Journal of Soft Computing Exploration
Homepage: htttps://shmpublisher.com/index.php/joscex
p-ISSN: 2746-7686
e-ISSN: 2746-0991

.

Web-based sentiment analysis of environmental issues on social media X: A comparison of svm and random forest

Subandi[1], Agus Setiyo Budi Nugroho[2], Politeknik Hasnur[3]
[1]Department of Smart City Information System, Politeknik Negeri Banjarmasin, Indonesia
[2]Department of Computer Science, Politeknik Negeri Banjarmasin, Indonesia
[3] Department of Multimedia Engineering Technology, Hasnur Polytechnic, Indonesia

Article Info

Article history:

Received April 17, 2026
Revised May 5, 2026
Accepted May 5, 2026

Keywords:

Sentiment analysis
Support vector machine
Random forest
Social media X
Machine learning

ABSTRACT

This study aims to compare the performance of Support Vector Machine (SVM) and Random Forest (RF) in classifying public sentiment toward environmental issues on social media X (formerly Twitter) and to develop a web-based system for sentiment monitoring and visualization. A total of 47,245 tweets from 2021–2025 were collected using 24 environmental keywords. The data were processed through text cleaning, tokenization, stopword removal, and stemming. Sentiment labeling was performed automatically using a lexicon-based approach with the InSet dictionary, resulting in positive, negative, and neutral classes. After filtering, 13,063 tweets were used for model training. Classification employed TF-IDF features and 5-fold cross-validation. The results indicate that SVM outperformed RF with an accuracy of 83%, compared to 81%. Both models performed well in identifying sentiment polarity, although challenges remain in classifying neutral sentiment. The novelty of this study lies in integrating lexicon-based labeling with machine learning and implementing it in a web-based system for automated analysis and visualization. Practically, this system supports stakeholders in monitoring public opinion and enables data-driven decision-making in environmental policy and management.

Corresponding Author:

Muhammad Hidayat,
Department of Multimedia Engineering Technology,
Politeknik Hasnur, Kalimantan Selatan, Indonesia
Email: hidayatsyahid117@gmail.com
https://doi.org/10.52465/joscex.v7i2.55

1. INTRODUCTION

Environmental issues are one of the global problems that continue to evolve and are a serious concern for the public, particularly in Indonesia. Various problems such as water pollution, air pollution, waste management, forest fires, and floods not only impact the ecosystem but also affect human quality of life. These phenomena often trigger broad and dynamic public responses in the digital space, especially through social media. Social media, particularly platform X (formerly Twitter), has become a primary means for people to express opinions, criticism, and support regarding various environmental issues. The real-time, open, and text-based characteristics of the platform make X a rich data source for accurately capturing public perception and

emotion. This aligns with research showing that social media plays an important role as an indicator of public opinion on social and environmental issues [1].

Nevertheless, the high volume of data generated from social media poses its own challenges in the analysis process. Given the large volume of unstructured data, a manual approach becomes inefficient in terms of time and resources. Therefore, an automated approach capable of processing large amounts of data quickly and accurately is needed. In this context, text mining and machine learning techniques[2] become relevant solutions for extracting meaningful information from unstructured text data [3].

One widely used method in public opinion processing is sentiment analysis, which is the process of classifying text into sentiment categories such as positive, negative, and neutral [4]. This analysis enables systematic identification of public perceptions regarding an issue. Various machine learning algorithms have been widely used for this purpose, including Support Vector Machine (SVM) [5] and Random Forest (RF), which are known to perform well in text classification [6]. The primary objective of this study is to develop a web-based information system for sentiment analysis of environmental issues on social media platform X using SVM and RF algorithms, as well as to compare their performance in terms of classification accuracy and effectiveness. This system is expected to enable faster, more accurate, and systematic understanding of public perception while supporting data-driven decision-making on environmental issues.

Several recent studies indicate that the SVM algorithm achieves high accuracy in sentiment analysis based on short texts such as X , while RF excels at handling complex and heterogeneous data and provides stable classification results [7]. Furthermore, comparative studies also show that traditional machine learning methods such as SVM and RF remain competitive compared to more complex methods, especially in terms of computational efficiency and model interpretability [8]. The uniqueness of this research lies in the integration of machine learning-based sentiment analysis with a web-based information system that can be used in real time by end users, rather than being limited to model experimentation.

Other relevant research also shows that using a combination of several classification algorithms can improve sentiment analysis system performance compared to using a single method [9]. This indicates an opportunity to develop more optimal sentiment analysis systems through comparative or hybrid approaches. Based on this background, this study aims to design and build a web-based information system capable of performing sentiment analysis on environmental issues on social media X [10] using the SVM and RF algorithms. This system is expected to help understand public perception more quickly, accurately, and systematically, as well as contribute to data-driven decision-making related to environmental issues. On the other hand, research that integrates sentiment analysis into a web-based information system is still limited. Most studies stop at the model experimentation stage without implementing a system that can be directly used by end users. In terms of system development, a web-based system integrated with a machine learning model can enhance the practical and real-time utilization of analysis results [11].

Although numerous studies on social media sentiment analysis have been conducted, most still focus on general topics such as products, services, politics, or broad public opinion [12]. Research specifically examining environmental issues in Indonesia remains relatively limited, particularly studies that utilize data from platform X as the primary source. Furthermore, some studies only apply a single classification algorithm without conducting an in-depth comparative evaluation of multiple methods. The urgency of this research stems from the increasing environmental challenges in Indonesia, which require rapid and data-driven understanding of public opinion to support effective policy-making..

Another limitation is the lack of a structured data mining development framework, which often results in research processes that are not systematically documented. Approaches such as Cross-Industry Standard Process for Data Mining (CRISP-DM) which encompasses the stages of business understanding [13], data understanding, data preparation, modeling, evaluation, and deployment are still rarely applied comprehensively in sentiment analysis research, particularly in the context of environmental issues in Indonesia. This study aims to address these gaps by conducting sentiment analysis on public opinion regarding environmental issues on social media X, using SVM and RF algorithms [14]. By implementing the full CRISP-DM lifecycle, including web-based deployment, this study provides both methodological and practical contributions, ensuring a systematic and applicable sentiment analysis framework. Thus, the main contributions of this study are: (1) sentiment analysis of environmental issues using social media data in Indonesia, (2) a comprehensive comparative evaluation of SVM and RF algorithms [15], and (3) the development of a web-based system that enables practical and real-time utilization of sentiment analysis results.

2. METHOD

This research employs the CRISP-DM framework to guide the development of a machine learning-based sentiment analysis system. CRISP-DM consists of six well-defined phases and is extensively utilized in data mining projects owing to its systematic approach, adaptability, and compatibility with contemporary

technologies, including web-based applications [16]. The CRISP-DM framework process is shown in Figure 1.

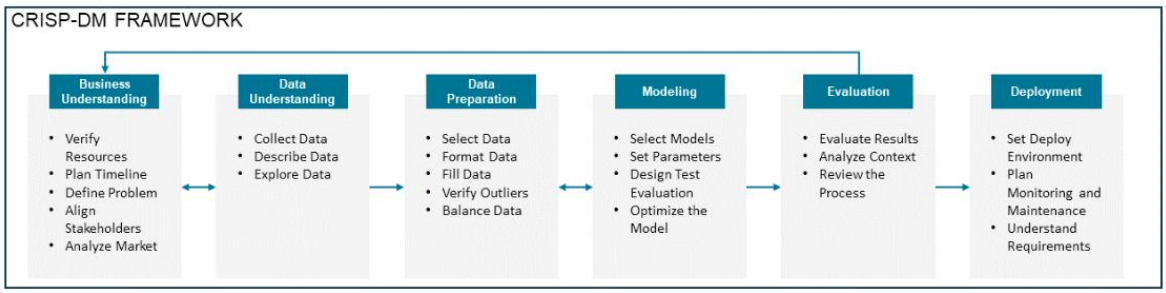

Figure 1. Stages of the CRISP-DM method

Business Understanding

This phase aims to clearly define the research objectives and the requirements of the system to be developed. In the context of this study, the objectives to be achieved include: first, conducting sentiment analysis on environmental issues emerging on the social media platform X; second, building a web-based system capable of monitoring public opinion both in real-time and periodically; and third, providing support for data-driven decision making. In line with the CRISP-DM approach, a deep understanding of the problem domain is a crucial factor to ensure that the resulting sentiment analysis model remains relevant and aligned with the real needs of stakeholders [17].

Data Understanding

This phase encompasses the processes of data collection, exploration, and initial characterization as a fundamental foundation within the CRISP-DM framework. The data used in this study were obtained from the social media platform X through a data scraping process based on keywords related to environmental issues. The collected data consist of public opinion texts that reflect various societal perspectives on environmental problems. Subsequently, a sentiment distribution analysis was conducted to identify the balance among sentiment classes (positive, negative, and neutral) and to detect potential data imbalance that may affect model performance. Modern CRISP-DM emphasizes that data exploration is a critical stage [18], given that most data science project failures are caused by a lack of understanding of data characteristics and quality from the early stages. The formulas for calculating the count and percentage per sentiment class are presented as follows, as shown in Equation (1).

$$n_c = \sum_{i=1}^{N} [y_i = c] \tag{1}$$

Where n_c = the number of tweets in sentiment class C (positive, negative, or neutral), N is the total number of tweets collected, y_i is the sentiment label of the i-th tweet, and $[y_i = c]$ is an indicator function that takes the value 1 if the label of the i-th tweet is equal to class c, and 0 otherwise.

Data Preparation

This stage includes a series of preprocessing and data transformation processes aimed at preparing raw data into a format ready for use in modeling [19]. In this study, four main steps were carried out. First, text cleaning was performed to remove irrelevant elements such as links, punctuation, numbers, emoticons, and words that lack substantive meaning, resulting in clean text [20]. Second, sentiment labeling (label encoding) was applied to three categories negative, neutral, and positive to convert categorical data into numerical representations that can be processed by classification algorithms [21]. Third, feature extraction was conducted using the TF-IDF (Term Frequency-Inverse Document Frequency) method, utilizing a combination of unigrams and bigrams to capture the contextual meaning of each tweet more comprehensively [22]. as shown in Equation (2).

$$\text{"TF-IDF" (t,d)="TF" (t,d)}\times \text{lo g}_{f0}\text{(N/("DF" (t)))} \tag{2}$$

Where "TF"(*t,d*) is the frequency of term t appearing in document *d*, *N* is the total number of documents, and "DF"(*t*) is the number of documents that contain term *t*. Fourth, data imbalance was handled using the SMOTE (Synthetic Minority Oversampling Technique) to improve the representation of the minority class [23]. as shown in Equation (3).

$$x_{new} = xi + \lambda \times (xzi - xi) \tag{3}$$

Where x_i is a minority sample, x_{zi} is its nearest neighbor (k-nearest neighbors), and λ is a random number between 0 and 1. It should be noted that the data preparation stage is the most time-consuming phase within the CRISP-DM framework; nevertheless, this stage largely determines the quality of the model to be produced at the end of the research [19].

Modeling

This stage aims to build sentiment classification models using two algorithms, namely Support Vector Machine and Random Forest, to compare their performance in analyzing public opinion on environmental issues on social media X [24]. The selection of these two algorithms is based on their respective advantages in handling high-dimensional text data.

Support Vector Machine (SVM)

SVM works by finding the optimal hyperplane that separates data between sentiment classes (positive, negative, neutral) in a high-dimensional feature space [25]. The optimal hyperplane is defined as the plane that maximizes the margin, which is the distance between the hyperplane and the nearest support vectors from each class. The SVM decision function can be expressed as shown in Equation (4).

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i \, y_i K(x_i, x) + b\right) \tag{4}$$

Where $f(x)$ = the predicted class for data *x*, α_i = the Lagrange coefficients obtained from the optimization process, y_i = the class label of the *i* th training data, $K(x_i, x)$ = is the kernel function that maps the data into a higher dimensional space, and b is the bias or intercept.

Random Forest (RF)

Random Forest is an ensemble learning method based on decision trees that combines many decision trees to improve accuracy and reduce overfitting [26]. The final prediction of RF for classification tasks is determined through a majority voting mechanism, as shown in Equation (5).

$$\hat{y} = "mode" \left(\{T\_1\,(x), T\_2\,(x), \ldots, T\_k\,(x)\}\right) \tag{5}$$

Where $\hat{y}$ = the final predicted class, $T_j(x)$ = the prediction from the j-th decision tree, k = the number of trees in the forest (n_estimators), Each decision tree in RF is built using bootstrap aggregating (bagging) data samples, and at each split, only a random subset of features is considered [27]. his is stated as shown in Equation (6).

$$m = \sqrt{M} \; or \; m = log_{\,2}(M) + 1 \tag{6}$$

Where M is the total number of features and m is the number of features randomly selected at each split.

Evaluation

The evaluation stage aims to measure the performance of the sentiment classification models built using the SVM and RF algorithms [28]. In the modern CRISP-DM framework [29], evaluation does not only focus on accuracy alone but also considers the stability and generalization capability of the model on unseen data. Predictive evaluation needs to consider the reliability aspects of individual outcomes, especially when the model is used to support decision making [30]. CRISP-DM has evolved into a flexible and adaptive framework that allows the integration of more comprehensive evaluation metrics beyond accuracy [31].

To obtain a comprehensive assessment, several evaluation metrics are employed, namely accuracy, precision, recall, and F1-score. These metrics provide complementary perspectives in evaluating classification performance. Accuracy represents the proportion of correctly classified instances over the total data. Precision

reflects the model's ability to correctly identify positive predictions, while recall measures the model's capability to capture all relevant instances of a particular class. The F1-score is the harmonic mean of precision and recall, balancing both metrics [32]. Mathematically, these evaluation metrics are defined as follows.

The accuracy equation can be seen in Equation (7).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

The precision equation can be seen in Equation (8).

$$Precision = \frac{TP}{FP + TP} \tag{8}$$

The recall equation can be seen in Equation (9).

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

The F1-score equation can be seen in Equation (10).

$$F1 - Score = 2 \times \frac{Precision+Recall}{Precision \times Recall} \tag{10}$$

The use of multiple evaluation metrics is essential in sentiment analysis, particularly when dealing with imbalanced class distributions. Relying solely on accuracy may lead to misleading conclusions, as a model can achieve high accuracy while failing to correctly classify minority classes, such as neutral sentiment.

Deployment

The deployment stage aims to bridge the gap between experimental research and practical utility by transforming the developed model into a functional system that can be accessed by end-users [33]. In the modern CRISP-DM framework, deployment is an important contribution that transforms analytical models into solutions that are accessible and usable by stakeholders. In this study, the best-performing model resulting from the evaluation stage is integrated into a web-based system to monitor public opinion on environmental issues on social media X in real-time. This system is designed to provide data-driven decision support for stakeholders, such as environmental policy managers, researchers, and environmental activists [34].

Technically, the system is developed using a client-server architecture, where the frontend is responsible for user interaction and the backend handles data processing and sentiment classification. The frontend interface is designed to be responsive and user-friendly, allowing users to input text manually or upload datasets for analysis. The backend integrates the trained machine learning model (SVM or RF) and processes incoming data through a pipeline consisting of text preprocessing, feature extraction using TF-IDF [35], and classification. The system provides several core functionalities, including: (1) Text Input and Dataset Upload, Users can analyze single text inputs or bulk tweet data. (2) Real-Time Sentiment Classification, The system processes input data and returns sentiment labels (positive, negative, neutral) along with confidence scores. (3) Data Visualization, The system generates visual outputs such as WordCloud for overall, positive, and negative sentiments, as shown in Figure 9. (4) Downloadable Results, Users can download the generated visualizations and analysis results for further use. (5) Session Management, The system includes a session reset feature to allow users to perform new analyses efficiently.

From a processing perspective, when a user submits input, the data is sent to the server via an HTTP request. The backend then applies preprocessing (cleaning, normalization), transforms the text into numerical vectors using TF-IDF [36], and feeds the data into the trained classification model. The prediction results are then returned to the frontend and displayed in both textual and visual formats. This deployment ensures that the developed model is not only theoretically validated but also practically applicable, enabling stakeholders to monitor public sentiment dynamically and support data-driven decision-making.

3. RESULTS AND DISCUSSIONS

Data Collection Process from Social Media X

The data collection process was conducted as the initial stage of this research, targeting public discussions on environmental issues across Indonesia. The collection period spanned from 2021 to 2025, capturing public discourse over a five-year period to enable longitudinal analysis of opinion trends. Several studies have confirmed that social media platforms, particularly X (formerly Twitter), serve as vibrant and informative repositories of public opinion on environmental challenges.

To collect tweets containing specific environmental keywords, a Python-based crawling script was developed using two complementary approaches. First, the Snscrape library was employed as the primary tool because it does not require official API authentication and remains functional for public tweet retrieval without rate limiting. Second, the Selenium library was used as a fallback method to handle dynamic content and bypass platform changes that may affect Snscrape's functionality. Selenium enables browser automation to mimic human user activities such as page navigation, keyword searching, automatic scrolling, and dynamic tweet extraction. The combination of these approaches ensures robust data collection without relying on X's official API, which imposes significant rate limitations for academic research.

A total of 24 keywords related to environmental issues were employed to direct the crawling process toward topics actively discussed by X users. As presented in Figure 2, these keywords encompassed various environmental themes, including waste management ("daur ulang sampah," "bank sampah," "sampah plastik"), water pollution ("sungai bersih," "pencemaran air," "sungai tercemar"), air quality ("polusi udara"), deforestation ("reboisasi," "kebakaran hutan"), and natural disasters ("banjir bandang," "solusi banjir").

```
# ==========================================
# KEYWORD DEFINITION
# ==========================================
# Daftar 24 kata kunci isu lingkungan (Table 1)
environmental_keywords = [
    "gotong royong", "daur ulang sampah", "sungai bersih",
    "pengelolaan sampah", "gerakan hijau lingkungan", "lingkungan bersih",
    "pengolahan limbah", "pengelolaan sampah plastik", "pencemaran air",
    "bank sampah", "solusi banjir", "csr lingkungan", "reboisasi",
    "penanaman pohon", "pembersihan sungai", "polusi udara",
    "sampah plastik", "banjir bandang", "kebakaran hutan",
    "limbah industri", "tumpukan sampah", "buang sampah sembarangan",
    "sungai tercemar", "tempat pembuangan sampah"
]
```

Figure 2. List of 24 environmental keywords used for data crawling

The dataset obtained from the crawling process was stored in CSV (Comma-Separated Values) format to facilitate seamless integration with subsequent processes such as text preprocessing and sentiment labeling. The CSV format was chosen due to its simplicity, wide compatibility, and ease of use with the pandas library in Python for data manipulation and analysis. Furthermore, this format is fully compatible with the web based system developed for public opinion monitoring. As shown in Figure 3, the dataset comprises 47,249 rows (tweets) and 4 columns, including username, tweet (tweet content), waktu (timestamp), and tweet_link (URL of the tweet). This structured representation enables efficient data handling and ensures reproducibility of the analysis.

	username	tweet	waktu	tweet_link	polarity
0	saveforest	saveforest · Masalah perubahan iklim semakin p...	12-02-2022 06:51	https://x.com/saveforest/status/75740906668620...	negative
1	climate_id	climate_id · Program energi terbarukan semakin...	21-08-2024 12:55	https://x.com/climate_id/status/34586435367889...	positive
2	eco_warrior	eco_warrior · Masalah sampah plastik semakin p...	14-01-2022 22:37	https://x.com/eco_warrior/status/7390022817412...	negative
3	eco_news	eco_news · Program energi terbarukan semakin b...	11-03-2021 18:50	https://x.com/eco_news/status/287544677673599777	positive
4	eco_action	eco_action · Program banjir perkotaan semakin ...	29-12-2022 17:43	https://x.com/eco_action/status/84319252092357...	positive
5	lingkungan_id	lingkungan_id · Program reboisasi hutan semaki...	26-02-2021 10:21	https://x.com/lingkungan_id/status/41056964816...	positive
6	forest_watch	forest_watch · Program pencemaran air semakin ...	08-05-2022 03:37	https://x.com/forest_watch/status/949264354112...	positive
7	eco_warrior	eco_warrior · Program sampah plastik semakin b...	18-06-2025 21:18	https://x.com/eco_warrior/status/2234617390945...	positive
8	climate_id	climate_id · Program pencemaran air semakin ba...	19-08-2024 10:55	https://x.com/climate_id/status/65467086741155...	positive
9	forest_watch	forest_watch · Isu banjir perkotaan masih menj...	03-09-2022 19:27	https://x.com/forest_watch/status/336763411287...	neutral
47240	saveforest	saveforest · Isu deforestasi masih menjadi per...	28-01-2024 12:22	https://x.com/saveforest/status/74814102925202...	neutral
47241	eco_news	eco_news · Program pencemaran air semakin baik...	03-06-2021 23:51	https://x.com/eco_news/status/703412974944112087	positive
47242	forest_watch	forest_watch · Masalah deforestasi semakin par...	10-01-2023 05:34	https://x.com/forest_watch/status/683052480193...	negative
47243	saveearth	saveearth · Program polusi udara semakin baik ...	16-05-2024 14:28	https://x.com/saveearth/status/464191738031021204	positive
47244	eco_action	eco_action · Masalah sampah plastik semakin pa...	15-02-2024 19:52	https://x.com/eco_action/status/33191344948148...	negative
47245	bumi_id	bumi_id · Isu pencemaran air masih menjadi per...	15-05-2022 06:29	https://x.com/bumi_id/status/819003552316761401	neutral
47246	hijaukan_id	hijaukan_id · Masalah sampah plastik semakin p...	25-07-2023 08:07	https://x.com/hijaukan_id/status/5671213000819...	negative
47247	bumi_lestari	bumi_lestari · Masalah pencemaran air semakin ...	21-03-2025 06:29	https://x.com/bumi_lestari/status/115600972954...	negative
47248	eco_update	eco_update · Program perubahan iklim semakin b...	03-06-2025 11:15	https://x.com/eco_update/status/81738316206518...	positive
47249	eco_news	eco_news · Isu energi terbarukan masih menjadi...	18-12-2024 05:03	https://x.com/eco_news/status/619176088952056336	neutral

Figure 3. Sample of the crawled dataset in CSV format (47,249 tweets)

Sentiment Distribution Analysis

Prior to sentiment labeling, the raw dataset of 47,245 tweets underwent a rigorous data normalization process to ensure data quality. Tweets that were duplicates, non-Indonesian, irrelevant to environmental issues, or lacking substantive content were systematically removed. This filtering stage yielded a normalized dataset of 13,063 tweets, representing a 72.4% reduction from the raw data a proportion consistent with standard text mining practices for social media data. The distribution of sentiments across this normalized dataset is presented in Table 1.

Table 1. Distribution of sentiment classes

Sentiment Class	Number of Tweets	Percentage (%)
Negative	9,510	72.8%
Neutral	673	5.2%
Positive	2,880	22.0%
Total	**13,063**	**100%**

Model Training and Evaluation

Model training was conducted using two widely adopted machine learning algorithms for text classification tasks: SVM and RF. Both algorithms were selected based on their proven effectiveness in sentiment analysis and their ability to handle high-dimensional sparse data, such as TF-IDF transformed text features. The dataset used for model training consisted of 13,063 tweets that had undergone the complete preprocessing pipeline described in section sentiment distribution analysis, including text cleaning, tokenization, stopword removal, stemming, and TF-IDF feature extraction with unigram and bigram combinations.

As shown in Figures 4 and 5, the SVM model achieved a higher accuracy of 0.83, with excellent performance on the negative class as well, specifically a precision of 0.94, a recall of 0.88, and an F1-score of 0.91. Performance on the positive class was also relatively high compared to RF. However, similar to RF, performance on the neutral class remains low, with an F1-score of only 0.37. Although the overall accuracy of the model is relatively moderate, this result should be interpreted with caution, particularly in the context of imbalanced data distribution. As shown in the evaluation results , the model demonstrates strong performance in the dominant classes, especially the negative class, achieving high precision, recall, and F1-score values.

Akurasi: 0.83
Laporan Klasifikasi:

	precision	recall	f1-score	support
negative	0.94	0.88	0.91	9510
neutral	0.29	0.49	0.37	673
positive	0.73	0.76	0.75	2880
accuracy			0.83	13063
macro avg	0.65	0.71	0.67	13063
weighted avg	0.86	0.83	0.84	13063

Figure 4. Model evaluation support vector machine

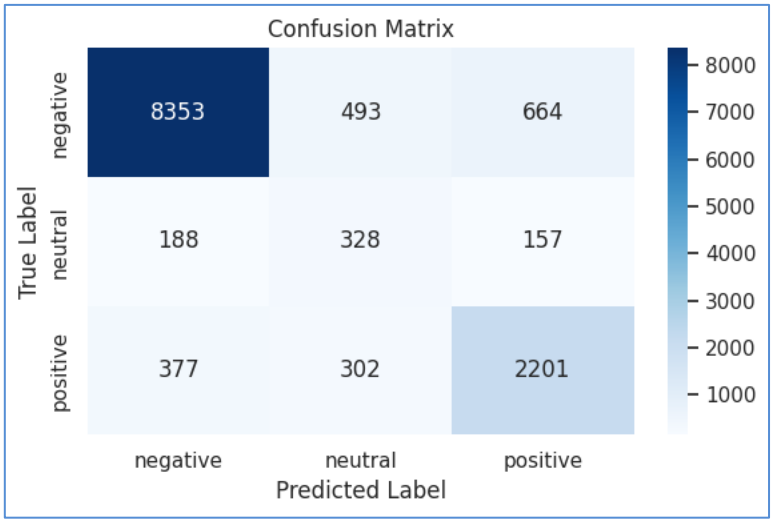

Figure 5. Confusion matrix support vector machine

As shown in Figures 6 and 7, the Random Forest (RF) model achieved an accuracy of 0.81, with its best performance observed in the negative class, where precision reached 0.88, recall 0.90, and F1-score 0.89. However, the model shows relatively low performance in the neutral class (precision of 0.41 and recall of 0.43), indicating difficulty in distinguishing neutral sentiment from other classes. Despite this limitation, the model still provides meaningful insights, particularly in identifying strong sentiment polarity (positive and negative), which is often more critical in practical applications such as public opinion monitoring. Therefore, the model remains valuable for supporting decision-making processes, even with a moderate overall accuracy.

Akurasi Random Forest setelah tuning: 0.82
Laporan Klasifikasi Random Forest setelah tuning:

	precision	recall	f1-score	support
negative	0.88	0.90	0.89	6340
neutral	0.42	0.45	0.43	449
positive	0.69	0.62	0.66	1920
accuracy			0.82	8709
macro avg	0.66	0.66	0.66	8709
weighted avg	0.81	0.82	0.82	8709

Figure 6. Model evaluation random forest

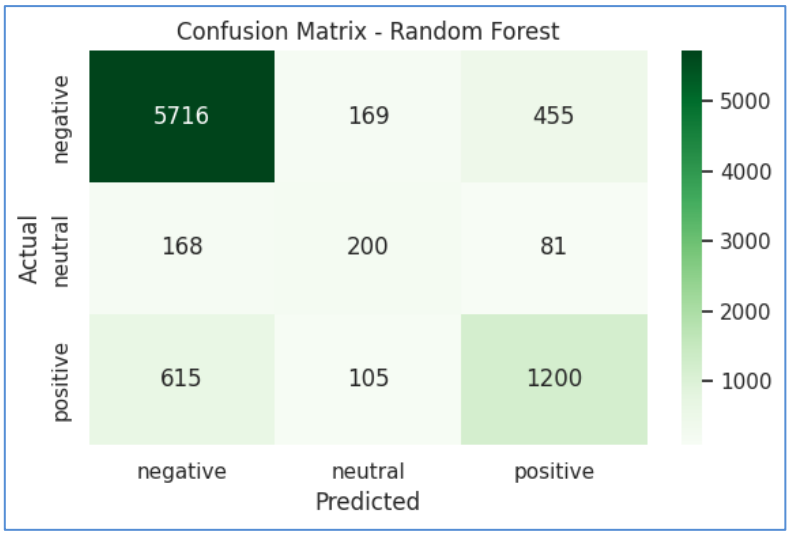

Figure 7. Confusion matrix random forest

Web Based System Implementation

The optimal model derived from the CRISP-DM framework was successfully deployed as a web-based sentiment analysis system. This system enables users to perform real-time sentiment classification on tweet data related to environmental issues. The web interface was intentionally designed with a minimalist and responsive layout to enhance usability and accessibility across various devices (desktop, tablet, and mobile). As shown in Figure 8, the interface provides an intuitive user experience, allowing users to input text or upload datasets and receive instant sentiment predictions (positive, negative, or neutral) along with confidence scores.

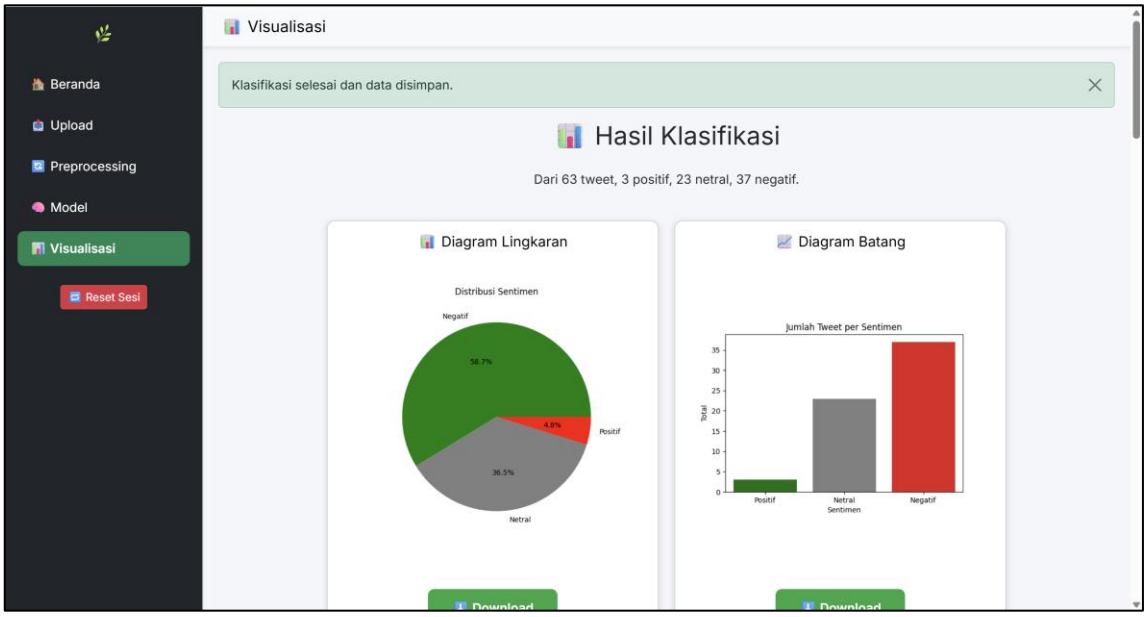

Figure 8. Results visualization report page

Additionally, as shown in Figure 9. Results Visualization WordCloud, the system provides interactive WordCloud visualizations for all tweets, positive sentiments, and negative sentiments. These visualizations help users quickly identify dominant keywords and themes within each sentiment category. Each WordCloud can be downloaded individually using the provided download buttons, and the "Reset Sesi" button allows users to restart the analysis session conveniently.

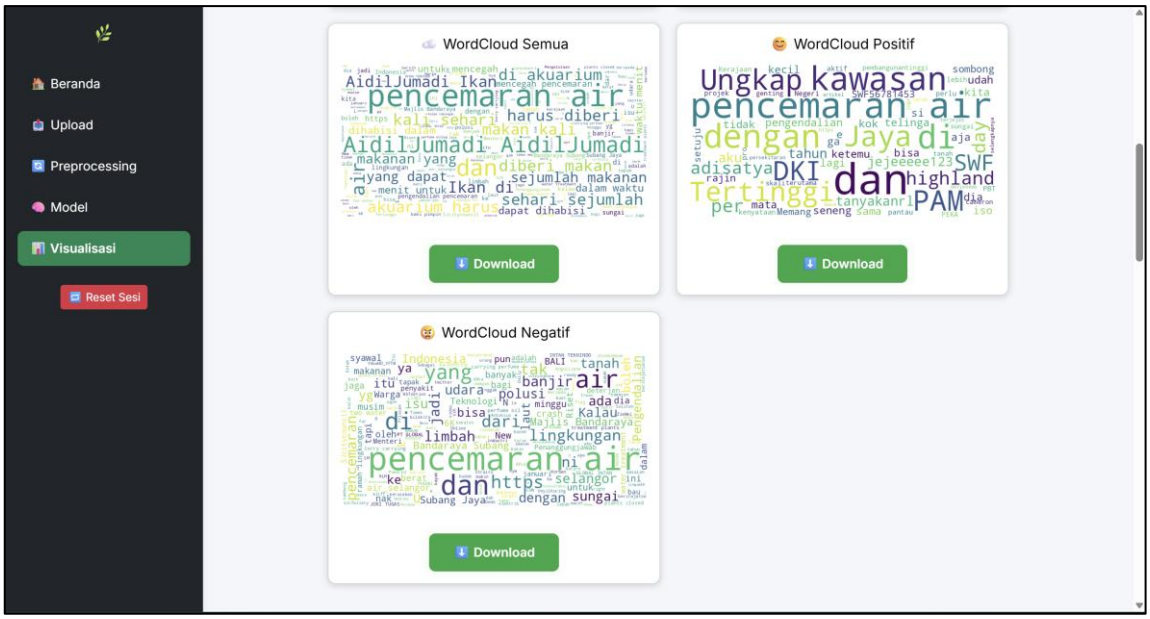

Figure 9. Results Visualization WordCloud

By transforming complex machine learning outputs into accessible visual insights, the system enhances usability and bridges the gap between analytical models and practical applications. This makes the proposed system particularly useful for supporting data-driven policy evaluation, environmental awareness campaigns, and public communication strategies.

Comparison with Previous State-of-the-Art Studies

To position this research within the current body of knowledge, a comparative analysis was conducted against three relevant studies that employed machine learning algorithms for sentiment analysis on social media data. The comparison focuses on key aspects including dataset characteristics, preprocessing techniques, classification algorithms, evaluation methods, and implementation into deployable systems. Table 2 presents the comparative summary.

Table 2. Comparative analysis of sentiment analysis studies on social media

Aspect	Widyanto et al. (2023) [1]	Aryanti & Suria (2025) [10]	Fahri & Gunawan (2025) [7]	**This Study**
Platform	Twitter (X)	Twitter (X)	Twitter (X)	**Twitter (X)**
Domain	Health Law (RUU Kesehatan)	Employment Termination (PHK)	Women in Workplace	**Environmental Issues**
Dataset Size	Not specified	Not specified	Not specified	**13,063 labeled tweets**
Time Period	Not specified	Not specified	Not specified	**2021–2025 (5 years)**
Algorithms	Naïve Bayes, SVM	IndoBERT, SVM, RF, Decision Tree	SVM	**SVM, RF**
Framework	Not explicitly used	Not explicitly used	Not explicitly used	**Full CRISP-DM (6 stages)**
Web-Based Deployment	Not implemented	Not implemented	Not implemented	**Implemented (Streamlit/Flask)**
Practical Contribution	Limited	Limited	Limited	**Decision support for environmental monitoring**
Real-Time Analysis	No	No	No	**Yes**

4. CONCLUSION

This study demonstrates the effectiveness of supervised machine learning approaches for analyzing public sentiment on environmental issues from social media platform X. Among the evaluated models, Support Vector Machine (SVM) achieved the best performance with an accuracy of 83%, slightly outperforming

Random Forest (81%), indicating its suitability for handling high-dimensional textual data. Despite limitations in classifying neutral sentiment, both models showed strong capability in identifying dominant sentiment polarity (positive and negative), which is crucial for understanding public opinion trends. The study also contributes by implementing a web-based sentiment analysis system that transforms complex analytical results into accessible visual insights, enabling stakeholders to monitor environmental issues in real time. This practical implementation highlights the applicability of machine learning in supporting data-driven decision-making. From a broader perspective, the findings of this study can assist policymakers, researchers, and environmental organizations in identifying public concerns, evaluating environmental policies, and designing more responsive interventions. Future research is recommended to improve model performance, particularly in handling neutral sentiment, by incorporating advanced techniques such as deep learning models or contextual language representations.

REFERENCES

[1] W. Tetrian, R. Ina, and W. Arief, "A Comparison of Naïve Bayes and SVM for Sentiment Analysis of the Health Bill on Twitter," *SINTECH (Science and Information Technology) Journal*, vol. 6, no. 3, pp. 147–161, Dec. 2023, doi: 10.31598/sintechjournal.v6i3.1433.

[2] A. I. Hamid and A. M. Abdulazeez, "Sentiment Analysis Based on Machine Learning Techniques: A Comprehensive Review," *Indonesian Journal of Computer Science*, vol. 13, no. 3, Jun. 2024, doi: 10.33022/ijcs.v13i3.4049.

[3] A. Hermawan, I. Jowensen, J. Junaedi, and Edy, "Implementation of Text Mining for Sentiment Analysis on Twitter Using the Support Vector Machine Algorithm," *JST (Jurnal Sains dan Teknologi)*, vol. 12, no. 1, pp. 129–137, Apr. 2023, doi: 10.23887/jstundiksha.v12i1.52358.

[4] J. Benchimol, S. Kazinnik, and Y. Saadon, "Text mining methodologies with R: An application to central bank texts," *Machine Learning with Applications*, vol. 8, p. 100286, Jun. 2022, doi: 10.1016/j.mlwa.2022.100286.

[5] J. Ipmawati, S. Saifulloh, and K. Kusnawi, "Sentiment Analysis of Tourist Attractions Based on Google Maps Reviews Using the Support Vector Machine Algorithm," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 1, pp. 247–256, Jan. 2024, doi: 10.57152/malcom.v4i1.1066.

[6] H. T. H. Jaya, R. Yova, A. A. Rizki, M. G. Raditia, N. A. Wija, and A. M. Wijaya, "Sentiment analysis of twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier," *Procedia Comput. Sci.*, vol. 197, pp. 660–667, 2022, doi: 10.1016/j.procs.2021.12.187.

[7] F. M. D. Hilal and D. Gunawan, "Analysis of X Users' Sentiment Toward Women in the Workplace Using Machine Learning Algorithms," *Journal of Technology and Informatics (JoTI)*, vol. 7, no. 2, Sep. 2025, doi: 10.37802/joti.v7i2.1087.

[8] D. A. Dzulhijjah, S. Hafidz, H. A.S. Wahyudi, Y. A. Almi, and E. Utami, "A Comparison of the Random Forest and KNN Methods in Twitter Sentiment Analysis," *Smart Comp: Jurnalnya Orang Pintar Komputer*, vol. 12, no. 3, pp. 767–772, Jul. 2023, doi: 10.30591/smartcomp.v12i3.5106.

[9] A. A. Nida Nur and S. Ozzi, "Sentiment Analysis of Layoffs in Indonesia: A Comparison of Indobert with SVM, Random Forest, and Decision Trees Using TF-IDF Optimization," *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 10, no. 2, pp. 1158–1176, Jul. 2025, doi: 10.36341/rabit.v10i2.6364.

[10] F. Novianti and K. R. N. Wardani, "Analysis of Public Sentiment Toward Traveloka Tweet Data During Antigen Rapid Testing Using the Naïve Bayes Algorithm," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 8, no. 3, pp. 922–933, Aug. 2023, doi: 10.29100/jipi.v8i3.3973.

[11] S. E. Septa, A. S. Ruli, V. C. Bella, and P. A. Nugroho, "Implementation of Machine Learning in an LLM-Integrated Diet Plan Prediction and Recommendation System," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 11, no. 2, pp. 144–151, Sep. 2025, doi: 10.25077/TEKNOSI.v11i2.2025.144-151.

[12] A. S. Jaiswal, D. V. Bhavsagar, K. Dhole, S. Chaurasia, M. Daph, and S. Chourasia, "Sentiment Analysis of Election Result Prediction using Twitter Data By NLP and ML," in *2024 International Conference on Innovations and Challenges in Emerging Technologies (ICICET)*, IEEE, Jun. 2024, pp. 1–4. doi: 10.1109/ICICET59348.2024.10616319.

[13] S. N. D.Husna and S. Sari, "Analysis of Public Sentiment Toward Plastic Waste Reduction Campaigns Using the Naïve Bayes Algorithm," *JURNAL FASILKOM*, vol. 15, no. 2, pp. 202–212, Aug. 2025, doi: 10.37859/jf.v15i2.9574.

[14] A. Nursikuwagus, Suherman, H. Purwanto, and T. Hartono, "Support Vector Machine to Classify Sentiment Reviews on Google Play Store," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 11, no. 3, pp. 724–732, Feb. 2026, doi: 10.33480/jitk.v11i3.7282.

[15] E. Wincent, J. Jovito, R. D. Bee, and E. D. Madyatmadja, "Harmonizing Public Sentiment Analysis on Indonesia's New Capital (IKN) on X Using Machine Learning," in *2025 7th International Conference on Cybernetics and Intelligent System (ICORIS)*, IEEE, Sep. 2025, pp. 1–5. doi: 10.1109/ICORIS67789.2025.11295975.

[16] R. A. Casonatto, T. De Pádua Grillo Souza, and A. M. Mariano, "Quality and Risk Management in Data Mining: A CRISP-DM Perspective.," *Procedia Comput. Sci.*, vol. 242, pp. 161–168, 2024, doi: 10.1016/j.procs.2024.08.257.

[17] N. Lundén, E. T. Bekar, A. Skoogh, and J. Bokrantz, "Domain Knowledge in CRISP-DM: An Application Case in Manufacturing," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 7603–7608, 2023, doi: 10.1016/j.ifacol.2023.10.1156.

[18] S. Maataoui, G. Bencheikh, and G. Bencheikh, "Predictive Maintenance in the Industrial Sector: A CRISP-DM Approach for Developing Accurate Machine Failure Prediction Models," in *2023 Fifth International Conference on Advances in Computational Tools for Engineering Applications (ACTEA)*, IEEE, Jul. 2023, pp. 223–227. doi: 10.1109/ACTEA58025.2023.10193983.

[19] A. M. Shimaoka, R. C. Ferreira, and A. Goldman, "The evolution of CRISP-DM for Data Science: Methods, Processes and Frameworks," *SBC Reviews on Computer Science*, vol. 4, no. 1, pp. 28–43, Oct. 2024, doi: 10.5753/reviews.2024.3757.

[20] Z. Li and Z. Zou, "Punctuation and lexicon aid representation: A hybrid model for short text sentiment analysis on social media platform," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 3, p. 102010, Mar. 2024, doi: 10.1016/j.jksuci.2024.102010.

[21] Y. Durachman, S. J. Putra, H. Nanang, and H. T. Sukmana, "Analysis Sentiment of Public Opinion on Social Media Using Naïve Bayes and TF-IDF Algorithms," in *2024 3rd International Conference on Creative Communication and Innovative Technology (ICCIT)*, IEEE, Aug. 2024, pp. 1–6. doi: 10.1109/ICCIT62134.2024.10701191.

[22] A. D. Adyatma, L. Afuan, and E. Maryanto, "The Effect Of Unigram And Bigram In The Naïve Bayes Multinomial For Analyzing Of Comment Sentiment Of Gojek Application In Google Play Store," *Jurnal Teknik Informatika (Jutif)*, vol. 4, no. 6, pp. 1535–1540, Dec. 2023, doi: 10.52436/1.jutif.2023.4.6.1310.

[23] S. F. Taskiran, B. Turkoglu, E. Kaya, and T. Asuroglu, "A comprehensive evaluation of oversampling techniques for enhancing text classification performance," *Sci. Rep.*, vol. 15, no. 1, p. 21631, Jul. 2025, doi: 10.1038/s41598-025-05791-7.

[24] R. A. Kumar, R. Karnati, K. S. Goud, N. Ravula, and V. Murthy, "A Novel Technique for Analyzing the Sentiment of Social Media Posts Using Deep Learning Techniques," 2024, pp. 263–273. doi: 10.1007/978-3-031-48888-7_22.

[25] I. M. Fadhil and Y. Sibaroni, "Topic Classification in Indonesian-language Tweets using Fast-Text Feature Expansion with Support Vector Machine (SVM)," in *2022 International Conference on Data Science and Its Applications (ICoDSA)*, IEEE, Jul. 2022, pp. 214–219. doi: 10.1109/ICoDSA55874.2022.9862899.

[26] G. M, V. Asha, K. Vinisha, J. H V, and J. R, "Predicting Crop Yields with Random Forest: A Data-Driven Approach," in *2025 3rd International Conference on Inventive Computing and Informatics (ICICI)*, IEEE, Jun. 2025, pp. 223–228. doi: 10.1109/ICICI65870.2025.11069799.

[27] Y. Liao, Q. Wu, and X. Yan, "Invariant Random Forest: Tree-Based Model Solution for OOD Generalization," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 12, pp. 13772–13781, Mar. 2024, doi: 10.1609/aaai.v38i12.29283.

[28] P. M. H. Tridharma and A. D. Indriyanti, "Sentiment Analysis of Public Figures on X Using Naïve Bayes and SVM," *Journal of Emerging Information Systems and Business Intelligence*, vol. 7, no. 2, pp. 307–312, 2026.

[29] S. M. T. Hayat and W. Yani, "Machine Learning Application Development Guidelines Using CRISP-DM and Scrum Concept," in *2023 IEEE International Conference on Data and Software Engineering (ICoDSE)*, IEEE, Sep. 2023, pp. 168–173. doi: 10.1109/ICoDSE59534.2023.10291438.

[30] S. Staudinger, C. G. Schuetz, and M. Schrefl, "A Reference Process for Assessing the Reliability of Predictive Analytics Results," *SN Comput. Sci.*, vol. 5, no. 5, p. 563, May 2024, doi: 10.1007/s42979-024-02892-4.

[31] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Comput. Sci.*, vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[32] A. Gupta, D. Ather, R. Kler, N. Chaudhary, G. Singh, and Chitra, "Consumer Feedback Analysis on iPhones: EDA and NLP Model Comparison for Sentiment Detection," in *2024 International Conference on Intelligent & Innovative Practices in Engineering & Management (IIPEM)*, IEEE, Nov. 2024, pp. 1–5. doi: 10.1109/IIPEM62726.2024.10925785.

[33] A. Yazid, A. M.Helmy, and R. Jiki, "Web-based CNN Application for Arabica Coffee Leaf Disease Prediction in Smart Agriculture," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 1, pp. 71–79, Feb. 2023, doi: 10.29207/resti.v7i1.4622.

[34] X. Hu, "Research on Integration of Intelligent Algorithms in Big Data-Driven Decision Support Systems," in *2024 IEEE 7th International Conference on Information Systems and Computer Aided Education (ICISCAE)*, IEEE, Sep. 2024, pp. 1195–1199. doi: 10.1109/ICISCAE62304.2024.10761383.

[35] E. Yang and Z. Long, "Research on the Weighting Method Based on Tf-IDF and Apriori Algorithm," in *2023 IEEE 6th International Conference on Information Systems and Computer Aided Education (ICISCAE)*, IEEE, Sep. 2023, pp. 1003–1005. doi: 10.1109/ICISCAE59047.2023.10393523.

[36] T. Liu, "Research on Chinese News Text Classification Based on Improved TF-IDF and MNB," in *2025 International Conference on Power, Electrical Engineering, Electronics and Control (PEEEC)*, IEEE, Dec. 2025, pp. 1003–1007. doi: 10.1109/PEEEC67807.2025.00174.