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

Vol. 7, No. 2, April 2026, Page. 1007-1027
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5415

Analysis of Public Sentiment Indonesia's Personal Data Protection Law: A Comparison of SVM and IndoBERT on X Platform

Yulia Kurniawati[*1], Ricky Bahari Hamid[2], Dana Indra Sensuse[3], Sofian Lusa[4], Prasetyo Adi Wibowo Putro[5], Sofiyanti Indriasari[6]

[1,2,3]Faculty of Computer Science, Universitas Indonesia, Indonesia
[4]Department of Tourism, Trisakti Institute of Tourism, Indonesia
[5]Cryptographic Engineering, Polytechnic of Cyber and State Cryptography, Indonesia
[6]Software Engineering Technology, School of Vocational Studies IPB University, Indonesia

Email: [1]yulia.kurniawati@ui.ac.id

Received : Oct 31, 2025; Revised : Dec 12, 2025; Accepted : Dec 15, 2025; Published : Apr 15, 2026

Abstract

The high number of data misuses, thefts, and leaks led to the enactment of the PDP Law, which regulates the rights and obligations of data owners and electronic system providers. The purpose of this study is to examine the public's response to the implementation of the law through the X platform, using tweet harvest as a scraping tool, and to evaluate model performance through a comparative approach between SVM and BERT. The feature extraction used in this study is TF-IDF for SVM and BERT with IndoBERT. The accuracy results indicate that BERT is better with an accuracy of 86% compared to SVM with a training and test data ratio of 85:15. This advantage is because BERT can understand linguistic context that SVM cannot. On the other hand, SVM has advantages in computational efficiency and faster processing, making it a suitable choice in situations with limited computational resources.
The sentiment analysis result revealed that data protection, digital footprint and the institution's role were the most frequently discussed topics. Furthermore, periodic or real-time evaluations can be conducted on the public's response to the PDP Law to ensure it remains aligned and relevant to technological developments and societal needs.

Keywords : *IndoBERT, Privacy Data Protection, Public Sentiment, Sentiment Analysis, Support Vector Machine (SVM), UU PDP*

1. INTRODUCTION

The large number of people in Indonesia who rely on gadgets for communication, work, learning, and entertainment has resulted in a vast amount of information being stored online through various application uses[1]. The exposure of data from two million customers of state-owned banks in 2021, the misuse of data on one of Indonesia's largest e-commerce platforms in 2020, and privacy-related incidents in Indonesia illustrate the importance the enactment of data protection laws as a form of government presence in protecting the data privacy of the public[2]. In 2022, the Indonesian House of Representatives (DPRD) passed the Personal Data Protection Law (Law No. 27 on Personal Data Protection), which will come into effect on October 17, 2024. This law sets comprehensive standards for personal data management, requiring all organizations acting as data controllers to adhere to the principles of transparency, accountability, and integrity[3]. When the PDP Law was announced, the public's response was varied, particularly on the X platform (formerly Twitter), including doubts about the readiness of stakeholders to comply with the law and concerns about the government's readiness to implement it, especially since no personal data protection agency had been established[4].

In Indonesia, various social media platforms are used to exchange opinions, including YouTube, specifically in the comment section, to respond to the availability of the COVID-19 vaccine[5]. Other media, such as Twitter, now known as Platform X, is one of the social media platforms that allows the

Vol. 7, No. 2, April 2026, Page. 1007-1027
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5415

public to express their aspirations regarding public policies that promote healthy democracy [6]. Sentiment analysis can be used as one method to monitor and evaluate government policies in the future[7]. Sentiment analysis plays a role in gauging public sentiment and assessing public acceptance of implemented policies[8], For example, by looking at negative sentiment, the government can maximize or focus on improving its policies to align with the issues faced by society [9] . Sentiment analysis can also be used to track changes in public opinion over time because public sentiment is dynamic and can shift based on social context or specific times[10]. Sentiment analysis can also be used to group user comments based on topic similarity, representing user persona groups[11]. One categorization that can be used is Aspect-based sentiment analysis, which can help the government understand the aspirations expressed by the public specifically based on topic[12]. Sentiment analysis can also be used automatically in the form of websites that can assess sentiment based on user input [13]. Sentiment analysis can also be done across platforms, for example, by looking at sentiment on YouTube and Facebook to gain a broader view of public response [14].

The selection of an appropriate classification model and word embedding needs to be considered to provide balanced performance in sentiment analysis [15], Therefore, some studies compare various models that are deemed effective to determine which one achieves the highest accuracy in sentiment analysis [16]. The division of data, for example, into how much training data and test data, also affects the performance of sentiment classification [17]. Additionally, back translation and adjusting stratified K-fold cross also play a role in influencing model accuracy [18]. Additionally, combining the two models is also possible to obtain optimal results [19].

Sentiment analysis related to the PDP Law was previously conducted using the Multinomial Naïve Bayes Algorithm, comparing the accuracy results of manual labeling (72%) and automatic labeling (74%). From that study, it was found that the accuracy of automatic labeling results was 2% higher than manual labeling [13]. Based on the literature study conducted, the researcher compared it with previous studies that discussed sentiment analysis on the topic of policy in Indonesia in Based on the comparison conducted in **Kesalahan! Bukan swa-referensi bookmark yang valid.**, it was found through gap analysis that SVM has the highest accuracy and is used most frequently compared to other algorithms. Additionally, it was found that combining SVM with TF-IDF as a feature extraction method proved superior to Bag of Words [22]. While the use of IndoBERT is still rare, its accuracy is higher compared to LR, Support Vector Classifier, Random Forest, LGBM Classifier, XGBoost, AdaBoost, and Decision Tree with the same data [34]. Additionally, mBERT can be used in various languages [35], while IndoBERT can understand the linguistic context more accurately because it is a pre-trained model specifically for the Indonesian language [36], Additionally, in further research, there is IndoGovBERT, which is specifically trained for classification based on Indonesian government documents, particularly in SDG's[37] . Therefore, in this study, the SVM classification algorithm with TF-IDF and IndoBERT with BERT Embedding was used to analyze sentiment toward the PDP Law. The comparative approach between SVM and IndoBERT on the Indonesian language dataset provides new empirical evidence regarding the effectiveness of the classification models. Additionally, pre- and post-implementation analyses were conducted to address the temporal gap from the enactment of the PDP Law.

This research was conducted to examine public opinion on the PDP Law and answer the following research questions :

RQ1 : What is the public's response and opinion on the implementation of the PDP Law in Indonesia ?

RQ2 : How do BERT and SVM perform comparatively in sentiment analysis of the PDP Law in Indonesia?

Vol. 7, No. 2, April 2026, Page. 1007-1027
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5415

Table 1.

Based on the comparison conducted in **Kesalahan! Bukan swa-referensi bookmark yang valid.**, it was found through gap analysis that SVM has the highest accuracy and is used most frequently compared to other algorithms. Additionally, it was found that combining SVM with TF-IDF as a feature extraction method proved superior to Bag of Words [22]. While the use of IndoBERT is still rare, its accuracy is higher compared to LR, Support Vector Classifier, Random Forest, LGBM Classifier, XGBoost, AdaBoost, and Decision Tree with the same data [34]. Additionally, mBERT can be used in various languages [35], while IndoBERT can understand the linguistic context more accurately because it is a pre-trained model specifically for the Indonesian language [36], Additionally, in further research, there is IndoGovBERT, which is specifically trained for classification based on Indonesian government documents, particularly in SDG's[37] . Therefore, in this study, the SVM classification algorithm with TF-IDF and IndoBERT with BERT Embedding was used to analyze sentiment toward the PDP Law. The comparative approach between SVM and IndoBERT on the Indonesian language dataset provides new empirical evidence regarding the effectiveness of the classification models. Additionally, pre- and post-implementation analyses were conducted to address the temporal gap from the enactment of the PDP Law.

This research was conducted to examine public opinion on the PDP Law and answer the following research questions :

RQ1 : What is the public's response and opinion on the implementation of the PDP Law in Indonesia ?

RQ2 : How do BERT and SVM perform comparatively in sentiment analysis of the PDP Law in Indonesia?

Table 1. Comparison of Classification Method Usage in Policy Sentiment Analysis in Indonesia

Classification Methods	Highest Accuracy	Reference Journal	Description
SVM	98.75%	[20][21][8][18][12] [22][23][24][25][26] [27]	It has the highest accuracy and is the most frequently used [20]
LSTM	97.49%	[11][15][28]	On the same data, LSTM is better than Naïve Bayes [11]
CNN	96%	[10][29]	In the same dataset, the accuracy of CNN is lower than that of SVM [18]
Naïve Bayes	89.2%	[13][30][31][32]	In the same dataset, the accuracy value of NB is lower than that of SVM [8] [12] [21][27]
Decision Tree	81%	[17]	In the same dataset, the accuracy of Decision Tree is lower than that of SVM [8][12]
Random Forest	79%	[9][33]	Using SMOTE can improve the accuracy of Random Forest [33]. Meanwhile, for the same data, the accuracy value of random forest is higher than Naïve Bayes and Decision Tree [9].
IndoBERT	78.99%	[34]	On the same dataset, IndoBERT demonstrates superior accuracy than LGBM Classifier, Support Vector Classifier, LR, XGBoost, AdaBoost, Random Forest and Decision Tree [34]

Vol. 7, No. 2, April 2026, Page. 1007-1027
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5415

XGBoost	75.9%	[6]	With the same data, XGBoost has higher accuracy compared to Naïve Bayes and Random Forest [6]

2. LITERATURE REVIEW

2.1. Sentiment Analysis

Sentiment analysis uses NLP and text analysis techniques to systematically identify, extract information, measure, and study the meaning or opinions from text. [38] Machine learning (ML) and deep learning (DL) approaches are widely used in sentiment analysis. DL uses more complex neural networks than ML for feature extraction, while ML is trained on labeled datasets to recognize textual patterns representing sentiment. [39]. By identifying complex patterns and relationships in technical datasets, techniques such as SVM, Naive Bayes, CNN, and LSTM are well-suited for performing analysis to support data-driven decisions, for example, in marketing, customer service, and politics. [40].

2.2. Pre-Processing Stages

The preprocessing stage includes various techniques such as removing unnecessary parts, for example, through noise reduction, normalization, quality assessment using criteria, and quality protection to verify whether the data is suitable for effective analysis. This step can improve the performance and accuracy of multimodal biometric systems, for example, in terms of feature extraction optimization [41].

2.3. Support Vector Machine (SVM)

SVM is widely used in classification tasks, particularly in sentiment analysis, because it can effectively accommodate high-dimensional spaces, making it suitable for classifying positive, negative, and neutral sentiments in textual data, especially when the data can be linearly separated and can also be adapted for non-linear data using kernel functions, particularly when the dataset is structured [42]. SVM uses both bag-of-words and TF-IDF for feature extraction to find the data points closest to the hyperplane, called support vectors, which are the model's decision boundaries [43]. In SVM, the preprocessing stage influences the model's performance in classification and helps reduce complexity [44].

2.4. BERT dan mBERT

BERT is effectively used for sentiment analysis because it can understand the context of words. Model optimization can be achieved through data cleaning, encoding, and formatting text to match BERT's input, which can then be used with ReLU and SoftMax to generate class probabilities and dropout to prevent overfitting [45]. MBERT (multilingual BERT) is a BERT-based pre-training model developed for various languages using masked language modeling and next sentence prediction, allowing it to understand cross-lingual context and making it suitable for use with limited datasets [46]. MBERT can be used for sentiment analysis on code-mixed text, which is text that combines two or more languages in a single sentence, commonly used on social media. This is achieved through a wordpiece tokenizer and contextual embeddings that understand the meaning of words based on their surrounding context [47].

In the 2024 study by Nabillah et al., "Indonesian Multilabel Classification using IndoBERT Embedding and mBERT Classification," explained that mBERT (Multilingual BERT) is a multilingual version of BERT, including Indonesian. Unlike single-language BERT models, mBERT is generalized across various languages, making it suitable for cross-lingual use. Additionally, mBERT is considered good for applications requiring contextual understanding in diverse linguistic contexts [35].

Vol. 7, No. 2, April 2026, Page. 1007-1027
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5415

3. METHOD

Figure 1 shows the research methodology flowchart, which begins with data collection from the X platform using tweet harvesting, followed by data preprocessing, data labeling with InsetLexicon, feature extraction with TF-IDF for SVM and IndoBERT for BERT, sentiment classification, model evaluation based on the classification results, and finally, a discussion of the sentiment analysis and model evaluation findings.

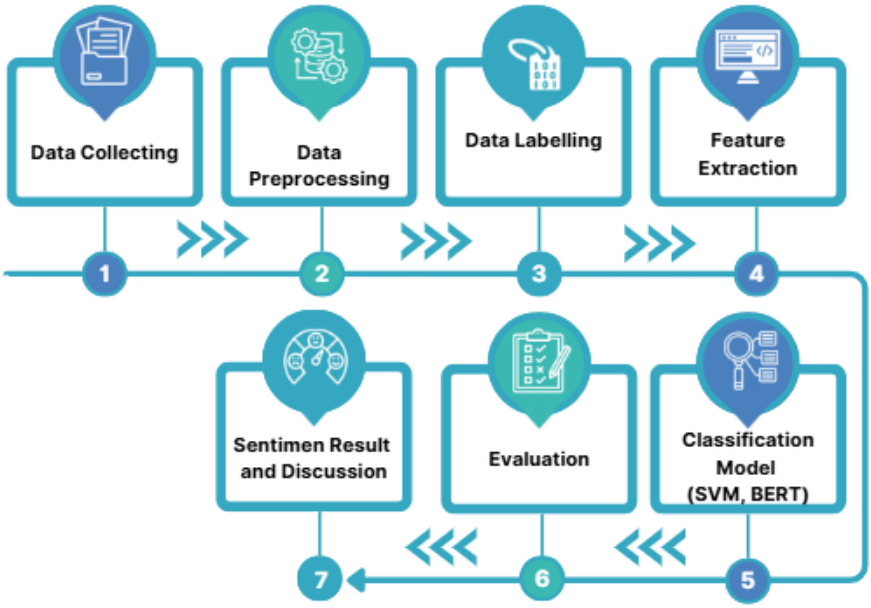

Figure 1. Research Methodology Flow

3.1. Data Collection

Data collection was the first step in the research, followed by preprocessing to ensure its relevance and quality [48]. The data used is uploaded by users on the X platform, a micro-blogging site that allows users to send short messages, often called tweets, which can be used for various research purposes [49]. During the data collection stage, the tweet harvest method was employed, which uses Playwright as an automation framework for the scraping process. The scraping process is carried out with the provision that only public data available on the X platform is collected, excluding personal data such as email addresses, phone numbers, or specific user identities. Instead, only publicly available upload text is utilized. From the scraping results, 3,345 tweet data points were obtained, and after removing duplicates, 2,852 tweets remained. The keywords used can be seen in Table 2. Data was used from the time before and after the PDP Law came into effect. Based on this approach, we set the data collection period as follows: September 22, 2022–November 1, 2022 and July 1, 2024–November 4, 2024. This was done to see public sentiment when the PDP Law was passed, as well as during the period leading up to and after its enactment. This was considered because the PDP Law was passed on September 20, 2022, so data was collected from September to November 2022 to see public opinion after the law was passed. Additionally, the PDP Law will be implemented starting October 17, 2024, so data was collected from 4 months before its enactment until the end of this research.

Table 2. Details of the Tweet Data retrieved

No	Keywords	Total
1	UU PDP	1.544
2	UNDANG-UNDANG PDP	102
3	UU PERLINDUNGAN DATA PRIBADI	1.699

Total	3.345

3.2. Data Labelling

NLTK is a widely used Python library for natural language processing (NLP), including automatic labelling of user uploads in the Twitter application [49]. However, NLTK was originally developed for English, making it less optimal for other languages. Therefore, in this study, we used the InSet Lexicon, which is a lexicon-based approach that classifies text based on predefined sentiment words specifically designed for the Indonesian language [50]. In this study, a threshold of -5 to 5 was used [50], with the provision that a score less than 0 is negative, equal to 0 is neutral, and greater than 0 is positive. The labelling process yielded 1,218 positive tweets, 950 negative tweets, and 684 neutral tweets.

3.3. Data Pre-Processing

Preprocessing processes data to improve the quality and effectiveness of model training, making the model more accurate and efficient. Good preprocessing can reduce existing inconsistencies and noise [51]. In this study, six preprocessing stages were performed: noise removal to eliminate irrelevant elements, case folding to convert text data to lowercase, normalization to change informal/slang words, tokenization to divide sentences into tokens, stopword removal to eliminate common words with no significant meaning, and stemming to extract the root words from the text [35].

3.4. Feature Extraction

To transform features into a new feature group that represents information from data with more concise dimensions, feature extraction is used [52]. For SVM, researchers used the TF-IDF term weighting method, while for BERT, IndoBERT was used.

3.4.1. IndoBERT

The IndoBERT tokenizer converts text into tokens, producing input features and an attention mask as output. This data is then processed by a pre-trained model. IndoBERT can understand context and classify it into predefined categories specifically designed for the Indonesian language [53]. Feature extraction for IndoBERT was performed using the IndoBERT tokenizer, which converts text into tokens and generates input features and attention masks. After tokenization, this data was processed into the IndoBERT model, which had been trained for sentiment classification. This model utilizes the representation of the special [CLS] token as the main feature for identifying the overall context of the text. The final result of the prediction is negative, neutral, or positive sentiment [35].

3.4.2. TF-IDF

TF-IDF combines word frequency, to see how often a word appears in a document, with document frequency, which reduces the significance of common words found in multiple documents. This approach prioritizes terms relevant to the document's context. This research extends TF-IDF by exploring rarely occurring words and demonstrating through SVM-based feature selection experiments a substantial contribution to distinguishing documents describing specific content from documents with general content [54]. TF-IDF calculates the probability of a word appearing in a text as TF, and IDF represents the weight of that word across all documents, as shown in equations 1, 2, and 3 [13].

$$TF - IDF = TF_{td} * IDF_t \quad (1)$$

$$TF\ (term) = \frac{Terms'\ frequency}{Word\ count\ in\ document} \quad (2)$$

Vol. 7, No. 2, April 2026, Page. 1007-1027
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5415

$$IDF\ (term) = \log(\frac{Total\ Data\ Document}{Terms\ in\ whole\ document}) \quad (3)$$

3.5. Sentiment Classification

The sentiment classification used is a ternary classification that categorizes sentiment into positive, negative, and neutral to gain an understanding of the opinions present in the text. Our reasoning is because the neutral category is not completely biased compared to binary classification [55]. In terms of improving accuracy and efficiency, machine learning, deep learning, and hybrid models can be applied, for example, in adjusting the domain of feature representation and imbalanced data [56]. In the classification process, text extraction is crucial for classification accuracy, while the main challenge is limited data sources because the available data is generally in an unstructured format [57]. Classification in this study was performed using SVM and BERT, where BERT can provide dynamic contextual representations to improve classification accuracy [58], while TF-IDF calculates the frequency of words in the dataset, which is then used as input for machine learning models to improve classification accuracy [59].

3.6. Model Evaluation

A confusion matrix is used to evaluate machine learning models, especially classification models. It is a table that compares predicted labels with actual labels for the entire dataset [60]. To ensure the model's performance on the data, the classification model was evaluated using a confusion matrix based on true and false predictions for sentiment classes and by measuring precision, recall, and the F1 matrix [61]. The calculation of precision, recall, accuracy, and F1-score can be seen in the equations (4),(5),(6) and (7) [33].

$$precision = \frac{TP}{TP+FP} \quad (4)$$

$$recall = \frac{TP}{TP+FN} \quad (5)$$

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$F1\ Score = 2\ X\ \frac{precision\ x\ recall}{precision+recall} \quad (7)$$

Additionally, in this study, K-Fold Cross Validation was used to divide the data into several parts, train the model on most of the data while iteratively testing the remaining part, specifically for pre-training IndoBERT. K-fold cross-validation is a supervised learning model performance evaluation method primarily used for hyperparameter optimization. As for the commonly used value of k, it is 5 or 10, considering the balance between estimation error accuracy and computational cost [62].

4. RESULT

4.1. Preprocessing

Initially, the data collection phase used a data range from the date the Personal Data Protection Law came into effect, which is October 17, 2024. However, it was found that the number of existing posts was less than a thousand, and it was feared that the data did not yet represent a public sentiment sample. Subsequently, we expanded the data collection range from the period when the law came into effect, which is from October 17, 2022, to two months after and three months before the official

Vol. 7, No. 2, April 2026, Page. 1007-1027
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5415

implementation. A total of 3,345 posts were collected from the X platform, and after cleaning, 2,852 posts were obtained. Post-labeling is done automatically using the Inset Lexicon. In the labeling process, the number of positive tweets was 1,218, negative tweets 950, and neutral tweets 684. The data was then divided into training, validation, and testing datasets, which were further processed in the classification stage using the BERT and SVM methods. The results obtained from both methods were evaluated using a confusion matrix.

4.2. LDA Topic Modelling

From the sentiment analysis results obtained, topic modeling was performed to map the analysis results we obtained by grouping them based on specific topics using LDA (Latent Dirichlet Allocation) for topic modeling. Figure 2 shows the distribution of topics based on the sentiment obtained, including "data breach," "institution," "digital footprint," "data protection," and "regulation." From this grouping, it was found that the topic of data protection dominates, followed by digital foot print and institution.

4.3. Data Set Ratio Analysis

For our dataset ratio, we used 80:20, which is a common ratio for training and testing data. Additionally, we also tried an 85:15 dataset ratio based on our reference to imbalanced data used in each classification method, namely SVM and BERT [63]. Furthermore, K-fold cross-validation was performed five times using the BERT method.

4.4. Data Set Ratio Analysis

For our dataset ratio, we used 80:20, which is a common ratio for training and testing data. Additionally, we also tried an 85:15 dataset ratio based on our reference to imbalanced data used in each classification method, namely SVM and BERT [63]. Furthermore, K-fold cross-validation was performed five times using the BERT method.

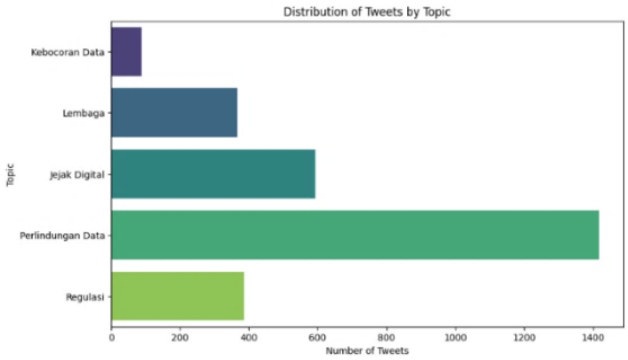

Figure 2. Distribution of Tweets by Topic

4.4.1. SVM Classification

In SVM, feature extraction is performed using TF-IDF. The data is then classified using the SVM model. Based on the classification that has been carried out, the precision, recall, and F1-score values for the 80:20 data are shown in Table 3, while Table 4 shows the values for data with an 85:15 ratio. The comparison between the two datasets shows that increasing the training data from 80% to 85% results in an increase from 76.53% to 76.63%. The most notable change is in the recall value for negative sentiment.

Table 3. Performance Evaluation Results of the SVM Model 80:20

	Precision	***Recall***	**F-1 *Score***
Negative	0.72	0.88	0.79

Vol. 7, No. 2, April 2026, Page. 1007-1027
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5415

Neutral	0.70	0.50	0.59
Positive	0.83	0.83	0.83

Table 4. Performance Evaluation Results of the SVM Model 85:15

	Precision	***Recall***	**F-1 *Score***
Negative	0.69	0.93	0.79
Neutral	0.74	0.45	0.56
Positive	0.85	0.83	0.84

Figure 3 shows the SVM confusion matrix with 80% training data, while Figure 4 shows the confusion matrix with 85% training data. Both models face challenges in accurately classifying neutral sentiment, as evidenced by higher misclassification rates. With the reduced dataset in the testing in Figure 4, negative sentiment is proven to have fewer mislabeled negative instances, while the distribution of neutral and positive sentiments remains largely unchanged.

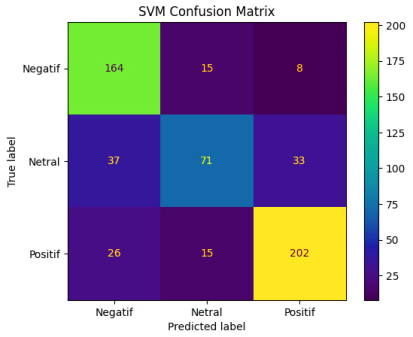

Figure 3 SVM Confusion Matrix 80:20

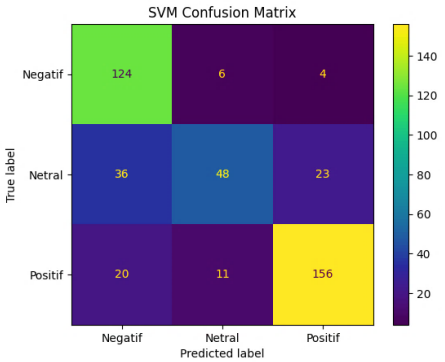

Figure 4. SVM Confusion Matrix 85:15

4.4.2. BERT Classification

Training and validation data were used for the fine-tuning process using IndoBERT, which was then applied to the previously separated test data. During the fine-tuning process, the researchers applied k-fold cross-validation with k=5, where for each fold, BERT generated model output. On average, the evaluation results from K-Fold Cross Validation yielded an accuracy of 0.9109 and an F1 score of 0.9105. The evaluation results from the fine-tuning process with 80% of the training and validation data can be seen in Table 5, while the confusion matrix for this data can be seen in Figure 5.

Table 5 Performance Evaluation Results of the BERT Model 80:20

K	***Precision***	***Recall***	**F-1 *Score***	***Accuracy***

Vol. 7, No. 2, April 2026, Page. 1007-1027
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5415

1	0.78	0.78	0.78	0.78
2	0.85	0.85	0.85	0.85
3	0.96	0.96	0.96	0.96
4	0.97	0.97	0.97	0.97
5	0.99	0.99	0.99	0.99

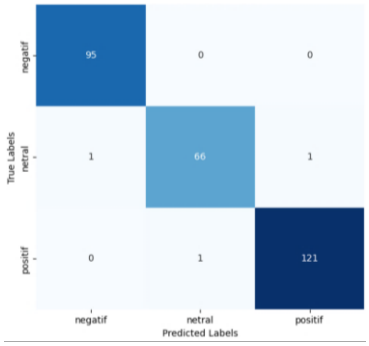

Figure 5. BERT fine tuning Confusion Metrix 80:20

From the model that has been created, testing is performed on 20% of the test data, with the precision, recall, and F-1 scores shown in Table 6. The BERT precision, recall, and f-1 Scores, while the confusion matrix can be seen in Figure 6.

Table 6. BERT Evaluation Score 80:20

	Precision	***Recall***	**F-1 *Score***
Negative	0.76	0.74	0.75
Neutral	0.56	0.64	0.60
Positive	0.86	0.81	0.84

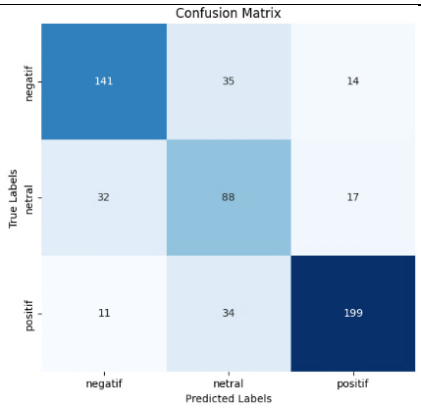

Figure 6. BERT Confusion Metrix 80:20

Because the achieved accuracy was considered insufficient, a different dataset composition was used, allocating 85% of the training and validation data and 15% of the test data [63]. However, the k-fold cross-validation applied in testing the BERT method remained at K=5.

Table 7. Performance Evaluation Results of the BERT Model 85:15

K	***Precision***	***Recall***	**F-1 *Score***	***Accuracy***
1	0.71	0.71	0.71	0.71
2	0.85	0.85	0.85	0.85

Vol. 7, No. 2, April 2026, Page. 1007-1027
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5415

3	0.95	0.95	0.95	0.95
4	0.99	0.99	0.99	0.99
5	0.99	0.99	0.99	0.99

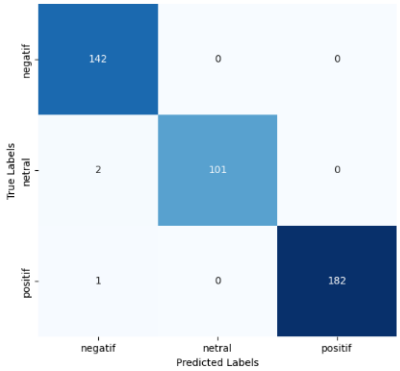

Figure 7. BERT Confusion Metrix 85:15

Table 7 displays the model evaluation, while Figure 7 shows the confusion matrix from the fine-tuning process with an 85:15 data combination. This indicates that the chosen division is more suitable for imbalanced datasets, as previously explained in the literature, which shows that this ratio yields better results. Next, the test data classification shown in Table 8 was performed for its precision, recall, and F-1 score values, while the confusion matrix can be seen in Figure 8.

Table 8. BERT Evaluation Score 85:15

	Precision	*Recall*	*F-1 Score*
Negative	0.88	0.85	0.86
Neutral	0.75	0.77	0.76
Positive	0.91	0.91	0.91

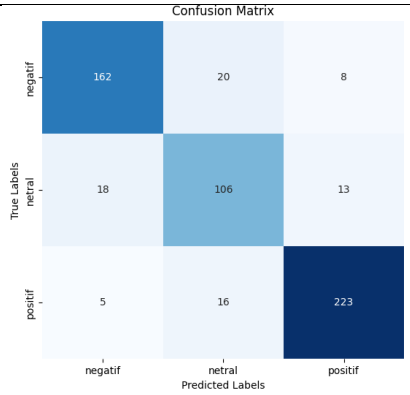

Figure 8. BERT Confusion Metrix 85:15

4.5. Classification Modeling Results

Overall, BERT showed better performance than SVM in both data-splitting scenarios. BERT's average F1-score was 0.91 ± 0.08 for 80:20 data splitting and 0.89 ± 0.11 for 85:15, while SVM showed lower average F1-scores of 0.74 ± 0.12 and 0.73 ± 0.12. Additionally, the accuracy values for SVM and BERT are shown in Table 9.

Table 9. Accuracy

	Data Train and Validation	*Accuracy*
SVM	80%	76.53%

Vol. 7, No. 2, April 2026, Page. 1007-1027
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5415

	85%	76.63%
BERT	80%	75%
	85%	86%

Table 9 shows that the size of the training data composition affects the classification results. It can be seen that in the comparison of training and test data at 85:15, both BERT and SVM have an increase in accuracy, with BERT being superior in this case with an accuracy value of 86%. Additionally, the availability of Indonesian language libraries that can understand context affects the accuracy of the classification performed. In this regard, BERT supports multilingual languages and has IndoBERT to support the context-based classification process, while in previous studies, SVM was considered good enough to perform sentiment analysis classification on laws in effect in Indonesia on the X platform [63].

5. DISCUSSIONS

5.1. Sentiment Analysis Content

Based on the analysis conducted, it was concluded that 43% had a positive sentiment, 33% had a negative sentiment, and 24% had a neutral sentiment, as shown in Figure 9. Meanwhile, example sentences from the posts can be seen in Table 10.

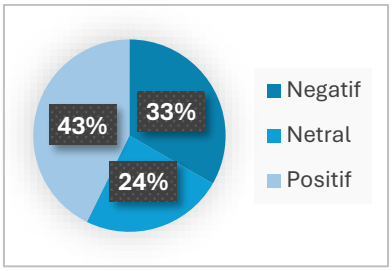

Figure 9. Sentiment on UU PDP

Based on the sentiment analysis percentages above, here are examples of posts on the X platform with the topic of data protection in positive, negative, and neutral sentiments, which can be seen in Table 10.

Table 10. Examples of Sentiment Sentences

Positive	kepatuhan terhadap UU Perlindungan Data Pribadi kita menciptakan lingkungan digital yang aman transparan dan bertanggung jawab. Yuk patuhi UU Perlindungan Data Pribadi biar info pribadi kita tetap aman dan gak disalahgunakan!
Neutral	Tanggal 18 Oktober 2024 akan menjadi hari pertama UU Perlindungan Data Pribadi (UU PDP) mulai berlaku setelah ditetapkan dan disahkan pada 17 Oktober 2022
Negative	Ayok buktikan mana uu nya bunyi nya gimana penerapan nya gimana terus gimana sama uu pdp Apa cuma omon doamg tuh uu Wkwkwkwk gaada harganya tuh uu

Based on the analysis conducted, the unique words that most frequently appear in negative sentiment, as shown in Figure 10, include "kena," "diskusi," "kpu," "salah," "legal," "cloud," "badan," "instansi," "kayak," dan "pilih" . Negative sentiment focuses on criticizing the suitability of the regulations that have been implemented. Meanwhile, in the positive sentiment shown in Figure 11, the

Vol. 7, No. 2, April 2026, Page. 1007-1027
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5415

most frequently occurring unique words include "komitmen," "pemutakhirdanatapelanggan","acceleratingrenewableenergy," "dukung," "langkah," "syukur," "keren," "nyaman," "dalam," dan "pres". Positive Sentiment, The public is focused on supporting and hoping that the PDP Law can provide strong legal protection. In the neutral sentiment shown in Figure 12, the most frequently occurring unique words include "enkripsi," "amsi," "ai," "bisnis," "maju," "hadap," "pers," "tribun," "peran," dan "etika". The public tends to be indifferent. One of the most discussed topics related to neutral sentiment is the establishment of an independent body for this PDP Law.

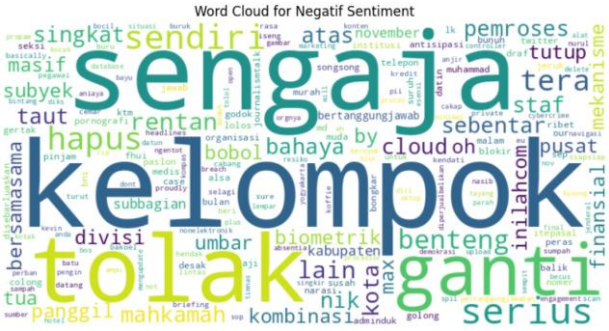

Figure 10. Word Cloud Negative Sentiment

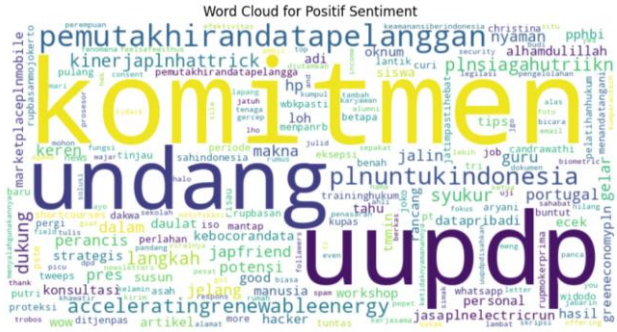

Figure 11. Word Cloud Positive Sentiment

Figure 12. Word Cloud Neutral Sentiment

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 1007-1027
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5415

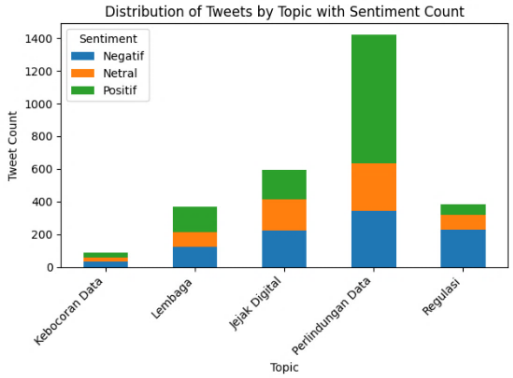

Figure 13. Distribution of Tweets by Topic with Sentiment Count

Figure 13 shows that the topic of data protection has 55.25% positive sentiment from 1,419 tweets, and institutions have 41.96% positive sentiment from a total of 313 tweets. On the other hand, the topic of regulation is dominated by negative sentiment, with 58.80% of the 386 tweets grouped under this topic expressing dissatisfaction or concern. For the topics of digital footprint and data breaches, both are relatively balanced between positive and negative sentiment obtained, although negative sentiment is slightly higher, with 37.44% negative sentiment from 593 tweets for digital footprint and 39.08% negative sentiment from 87 tweets for the data breaches topic.

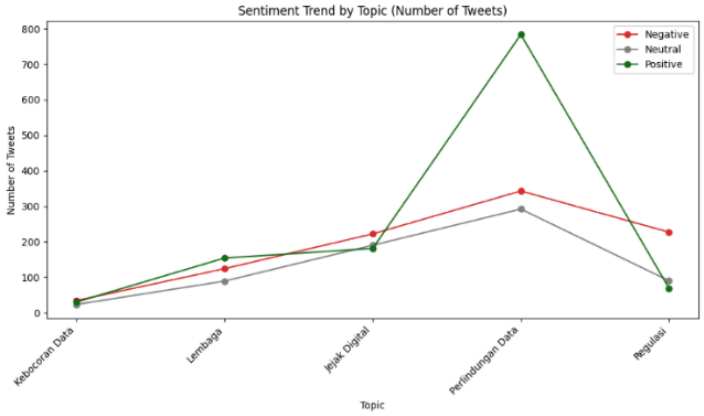

Figure 14. Sentiment Trend by Topic

Trend analysis in Figure 14 shows that the topic of data protection generally receives a positive response from the public. Conversely, the topic of regulation raises significant concerns, requiring policy strengthening to be well-received. For the topic of institutions, sentiment is relatively balanced with a slight positive bias. Meanwhile, regarding digital footprints, concerns have emerged about the risk of illegal use of personal data. The topic of data breaches shows a relatively balanced distribution of sentiment, although the number of analyzed tweets is limited.

However, both the sentiment analysis and the trends obtained previously have the potential for temporal bias and limitations due to scraping being done only on the X platform. This is because the data was obtained within a specific time range, so public opinion or sentiment will tend to be influenced by currently trending topics. For example, in the sentiment analysis conducted, data was collected within a time range after the Personal Data Protection Law (PDP Law) was passed and the PDP Law began to be implemented. Therefore, the opinions obtained tend to be unstable, and these opinions are not based on experience implementing the PDP Law but rather on public opinion regarding its passage and implementation.

Vol. 7, No. 2, April 2026, Page. 1007-1027
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5415

In the study conducted by Suragih et al. in 2021, data collection was carried out during two periods of PSBB because sentiment is time-bound [22]. For example, in previous studies, there were several studies on policies implemented during the pandemic/COVID-19, and it can be seen that the composition of public sentiment tended to differ across the time period from 2020 to 2021. For instance, it could be positive [9][16][19][24][30][25], negative [5][10][22][23][33] or neutral [6][17][64]. Additionally, different specific topics also influence sentiment; for example, sentiment on the topic of vaccines [10] may differ from sentiment on PSBB [16].

5.2. Classification Model

NLP is already widely used in various countries for sentiment analysis related to policies. For example, in Korea, KcBERT is used to perform sentiment analysis on the low birth rate policy in South Korea, using comments from the YouTube platform, with an accuracy of 82% [65]. In Bangladesh, sentiment analysis related to mob justice is conducted on the Facebook platform using DistilBERT, with an F1-score of 93% [66], Meanwhile, in India, several sentiment analyzes have been performed, such as using DistilBERT on Twitter, Facebook, online discussion platforms, and surveys with an accuracy of 92% [67], RNN and Transformer on Google Forms, social media, and news articles with an accuracy of 83.3% [68] Bi-LSTM on the Twitter plat(Soreng & Bandhu, 2025)f 92.8% [69]. In Indonesia, NLP is also used for sentiment analysis, for example, in green economy policies [70] and biodiversity [34]. Previous research shows that NLP performs well in a variety of languages, particularly for BERT, and is applicable on a variety of platforms.

In previous research to analyze sentiment in Indonesian biodiversity policy tweets, several traditional models were compared, and it was found that IndoBERTweet performed best with an accuracy of 78.99% [34]. This is in direct contrast to the research conducted by the author, however, in this study, the author used IndoBERT, and it was found that BERT outperformed SVM. This is because BERT can understand context in multiple languages. Although it has lower accuracy than BERT, SVM has advantages in computational efficiency and faster processing, making it suitable for environments with limited computing power, while BERT requires a longer fine-tuning and computational process.

5.3. Recommendation

Based on the sentiment analysis and modeling conducted by the researcher, several recommendations are provided, including strengthening digital security through periodic audits to ensure compliance with the PDP Law. Furthermore, education, including public awareness campaigns for various elements of society and government, needs to be enhanced. In addition, the establishment of an independent institution to oversee the implementation of the PDP Law needs to be prioritized because it is one of the issues of public concern, and mandatory security certification needs to be applied to companies that manage personal data. Continuous periodic evaluations are expected to maintain the alignment of the PDP Law with technological developments and community needs, particularly in terms of sanctions, which are still perceived as weak by the public, and to make the regulations more adaptable and increase public trust.

Based on the research conducted, the researchers recommend using a training-to-test data ratio of 85:15, which is suitable for imbalanced data composition, and selecting a method that can understand language context and support multiple languages to achieve better results. Furthermore, this sentiment analysis can be implemented as a real-time monitoring tool for public sentiment regarding policies that have been or will be implemented by the government, serving as a basis for evaluating and improving policies to align with community conditions and needs. Additionally, the researchers recommend using hybrid methods from machine learning and deep learning as the foundation for sentiment analysis to improve accuracy.

Vol. 7, No. 2, April 2026, Page. 1007-1027
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5415

6. CONCLUSION

Based on the sentiment analysis conducted, it was found that public sentiment tends to be positive at 43%, while negative sentiment is 33% and neutral sentiment is 24%. The topics generally discussed are data protection, regulation, institutions, data breaches, and digital footprints. Meanwhile, in terms of accuracy measurement, it was found that BERT performed better with an accuracy of 86% compared to SVM at 76.63% with a training and testing data composition of 85:15. This advantage is due to IndoBERT's ability, which has been optimized for the Indonesian language and supports context-based classification. Conversely, the performance of SVM is highly dependent on the text preprocessing stage, but it still offers advantages in terms of speed and relatively low computational requirements. This indicates that using transformer-based models is more relevant for analyzing Indonesian-language text. From the research conducted, it provides insight that NLP can be further used as the basis for algorithms or frameworks in sentiment analysis, particularly in examining public response or opinion toward government policies, and can serve as a database for decision-making, for example, by being implemented in a real-time monitoring system.

This study has limitations because the data only covers public uploads within a specific period and is only conducted on Platform X. Therefore, future research is recommended to expand the data scope to various platforms (cross-platform data), apply temporal analysis, and consider demographic segmentation and advanced NLP methods to make the sentiment analysis results regarding the PDP Law more comprehensive.

ACKNOWLEDGEMENT

The author expresses appreciation and gratitude to the Ministry of Communications and Digital for the financial support provided, enabling this research to be carried out successfully.

REFERENCES

[1] M. Amin *et al.*, "Security and privacy awareness of smartphone users in Indonesia," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, May 2021. doi: 10.1088/1742-6596/1882/1/012134.

[2] I. G. N. P. Widiatedja and N. Mishra, "Establishing an independent data protection authority in Indonesia: a future–forward perspective," *International Review of Law, Computers and Technology*, vol. 37, no. 3, pp. 252–273, 2023, doi: 10.1080/13600869.2022.2155793.

[3] Republik Indonesia, *Undang-Undang (UU) Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi*. Indonesia: Kementerian Komunikasi dan Digital RI (JDHI Komdigi), 2022. Accessed: Dec. 08, 2025. [Online]. Available: https://jdih.komdigi.go.id/produk_hukum/view/id/832/t/undangundang+nomor+27+tahun+2022

[4] CNBC Indonesia, "Lembaga Perlindungan Data Pribadi Belum Dibentuk, Ini Kata Angga Raka." Accessed: Nov. 03, 2024. [Online]. Available: https://www.cnbcindonesia.com/tech/20241021174853-37-581818/lembaga-perlindungan-data-pribadi-belum-dibentuk-ini-kata-angga-raka

[5] H. Al Jannah and D. Hermawan, "Analysis of Indonesian Society's Perceptions of the COVID-19 Vaccine in Youtube Comments Using Machine Learning Algorithms," in *2022 3rd International Conference on Artificial Intelligence and Data Sciences: Championing Innovations in Artificial Intelligence and Data Sciences for Sustainable Future, AiDAS 2022 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 72–77. doi: 10.1109/AiDAS56890.2022.9918796.

[6] M. N. Fakhruzzaman, S. Z. Jannah, S. W. Gunawan, A. I. Pratama, and D. A. Ardanty, "IndoPolicyStats: sentiment analyzer for public policy issues," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 1, pp. 482–489, Feb. 2024, doi: 10.11591/eei.v13i1.5263.

Vol. 7, No. 2, April 2026, Page. 1007-1027
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5415

[7] Z. D. W. Putra, "The Sentiments of Indonesian Urban Citizens Regarding the Lockdown-Like Policy During the COVID-19 Pandemic: A Path Towards an Urban E-Planning Process in a Pandemic Situation INTRODUCTION," *International Journal of E-Planning Research*, vol. 11, no. 1, 2022, doi: 10.4018/IJEPR.297515.

[8] E. A. Sukma, A. N. Hidayanto, A. I. Pandesenda, A. N. Yahya, P. Widharto, and U. Rahardja, "Sentiment Analysis of the New Indonesian Government Policy (Omnibus Law) on Social Media Twitter," in *Proceedings - 2nd International Conference on Informatics, Multimedia, Cyber, and Information System, ICIMCIS 2020*, Institute of Electrical and Electronics Engineers Inc., Nov. 2020, pp. 153–158. doi: 10.1109/ICIMCIS51567.2020.9354287.

[9] P. Maulana, I. Budi, and A. Budi Santoso, "Sentiment Analysis of Indonesian Government's Effort to Overcome the Unemployment Problem during COVID-19 Pandemic," in *ICOIACT 2022 - 5th International Conference on Information and Communications Technology: A New Way to Make AI Useful for Everyone in the New Normal Era, Proceeding*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 144–149. doi: 10.1109/ICOIACT55506.2022.9971853.

[10] D. S. A. Maylawati, M. T. S. Bilhaq, A. Wahana, D. R. Ramdania, E. Nurlatifah, and M. A. Ramdhani, "Analysis of Changes in Sentiment towards COVID-19 Vaccination in Indonesia Using the Convolutional Neural Network," in *2023 11th International Conference on Cyber and IT Service Management, CITSM 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/CITSM60085.2023.10455282.

[11] M. Pilliang, H. Akbar, and G. Firmansyah, "Sentiment Analysis for Super Applications in Indonesia: A Case Study of Gov2Go App," in *Proceedings of the International Conference on Electrical Engineering and Informatics*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 80–85. doi: 10.1109/IConEEI55709.2022.9972291.

[12] M. A. Khadija, I. S. D. Jayanti, and F. U. Nimah, "Towards Smart City: Aspect Based Sentiment Analysis of Indonesian Public Aspiration Complaints Data Using Machine Learning," in *Proceedings - International Conference on Informatics and Computational Sciences*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 215–220. doi: 10.1109/ICICoS62600.2024.10636859.

[13] D. Gunawan, S. N. Awwal, and F. N. H. Afif, "Sentiment Analysis of Personal Data Protection and Privacy Law in Indonesia Using Multinomial Naive Bayes Algorithm," in *2025 International Conference on Smart Computing, IoT and Machine Learning, SIML 2025*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/SIML65326.2025.11080902.

[14] D. Dennis *et al.*, "Sentiment Classification Against the Public Activity Restrictions Policy in Jakarta Using Machine Learning Models," Institute of Electrical and Electronics Engineers (IEEE), Feb. 2022, pp. 5–9. doi: 10.1109/icaibda53487.2021.9689741.

[15] D. Sugiarto, E. Utami, A. Yaqin, and S. Raharjo, "Sentiment Analysis of Cooking Oil Price Policy in Indonesia using Word2Vec Feature Extraction and Bidirectional Long Short Term Memory," in *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application, ICAICTA 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICAICTA59291.2023.10389977.

[16] L. Sandra and S. Aritonang, "Lockdown Countdown: Lockdown Sentiment Analysis on Twitter Using Artificial Neural Network," in *2021 International Conference on Data Science and Its Applications, ICoDSA 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 198–202. doi: 10.1109/ICoDSA53588.2021.9617212.

[17] L. Alfat, N. Uddin, M. Nasucha, M. Mohamad, and S. Masrom, "Analysis of Indonesian Opinion on the Arrival of COVID-19 Vaccine Using Machine Learning," in *Proceedings - 11th International Conference on Information Technology, Computer and Electrical Engineering, ICITACEE 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 326–331. doi: 10.1109/ICITACEE62763.2024.10761958.

[18] N. Hasanati, T. S. Utami, and R. H. Kusumaningtyas, "Sentiment Analysis on News Headlines of Nation's Capital Relocation Using CNN and SVM," in *2023 11th International Conference on Cyber and IT Service Management, CITSM 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/CITSM60085.2023.10455228.

Vol. 7, No. 2, April 2026, Page. 1007-1027
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5415

[19] S. F. Pane, J. Ramdan, A. G. Putrada, M. N. Fauzan, R. M. Awangga, and N. Alamsyah, "A Hybrid CNN-LSTM Model with Word-Emoji Embedding for Improving the Twitter Sentiment Analysis on Indonesia's PPKM Policy," in *Proceeding - 6th International Conference on Information Technology, Information Systems and Electrical Engineering: Applying Data Sciences and Artificial Intelligence Technologies for Environmental Sustainability, ICITISEE 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 51–56. doi: 10.1109/ICITISEE57756.2022.10057720.

[20] B. Waspodo, Q. Aini, F. R. Singgih, R. H. Kusumaningtyas, and E. Fetrina, "Support Vector Machine and Lexicon based Sentiment Analysis on Kartu Prakerja (Indonesia Pre-Employment Cards Government Initiatives)," in *2022 10th International Conference on Cyber and IT Service Management, CITSM 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/CITSM56380.2022.9935990.

[21] R. L. V. Nyoto and Y. Ruldeviyani, "Infiltration Wells Program in Jakarta: Twitter Sentiment Analysis," in *2022 1st International Conference on Information System and Information Technology, ICISIT 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 352–357. doi: 10.1109/ICISIT54091.2022.9872911.

[22] P. S. Saragih, D. Witarsyah, F. Hamami, and J. M. MacHado, "Sentiment Analysis of Social Media Twitter with Case of Large Scale Social Restriction in Jakarta using Support Vector Machine Algorithm," in *2021 International Conference Advancement in Data Science, E-Learning and Information Systems, ICADEIS 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICADEIS52521.2021.9701961.

[23] I. Firmansyah, M. H. Asnawi, S. A. Hasanah, R. Novian, and A. A. Pravitasari, "A Comparison of Support Vector Machine and Naïve Bayes Classifier in Binary Sentiment Reviews for PeduliLindungi Application," in *2021 International Conference on Artificial Intelligence and Big Data Analytics, ICAIBDA 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 140–145. doi: 10.1109/ICAIBDA53487.2021.9689771.

[24] G. Kanugrahan and A. F. Wicaksono, "Sentiment Analysis of Face-to-face Learning during Covid-19 Pandemic using Twitter Data," in *Proceedings - 2021 8th International Conference on Advanced Informatics: Concepts, Theory, and Application, ICAICTA 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICAICTA53211.2021.9640282.

[25] M. Rahardi, A. Aminuddin, F. F. Abdulloh, and R. A. Nugroho, "Sentiment Analysis of Covid-19 Vaccination using Support Vector Machine in Indonesia," 2022. doi: 10.14569/IJACSA.2022.0130665.

[26] D. F. Sebastian, H. Sulistiani, and A. R. Isnain, "SENTIMENT ANALYSIS OF PUBLIC OPINION ON THE RIGHT OF INQUIRY IN INDONESIA IN 2024 USING THE SUPPORT VECTOR MACHINE (SVM) METHOD," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 4, pp. 1025–1034, Jul. 2024, doi: 10.52436/1.jutif.2024.5.4.1968.

[27] T. Widyanto, I. Ristiana, and A. Wibowo, "Komparasi Naïve Bayes dan SVM Analisis Sentimen RUU Kesehatan di Twitter," *SINTECH Journal (Science and Information Technology Journal)*, vol. 6, no. 3, pp. 147–161, 2023, doi: 10.31598/sintechjournal.v6i3.1433.

[28] I. Nurma Yulita *et al.*, "Bidirectional Long Short-Term Memory for Analysis of Public Opinion Sentiment on Government Policy During the COVID-19 Pandemic," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 14, no. 11, p. 2023, 2023, doi: 10.14569/IJACSA.2023.0141189.

[29] A. L. Qosim, F. Kurniawan, U. Bahruddin, Z. Mubaraq, Suhartono, and M. Faisal, "Analysis Classification Opinion of Policy Government Announces Cabinet Reshuffle on YouTube Comments Using 1D Convolutional Neural Networks," in *3rd 2021 East Indonesia Conference on Computer and Information Technology, EIConCIT 2021*, Institute of Electrical and Electronics Engineers Inc., Apr. 2021, pp. 30–35. doi: 10.1109/EIConCIT50028.2021.9431884.

[30] I. H. Hidayat, R. E. Parwanto, and Rudy, "Sentiment Analysis on the Perception and Mindset of the People of Indonesia on the Use of Vaccines to Deal with the Covid-19 Pandemic using the Text Mining Method," in *Proceedings of 2022 International Conference on Information Management and Technology, ICIMTech 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 57–61. doi: 10.1109/ICIMTech55957.2022.9915072.

Vol. 7, No. 2, April 2026, Page. 1007-1027
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5415

[31] M. A. F. Bunyamin, T. H. Pudjiantoro, F. Renaldi, and A. I. Hadiana, "Analyzing Sentiments on Indonesia's New National Palace using The Combination of Naive Bayes and Sentiment Scoring," in *2022 International Conference on Science and Technology, ICOSTECH 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICOSTECH54296.2022.9829075.

[32] A. S. Setiawan, H. H. Nuha, and M. W. A. Bawono, "Classification of Sentiment Analysis Against Omnibus Law on Twitter Social Media and News Websites Using the Naïve Bayes Method," in *2022 5th International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 141–147. doi: 10.1109/ISRITI56927.2022.10052796.

[33] M. R. Pribadi, D. Manongga, H. D. Purnomo, I. Setyawan, and Hendry, "Sentiment Analysis of the PeduliLindungi on Google Play using the Random Forest Algorithm with SMOTE," in *2022 International Seminar on Intelligent Technology and Its Applications: Advanced Innovations of Electrical Systems for Humanity, ISITIA 2022 - Proceeding*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 115–119. doi: 10.1109/ISITIA56226.2022.9855372.

[34] M. T. Uliniansyah, A. Jarin, A. Santosa, and Gunarso, "Modeling sentiment analysis of Indonesian biodiversity policy Tweets using IndoBERTweet," *IAES International Journal of Artificial Intelligence*, vol. 14, no. 3, pp. 2389–2401, Jun. 2025, doi: 10.11591/ijai.v14.i3.pp2389-2401.

[35] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 1, pp. 1071–1078, 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.

[36] H. Imaduddin, F. Yusfida A'la, and Y. S. Nugroho, "Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, p. 113, 2023, doi: 10.14569/IJACSA.2023.0140813.

[37] A. Riyadi, M. Kovacs, U. Serdült, and V. Kryssanov, "IndoGovBERT: A Domain-Specific Language Model for Processing Indonesian Government SDG Documents," *Big Data and Cognitive Computing*, vol. 8, no. 11, Nov. 2024, doi: 10.3390/bdcc8110153.

[38] G. R. Usha and L. Dharmanna, "Sentiment Analysis on Business Data using Machine Learning," in *2021 2nd International Conference on Smart Technologies in Computing, Electrical and Electronics, ICSTCEE 2021 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICSTCEE54422.2021.9708593.

[39] L. Mahalakshmi and E. Anbalagan, "National Language Processing for Sentiment Analysis in Social Media-A Comprehensive Review," in *Proceedings of International Conference on Circuit Power and Computing Technologies, ICCPCT 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 504–508. doi: 10.1109/ICCPCT61902.2024.10672661.

[40] S. Sindhu, S. Kumar, and A. Noliya, "A Review on Sentiment Analysis using Machine Learning," *International Conference on Innovative Data Communication Technologies and Application, ICIDCA 2023 - Proceedings*, pp. 138–142, 2023, doi: 10.1109/ICIDCA56705.2023.10099665.

[41] R. Sharma, J. Sandhu, and V. Bharti, "Experimental Analysis of a Multimodal biometric System using Preprocessing and Feature Extraction Techniques and Their Impact on Analytical Results," *Proceedings - 2024 6th International Conference on Computational Intelligence and Communication Technologies, CCICT 2024*, pp. 212–219, 2024, doi: 10.1109/CCICT62777.2024.00043.

[42] S. A. Dzulkifli, M. N. M. Salleh, and K. H. Talpur, "Improved Weighted Learning Support Vector Machines (SVM) for High Accuracy," *ACM International Conference Proceeding Series*, pp. 40–44, 2019, doi: 10.1145/3372422.3372432.

[43] S. Thenappan, A. R. Krishnan, P. S. B. Murugan, A. S. Valarmathy, P. Chitra, and H. Sayyed, "Machine Learning Approaches for Sentiment Analysis in Social Media Data," in *7th International Conference on Electronics, Communication and Aerospace Technology, ICECA*

Vol. 7, No. 2, April 2026, Page. 1007-1027
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5415

2023 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 684–688. doi: 10.1109/ICECA58529.2023.10395612.

[44] X. Zhipeng, M. A. A. Aziz, and N. A. Razak, "Performance Evaluation of Support Vector Machine Algorithm in Object Classification Using Different Preprocessing Methods," in *2024 IEEE International Conference on Automatic Control and Intelligent Systems, I2CACIS 2024 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 169–174. doi: 10.1109/I2CACIS61270.2024.10649625.

[45] N. Al Hafidh and A. Al-Karawi, "Advanced Sentiment Analysis of Amazon Electronics Reviews Leveraging BERT: Model Optimization and Evaluation," in *Procedia Computer Science*, Elsevier B.V., 2025, pp. 3608–3618. doi: 10.1016/j.procs.2025.04.616.

[46] V. A. Manwar and A. B. Manwar, "mBERT: A Query Refinement Model for Marathi Word Sense Disambiguation," in *2025 IEEE International Students' Conference on Electrical, Electronics and Computer Science, SCEECS 2025*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/SCEECS64059.2025.10940800.

[47] S. K. Singh, A. Sharma, Sahil, D. Singh, S. Pandit, and U. Saghir, "Sentiment Analysis of English-Hindi Code-Mixed Text Using mBERT Model," in *Proceedings of the 2025 3rd International Conference on Inventive Computing and Informatics, ICICI 2025*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 552–556. doi: 10.1109/ICICI65870.2025.11069692.

[48] K. K. J, V. Saraswathi P, and R. T, "Advancing Sentiment Analysis with LLM: Comparative Study with Traditional ML on Social Media Reviews," Institute of Electrical and Electronics Engineers (IEEE), Jul. 2025, pp. 1328–1333. doi: 10.1109/iccsp64183.2025.11088527.

[49] Z. Yang and H. Men, "Natural Language Processing of COVID-19 Reports Involving China in New York Times - a Machine-based Framing Study of Media Language," *ACM International Conference Proceeding Series*, pp. 137–143, 2022, doi: 10.1145/3582768.3582785.

[50] F. Koto and G. Y. Rahmaningtyas, "InSet Lexicon: Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs," IEEE, 2017. doi: 10.1109/IALP.2017.8300625.

[51] C. P. Chai, "Comparison of text preprocessing methods," *Nat Lang Eng*, vol. 29, no. 3, pp. 509–553, 2023, doi: 10.1017/S1351324922000213.

[52] J. Li, M. S. Othman, H. Chen, and L. M. Yusuf, "Optimizing IoT intrusion detection system: feature selection versus feature extraction in machine learning," *J Big Data*, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-00892-y.

[53] I. K. D. Nuryana, L. I. D. Mawarni, and E. Juanara, "Early Detection of Environmental Issues from Social Media using IndoBERT and LDA: Case Study of Pollution and Deforestation in Indonesia," in *E3S Web of Conferences*, EDP Sciences, Aug. 2025. doi: 10.1051/e3sconf/202564505005.

[54] Y. Okumura, S. Hirokawa, and K. Takeuchi, "Significance of Low-Frequent Words in Concept Describing Document," *Proceedings - 2019 8th International Congress on Advanced Applied Informatics, IIAI-AAI 2019*, pp. 1035–1036, 2019, doi: 10.1109/IIAI-AAI.2019.00214.

[55] E. S. Alamoudi and N. S. Alghamdi, "Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings," *J Decis Syst*, vol. 30, no. 2–3, pp. 259–281, 2021, doi: 10.1080/12460125.2020.1864106.

[56] N. Keivandarian and M. Carvalho, "A Survey on Sentiment Classification Methods and Challenges," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, Association for Computational Linguistics (ACL), 2019, pp. 4171–4186. doi: 10.32473/flairs.36.133314.

[57] T. J. Sefara, M. Mbooi, K. Mashile, T. Rambuda, and M. Rangata, "A Toolkit for Text Extraction and Analysis for Natural Language Processing Tasks," in *5th International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems, icABCD 2022 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/icABCD54961.2022.9856269.

[58] T. Zhang and R. Zhang, "Revealing the power of BERT for text sentiment classification," in *4th IEEE International Conference on Automation, Electronics and Electrical Engineering,*

Vol. 7, No. 2, April 2026, Page. 1007-1027
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5415

AUTEEE 2021, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 14–17. doi: 10.1109/AUTEEE52864.2021.9668704.

[59] H. Pal and B. Bhushan, "Sentiment Analysis on Twitter Dataset using Voting Classifier," in *2024 International Conference on Electrical, Electronics and Computing Technologies, ICEECT 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICEECT61758.2024.10739316.

[60] J. Görtler, F. Hohman, D. Moritz, M. Kirchner, and K. Patel, "Confusion matriks," *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, May 2022, doi: 10.1145/3491102.3501823.

[61] N. A. P. Masaling, T. Lubis, A. Amalia, A. R. Lubis, and M. S. Lydia, "Category Based Sentiment Analysis for Basic Skin-Cares Using LDA and SVM Approach," *Proceeding - ELTICOM 2022: 6th International Conference on Electrical, Telecommunication and Computer Engineering 2022*, pp. 182–189, 2022, doi: 10.1109/ELTICOM57747.2022.10037876.

[62] S. Yerramilli and D. W. Apley, "Fractional Cross-Validation for Optimizing Hyperparameters of Supervised Learning Algorithms," *Technometrics*, 2025, doi: 10.1080/00401706.2025.2515926.

[63] S. A. Ajagbe, J. B. Awotunde, and H. Florez, "Intrusion Detection: A Comparison Study of Machine Learning Models Using Unbalanced Dataset," *SN Comput Sci*, vol. 5, no. 8, 2024, doi: 10.1007/s42979-024-03369-0.

[64] I. Nurma Yulita *et al.*, "Bidirectional Long Short-Term Memory for Analysis of Public Opinion Sentiment on Government Policy During the COVID-19 Pandemic," 2023. [Online]. Available: www.ijacsa.thesai.org

[65] S. Lee, R. H. Ali, T. A. Khan, M. Karunanithi, I. Ahmad, and R. Kouatly, "Analysing Public Perception of South Korea's Low Birth Rate Policies Using NLP-based Sentiment Analysis," in *SIST 2025 - 2025 IEEE 5th International Conference on Smart Information Systems and Technologies, Conference Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/SIST61657.2025.11139252.

[66] F. Akter, T. T. Aurpa, F. Islam, M. S. Islam, Hossneara, and M. Ashrafuzzaman, "Sentiment Analysis of Public Perception Toward Mob Justice," in *2025 IEEE International Conference on Quantum Photonics, Artificial Intelligence, and Networking, QPAIN 2025*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/QPAIN66474.2025.11171704.

[67] P. Sankireddy, N. S. Ramireddy, A. R. Penumada, J. Pechetti, and N. C. Nair, "Analyzing Public Feedback on Urban Infrastructure Projects," in *2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation, IATMSI 2025*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/IATMSI64286.2025.10984847.

[68] V. Joshi, N. Yede, P. Parihar, C. Dhule, R. Agrawal, and N. C. Morris, "Quantifying Effectiveness of Governmental Agriculture Policies using RNN-Transformer based Sentiment Analysis," in *3rd International Conference on Intelligent Data Communication Technologies and Internet of Things, IDCIoT 2025*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 1–7. doi: 10.1109/IDCIOT64235.2025.10915148.

[69] A. Soreng and K. C. Bandhu, "A Bi-LSTM and Attention-Based Sentiment Classifier for Enhancing Public Trust in COVID-19 Vaccination," in *2025 World Skills Conference on Universal Data Analytics and Sciences, WorldSUAS 2025*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/WorldSUAS66815.2025.11199276.

[70] W. R. E. Putri, S. Suhendro, R. Azhar, N. Desriani, and A. C. Pramana, "Exploring Public Sentiment on Green Economy Policy: A Natural Language Processing-Based Analysis," *International Journal of Energy Economics and Policy*, vol. 15, no. 2, pp. 560–565, Feb. 2025, doi: 10.32479/ijeep.18360.