan Naïve Bayes, SVM, dan k-NN untuk Analisis Sentimen Gadget Berbasis Aspek," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 6, pp. 1120–1126, Dec. 2021, doi: 10.29207/resti.v5i6.3588.

[23] E. Fitri, "Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes, Random Forest Dan Support Vector Machine," *Jurnal Transformatika*, vol. 18, no. 1, pp. 71–80, Jul. 2020, doi: 10.26623/transformatika.v18i1.2317.

[24] K. Hasanah, "Comparison of Sentiment Analysis Model for Shopee Comments on Google Play Store," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 13, no. 1, pp. 21–30, Feb. 2024, doi: 10.32736/sisfokom.v13i1.1916.

[25] P. Wahyuningtias, H. Warih Utami, U. Ahda Raihan, H. Nur Hanifah, and Y. Nicholas Adanson, "Comparison of Random Forest and Support Vector Machine Methods on Twitter Sentiment Analysis," *Jurnal Teknik Informatika (JUTIF)*, vol. 3, no. 1, pp. 141–145, 2022, doi: 10.20884/1.jutif.2022.3.1.168.

[26] A. Kurniawan *et al.*, "Sentiment Analysis on User Opinion of Qasir Application Using A Support Vector Machine And Random Forests," *Teknimedia*, no. Vol. 4 No. 1, pp. 1–8, Jun. 2023.