Jurnal Mandiri IT
Vol. 15 No. 2, October (2026), pp. 80-91

ISSN 2301-8984 (Print), 2809-1884 (Online)
Published by Institute of Computer Science

Sentiment analysis of public opinion toward government regulations concerning the implementation of tabungan perumahan rakyat (Tapera) using the convolutional neural network (CNN) model on social media X

M. Iqbal Noviansyah[1], Kusrini[2]
[1.2]Magister Informatika, Universitas Amikom Yogyakarta, Indonesia

A R T I C L E I N F O

Article history:

Received Jul 28, 2026
Revised Aug 3, 2026
Accepted Aug 13, 2026

Keywords:

Convolutional Neural Network;
InSet Lexicon;
Sentiment Analysis;
Tapera;
Word2Vec.

ABSTRACT

Sentiment analysis is one of the applications of *Natural Language Processing* (NLP) for identifying public opinion on social media. Although previous studies on Tapera sentiment analysis have applied conventional machine learning and transformer-based approaches, limited attention has been given to integrating automatic lexicon-based sentiment labeling with semantic word representation and deep learning classification within a unified framework. This study proposes an approach that integrates the *Indonesian Sentiment Lexicon* (InSet), *Word2Vec*, and *Convolutional Neural Network* (CNN) to classify public sentiment toward the Tabungan Perumahan Rakyat (Tapera) policy on social media X. A total of 1,291 tweets were processed through the preprocessing stage and automatically labeled using InSet. *Word2Vec* was employed to construct semantic word representations that were used as the initial weights of the CNN embedding layer. The proposed model was evaluated using four hold-out data split scenarios and compared with *BiLSTM* and *Random Forest*. The experimental results showed that the 80:20 configuration achieved the best performance with an accuracy of 74.52%, outperforming *BiLSTM* (71.43%) and *Random Forest* (69.88%). The novelty of this study lies in integrating automatic sentiment labeling using the *Indonesian Sentiment Lexicon* (InSet), semantic representation through *Word2Vec*, and CNN-based classification within a unified framework, together with an empirical evaluation of different train-test split ratios for Tapera policy sentiment analysis. These findings demonstrate that the proposed approach provides competitive sentiment classification performance while reducing dependence on manual sentiment annotation.

Corresponding Author:

M. Iqbal Noviansyah,
Magister Informatika,
Universitas Amikom Yogyakarta,
Jl. Ring Road Utara, Ngringin, Condongcatur, Kec. Depok, Kab. Sleman, Daerah Istimewa Yogyakarta, 55281, Indonesia
Email: iqbal2473@students.amikom.ac.id

1. INTRODUCTION

Fulfilling the need for decent housing is still one of the development challenges in Indonesia. Population growth accompanied by urbanization has led to an increase in the need for housing, while the availability of land and access to housing financing are still limited. This condition has resulted in an increase in the housing backlog, especially in urban areas, and encouraged the emergence of less livable settlements. Low-income people also face obstacles in obtaining home financing due to relatively strict credit requirements, so the home ownership gap is still a national development challenge (Asril et al., 2022; Sabitha, 2022). To overcome these problems, the

government established the Program Tabungan Perumahan Rakyat (Tapera) through Government Regulation of the Republic of Indonesia Number 21 of 2024 as an effort to expand access to housing financing for the community. Although it aims to improve people's welfare, the implementation of Tapera has given rise to various responses related to contribution obligations, transparency in fund management, and its impact on the community's economic conditions so that it has become one of the public policies that is widely discussed on social media (Tania et al., 2021).

The development of social media, such as *X (Twitter), Facebook, Instagram, YouTube*, and *TikTok*, has changed the way people express their opinions on various government policies. The volume of data generated every day is so large that social media becomes a rich source of data to understand public perception. However, the characteristics of unstructured data, the use of informal language, abbreviations, emojis, and writing variations make the manual analysis process inefficient and vulnerable to subjectivity. Therefore, a computational approach is needed that is able to systematically process large-scale text data (Fauzy, 2022). One of the widely used approaches is sentiment analysis, which is a technique in the field of *Natural Language Processing* (NLP) which classifies opinions into positive, negative, or neutral categories. The information produced can be used to measure public perception of a policy so that it becomes an evaluation material for decision makers. To obtain accurate classification results on social media text data, an algorithm is needed that is able to understand language characteristics and patterns of relationships between words.

Various studies have applied machine learning algorithms for sentiment analysis. *Support Vector Machine* (SVM) is one of the most widely used methods because it performs well in classifying high-dimensional data. However, its performance reached an accuracy of only 59% (Kurniasari & Fatta, 2021). I In the case of the Tapera policy, algorithms including *K-Nearest Neighbor, Support Vector Machine* (SVM), *Naïve Bayes*, and *Decision Tree* have been compared in previous studies, while other studies have evaluated the performance of SVM and *Random Forest* using data from platform X (Leidiyana et al., 2024; Musthafa et al., 2025). Both studies showed that SVM provides the best performance, with an accuracy of approximately 88%; however, its effectiveness still depends on feature engineering and the quality of the feature representations used (Leidiyana et al., 2024; Musthafa et al., 2025). As deep learning technology has evolved, the Convolutional Neural Network (CNN) has become increasingly popular because it can automatically learn feature representations through a convolution process without requiring complex feature engineering. Various studies have demonstrated that CNN achieves competitive performance, with accuracies of 85% (Gupta et al., 2021), 87.72% (Gandhi et al., 2021), and 92.90% on the Yelp dataset and 91.70% on the IMDB dataset (Cheng et al., 2020).

CNN performance is strongly influenced by the quality of the text representation used as input to the model. Therefore, word embedding techniques are widely applied to convert text into numerical vector representations that preserve semantic relationships between words. One of the most widely used methods is *Word2Vec*, which learns relationships between words based on their contextual usage, enabling semantically similar words to be represented in nearby vector spaces. Compared with other methods, such as *GloVe* and *FastText*, *Word2Vec* has relatively low computational complexity while still producing effective semantic representations. Dense word vector-based representations have been shown to improve the performance of various text classification models compared with word frequency-based representations (Çano & Morisio, 2019).

Although research on Tapera sentiment analysis has begun to emerge, several limitations remain in previous studies. Most studies have relied on conventional machine learning algorithms or transformer-based models. *IndoBERT* Lite has been applied to classify YouTube comments with an accuracy of 78% (Firdaus & Trisnawarman, 2025), whereas *Support Vector Machine* (SVM) achieved the best performance, with an accuracy of approximately 88%, in previous studies (Leidiyana et al., 2024; Musthafa et al., 2025). However, research integrating *lexicon-based sentiment analysis*, *Word2Vec*, and *Convolutional Neural Network* (CNN) for Tapera policy sentiment analysis remains limited. In addition, previous studies have generally evaluated their models using a single train-test split, leaving the effect of different data partitioning ratios on model performance insufficiently explored. These limitations indicate a research gap in the integration of automatic *lexicon-based sentiment labeling*, semantic word representation, and deep learning classification for Tapera policy sentiment analysis. To address these gaps, this study proposes a

sentiment analysis framework for public opinion on the Tapera policy using data collected from social media X by integrating the *InSet Lexicon, Word2Vec*, and CNN. The novelty of this study lies in combining automatic sentiment labeling using the *InSet Lexicon*, semantic representation through *Word2Vec*, and CNN-based classification within a unified framework, while empirically evaluating model performance across four train-test split scenarios (90:10, 80:20, 70:30, and 60:40). The proposed model is evaluated using accuracy, precision, recall, and F1-score to assess its classification performance under different data partitioning conditions. Thus, this study contributes to Tapera policy sentiment analysis by providing an integrated approach that reduces dependence on manual sentiment labeling and offers empirical evidence regarding the effect of train-test split ratios on CNN-based sentiment classification.

2. RESEARCH METHOD

This study uses a quantitative approach with the types of experimental research, computational laboratory and case studies (Sugiyono, 2019; Yin, 2017). The nature of the research is evaluative to measure the performance of the algorithm *Deep Learning* in the domain of sentiment analysis (Creswell & Creswell, 2017). The phenomenon that is studied in depth is the public perception of the government's policy regarding Tabungan Perumahan Rakyat (TAPERA). This research consists of five main stages, namely dataset collection, data preprocessing, sentiment labeling using *Lexicon-Based Sentiment Analysis (InSet Lexicon),* Feature representation using *Word2Vec*, as well as the classification and evaluation of models using *Convolutional Neural Network (CNN).*

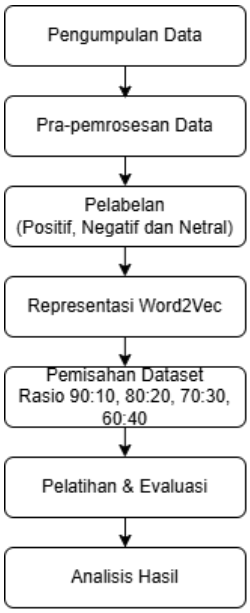

Figure 1. Sentiment analysis system flowchart

Dataset Collection

The dataset used is secondary data obtained from *the Kaggle repository* and consists of 1,291 tweets about the Tabungan Perumahan Rakyat (Tapera) policy on social media X. *full_text* attribute is used as the main source of analysis because it contains user opinions on the policy. The dataset was selected because it specifically contains public discussions related to Tapera and provides text instances suitable for sentiment classification. The dataset also contains positive, negative, and neutral sentiment categories after the lexicon-based labeling process, enabling the evaluation of a three-class classification model. Accordingly, the 1,291 tweets are treated as a representation of the observed public discourse contained in the selected dataset rather than as a statistically representative sample of the entire Indonesian population.

Data Pre-processing

The preprocessing stage aims to reduce *Noise*, standardize word forms, and produce more consistent representations of text, improving feature quality before sentiment labeling and classification processes. The stages carried out include *cleaning, case folding, normalization,*

tokenization, stopword removal, and *Vote* Using algorithms *Sastrawi*. This stage is a process commonly used in Indonesian sentiment analysis because it is able to produce a more consistent representation of text and improve the performance of classification models (Setiawan, 2024).

Data Labeling

Sentiment labeling on 1,291 tweet data was carried out automatically using a dictionary-based approach (*Lexicon-based approach*). This approach was chosen because it is considered more efficient and consistent for processing the selected dataset than manual labeling, which can be time-consuming and subject to annotator differences (Taboada et al., 2011). This study uses a lexicon of sentiments in Indonesian, such as *InSet (Indonesian Sentiment Lexicon*), which compiles a list of words and their polarity weights (Purwitasari et al., 2023). The resulting labels were then used as reference labels for supervised CNN training.

The labeling process is done by matching each word (*Stuart T*) in a tweet with a lexical dictionary. The total sentiment score of a tweet is calculated based on the accumulated polarity weight of the constituent word. The criteria for determining sentiment classes are set based on thresholds (*threshold*) polarity score as follows (S): Positive: If the polarity score accumulates $S > 0$. Negative: If the polarity score accumulates $S < 0$. Neutral: If the accumulation of polarity scores (no sentimental words or positive and negative word weights neutralize each other) $S = 0$.

***Word2Vec* Feature Representation**

After the sentiment labeling process, each word is represented into vector form using an algorithm *Word2Vec* With architecture *Skip-Gram*. *Word2Vec* Studying the semantic relationships between words based on the context in which they arise so that words that have similar meanings or contexts will have adjacent vector representations in the embedding space. This vector representation is then used as a matrix embedding on the CNN model to improve the model's ability to extract text features (Atandoh et al., 2023; Tang et al., 2023). So that each word is represented in a fixed-dimensional vector space that can be processed by the CNN network.

In *the Skip-Gram architecture*, the probability of the occurrence of a context word w_O based on the target word w_I is calculated using the following *Softmax function*.

$$P(w_O \mid w_I) = \frac{\exp(v'^{T}_{w_O} v_{w_I})}{\sum_{w=1}^{V} \exp(v'^{T}_{w} v_{w_I})} \quad (1)$$

With: v_{w_I} = input vector of the target word, v'_{w_O} = output vector of the context word, V = vocabulary size.

Furthermore, the sentence representation is formed from a set of word vectors generated by *Word2Vec*. If a sentence consists of words, then its representation can be expressed as n

$$X = [x_1, x_2, x_3, \ldots, x_n] \quad (2)$$

With: X = matrix *embedding* sentences, $x_i \in \mathbb{R}^d$ = *Word2Vec vector* dimensioned d for the word to-i, d = dimension *embedding.*

The *embedding* X matrix is then used as input to the *Embedding Layer* in the CNN architecture for feature extraction and sentiment classification.

Data Analysis and Modeling Methods

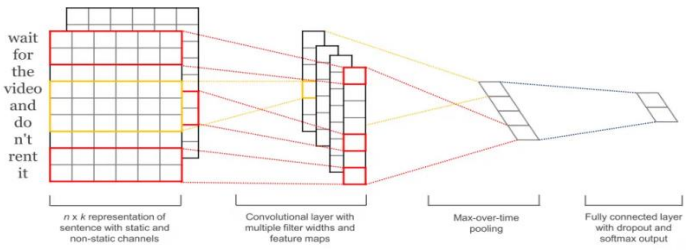

Figure 2. Architecture convolutional neural network

The *Convolutional Neural Network* (CNN) model is used to classify sentiment based on the word representation *of Word2Vec*. The model architecture consists of *the Input Layer, Embedding Layer, Conv1D Layer, Global Max Pooling Layer, Dropout Layer, Dense Layer*, and *Output Layer with* the Softmax *activation function*, as shown in Figure 2.

The Input Layer receives a sequence of padded tokens of fixed length (MAX_LEN). *The Embedding Layer* is initialized using the *Word2Vec* trained *embedding matrix* with vector dimensions of 100 (*trainable = True*). Furthermore, *Conv1D* uses 64 filters with a kernel size of 5 and *a ReLU* activation function to extract local features. The extraction results were summarized using *Global Max Pooling*, then *a Dropout* of 0.5 was applied to reduce *overfitting*. The next feature is processed by *the Dense Layer* which consists of 64 *neurons* with ReLU *activation*, followed by *a Dropout* of 0.3. In the *Output Layer*, 3 neurons with *Softmax activation* are used to generate a probability of three cential classes.

Convolution operations are expressed as follows.

$$c_i = f(W \cdot x_{i:i+h-1} + b) \quad (3)$$

by being W is a convolutional filter, x is a word vector, b is a biased, and f is an activation function *ReLU*. Models compiled using optimizers *Adam* with a learning rate of 0.0003, the loss function *categorical crossentropy*, and evaluation metrics *accuracy*. Instead of relying on a single train-test split, this study employed four hold-out scenarios (90:10, 80:20, 70:30, and 60:40) to investigate the robustness and generalization capability of the proposed model under different proportions of training and testing data. Varying the split ratio allows the model to be evaluated across different learning conditions, where a larger training set may improve feature learning while a larger testing set provides a more rigorous assessment of generalization performance. This experimental design also enables the identification of the most appropriate data partition for sentiment classification on relatively small Indonesian-language datasets. Evaluation was carried out using the *hold-out validation* with four data sharing scenarios, namely 90:10, 80:20, 70:30, and 60:40. to determine the effect of the proportion of training data on model performance. Use of combinations *Word2Vec* and CNN has been shown to improve the accuracy of sentiment classification over conventional text representations (Fahry et al., 2023).

Model Evaluation and Result Analysis

The classification performance was tested on the test data using the *confusion matrix tool*. The effectiveness of the model is assessed based on four key mathematical metrics:

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

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

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

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

The confusion matrix is used to calculate *the values of True Positive* (TP), *True Negative* (TN), *False Positive* (FP), and *False Negative* (FN). Based on these values, *the Accuracy, Precision, Recall,* and *F1-score metrics* are calculated to evaluate the model's ability to classify the three sentiment classes.

3. RESULTS AND DISCUSSIONS

Pre Processing Data

The dataset used in this study was obtained from *Kaggle* and consisted of 1,291 tweets regarding the Tebungan Perumahan Rakyat (Tapera) policy on social media X (Twitter). The *full_tex*t attribute is used as the primary source of analysis because it contains the user's opinion of the policy. The data obtained is still in the form of raw data with unstructured characteristics of social media text, such as the use of non-standard language, abbreviations, symbols, emojis, links (URLs), usernames, and other special characters. Therefore, the preprocessing data stage is

carried out to produce cleaner, consistent, and structured data so that it can improve the quality of text representation and the performance of the sentiment classification model (Fahry et al., 2023; Setiawan, 2025).

a. Cleaning Process, cleaning is carried out to remove URLs, usernames, symbols, numbers, punctuation, as well as case folding and normalizing words so that data is cleaner and more consistent. The results of the cleaning process are shown in Table 1.

Table 1. Process results cleaning

Full_text	Cleaning
BP TAPERA siap membantu tenaga kerja di Indonesia untuk memiliki hunian yang layak dan terjangkau https://t.co/xeiotqU4Qo	bp tapera siap membantu tenaga kerja di indonesia untuk memiliki hunian yang layak dan terjangkau
@hharuluv Tapera X mau diblokir ganti pakai aplikasi yang namanya gajelas dollar udah 16k ini lagi ada yang kasih wacana korban judol dikasih bansos	tapera kali mau diblokir ganti pakai aplikasi yang namanya dollar sudah ke ini lagi ada yang kasih wacana korban judol dikasih bansos
tapera tu bneran bakal ada kh .	tapera itu benaran bakal ada kah

b. Tokenizing Process, tokenizing aims to break down *the text of a tweet* into a set of words (*tokens*) that will be used in the next stage of analysis. The *tokenizing* results are shown in Table 2.

Table 2. Process results tokenizing

Cleaning	Tokenizing
BP TAPERA siap membantu tenaga kerja di Indonesia untuk memiliki hunian yang layak dan terjangkau	['bp', 'tapera', 'siap', 'membantu', 'tenaga', 'kerja', 'di', 'indonesia', 'untuk', 'memiliki', 'hunian', 'yang', 'layak', 'dan', 'terjangkau']
Tapera X mau diblokir ganti pakai aplikasi yang namanya gajelas dollar udah k ini lagi ada yang kasih wacana korban judol dikasih bansos	['tapera', 'kali', 'mau', 'diblokir', 'ganti', 'pakai', 'aplikasi', 'yang', 'namanya', 'dollar', 'sudah', 'ke', 'ini', 'lagi', 'ada', 'yang', 'kasih', 'wacana', 'korban', 'judol', 'dikasih', 'bansos']
tapera tu bneran bakal ada kh	['tapera', 'itu', 'benaran', 'bakal', 'ada', 'kah']

c. *Stopword Removal* Process
Stopword removal is done to remove common words that do not have a significant influence on sentiment. The results of this process are shown in Table 3.

Table 3. Process Results *Stopword Removal*

Tokenizing	*Stopword Removal*
['bp', 'tapera', 'siap', 'membantu', 'tenaga', 'kerja', 'di', 'indonesia', 'untuk', 'memiliki', 'hunian', 'yang', 'layak', 'dan', 'terjangkau']	['bp', 'tapera', 'membantu', 'tenaga', 'kerja', 'indonesia', 'memiliki', 'hunian', 'layak', 'terjangkau']
['tapera', 'kali', 'mau', 'diblokir', 'ganti', 'pakai', 'aplikasi', 'yang', 'namanya', 'dollar', 'sudah', 'ke', 'ini', 'lagi', 'ada', 'yang', 'kasih', 'wacana', 'korban', 'judol', 'dikasih', 'bansos']	['tapera', 'kali', 'diblokir', 'ganti', 'aplikasi', 'namanya', 'dollar', 'kasih', 'wacana', 'korban', 'judol', 'dikasih', 'bansos']
['tapera', 'itu', 'benaran', 'bakal', 'ada', 'kah']	['tapera', 'benaran', 'kah']

d. Stemming Process, stemming is done to change the word suffix into a basic form using the Literary library. The results of the voting process are shown in Table 4.

Table 4. Process results vote

Stopword Removal	Stemming
['bp', 'tapera', 'membantu', 'tenaga', 'kerja', 'indonesia', 'memiliki', 'hunian', 'layak', 'terjangkau']	bp tapera bantu tenaga kerja indonesia milik huni layak jangkau
['tapera', 'kali', 'diblokir', 'ganti', 'aplikasi', 'namanya', 'dollar', 'kasih', 'wacana', 'korban', 'judol', 'dikasih', 'bansos']	tapera kali blokir ganti aplikasi nama dollar kasih wacana korban judol kasih bansos
['tapera', 'benaran', 'kah']	tapera benar kah

Lexical Feature Mapping

Sentiment labeling was carried out using the *Lexicon-Based Sentiment Analysis* method by utilizing the Indonesian sentiment dictionary *InSet Lexicon*. Each word preprocessing result is matched with a dictionary of positive and negative words to determine the sentiment tendencies *of*

the tweet. *Tweets* are then classified into positive sentiment (*score > 0*), negative (*score < 0*), and neutral (*score = 0*). An example of sentiment labeling results is shown in Table 5.

Table 5. Sentiment score calculation

Full text	Score	Sentiment
bp tapera bantu tenaga kerja indonesia milik huni layak jangkau	2	Positive
tapera kali blokir ganti aplikasi nama dollar kasih wacana korban judol kasih bansos	-4	Negatives
tapera benar kah	0	Neutral

After the process *Lexicon-Based Sentiment Labeling*, each tweet is classified into positive, negative, or neutral sentiment based on value *score_lex*. The distribution of the number and percentage of each sentiment category is presented in Figure 3.

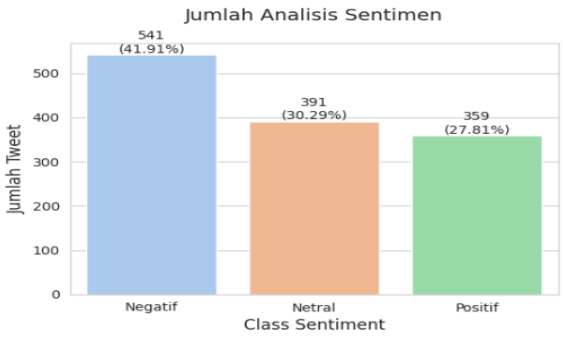

Figure 3. Sentiment label distribution

The results of the sentiment distribution showed that negative sentiment dominated with 541 tweets (41.91%), followed by neutral sentiment with 391 tweets (30.29%), and positive sentiment with 359 tweets (27.81%). The dominance of negative sentiment indicates that the observed public discourse in the dataset was more critical toward the Tapera policy. This pattern can be associated with concerns discussed around the policy, particularly mandatory contribution requirements, transparency in fund management, and its potential impact on people's economic conditions. Such issues can encourage critical responses on social media, where users can directly express disagreement or concern. Therefore, the negative sentiment distribution provides an indication of the concerns present in the observed online discourse, rather than implying that all Indonesian citizens hold the same view of the policy. This distribution is the basis for the training and evaluation process of the CNN model to classify public sentiment.

***Word2Vec* Representation**

Word representations are built using *the Word2Vec* method with the *Skip-Gram architecture*. The model was trained using *a vector size* of 100, *a window size* of 5, *min_count* of 1, *epochs* of 20, and *a seed* of 42 to produce a semantic representation of words based on the context of their appearance in the corpus *of tweets*. Before the training process, the tweets of *the polling* results are tokenized to form vocabulary and converted into numerical sequences. The results of *the Word2Vec* training generated 12,480 tokens with 3,199 unique words that make up the model's vocabulary. Vocabulary statistics are shown in Table 6.

Table 6. Vocabulary statistics

Remarks	Quantity
Total Token	12.480
Vocabulary Size	3.199
Unique Words	3.199

Visualization of representations *Word2Vec* Using *t-distributed Stochastic Neighbor Embedding (t-SNE)* shown in Figure 4. The visualisation results show that the word "tapera" is in the same environment as the words "rumah", "beli", "kredit", "pensiun", "kelola", "cair", "jangkau", and "sulit". The proximity shows that *Word2Vec* successfully learn semantic relationships based on the context of the occurrence of words in the corpus *tweet*. Most of the words around "tapera"

represent the main topics of discussion, namely housing financing, fund management, and various public opinions on the implementation of Tapera's policies. Thus, the representation *Word2Vec* able to produce *embedding* which maintains the contextual relationship between words and is subsequently used as an initial weight (*embedding matrix*) on *Embedding Layer* model CNN.

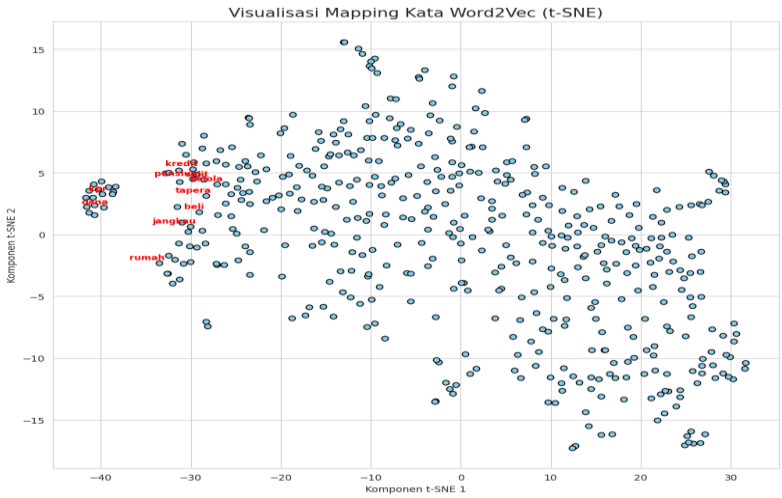

Figure 4. t-SNE visualization

Next, the vectors of *the Word2Vec* training are arranged into an embedding matrix, which is used as the *initial weights* in the *Embedding Layer* CNN model. By utilizing these semantic representations, CNN can extract features that are more informative than word-frequency-based representations, thus supporting the sentiment classification process more effectively.

Empirical Classification and Model Performance

Four experimental scenarios were conducted to evaluate the performance of the model *Convolutional Neural Network* (CNN) in classifying the sentiment of public opinion on the Public Housing Savings (Tapera) policy. Evaluation was carried out using the *hold-out validation* with four data sharing configurations, namely 90:10, 80:20, 70:30, and 60:40. The entire experiment uses the CNN architecture, the representation *Word2Vec*, as well as the same training parameters so that the difference in performance is only affected by the proportion of data sharing.

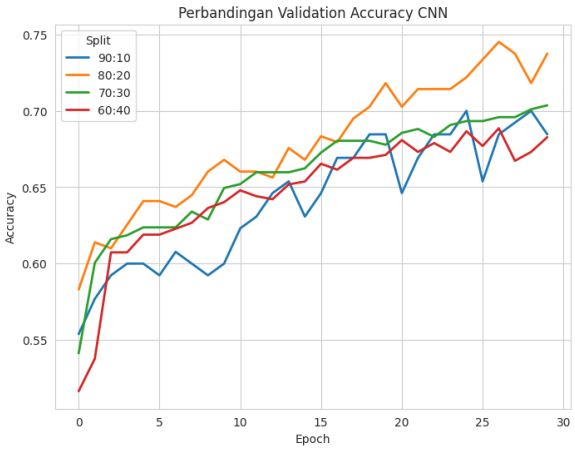

Figure 5**.** Comparison of CNN model accuracy of each data split

The results of CNN's evaluation of the four data-sharing configurations are shown in Table 7. The 80:20 configuration delivered the best performance with an accuracy of 74.51%, macro precision of 74.50%, macro recall of 73.27%, and macro F1-score of 73.74%. Meanwhile, the

90:10, 70:30, and 60:40 configurations resulted in accuracies of 68.46%, 69.59%, and 67.31%, respectively. The superior performance of the 80:20 configuration indicates a more appropriate balance between training capacity and test-set size for this dataset. The 90:10 configuration provides more training data but leaves a relatively small test set, which may provide a less stable representation of the diversity of tweet characteristics. Conversely, the 70:30 and 60:40 configurations allocate fewer samples for training, potentially limiting the CNN's ability to learn diverse semantic and local patterns. Thus, the 80:20 configuration provides sufficient training data while retaining an adequately sized test set for evaluating generalization.

Table 7. Comparing CNN's performance

Experiment	Train : Test	Accuracy	Precision	Recall	F1-Score
Experiment 1	90 : 10	68.46%	0.6819	0.6846	0.6809
Experiment 2	80 : 20	74.51%	0.7458	0.7451	0.7440
Experiment 3	70 : 30	69.58%	0.6963	0.6958	0.6942
Experiment 4	60 : 40	67.31%	0.6754	0.6731	0.6728

To improve the objectivity of the evaluation, the CNN model was compared with two baseline models, namely *Bidirectional Long Short-Term Memory* (BiLSTM) and *Random Forest*. The results of the comparison are shown in Table 8.

Table 8. Performance Comparison of CNN with Baseline Models

Model	Best Split	Accuracy	Precision	Recall	F1-Score
CNN (Proposed)	80:20	74.52%	0.7458	0.7452	0.7441
BiLSTM	80:20	71.43%	0.7400	0.7143	0.7204
Random Forest	80:20	69.88%	0.7231	0.6988	0.7009

The *BiLSTM* model achieved its best performance in the 80:20 configuration, with an accuracy of 71.43%, macro precision of 73.10%, macro recall of 71.20%, and macro F1-score of 71.43%. In the 70:30 and 60:40 configurations, *BiLSTM* obtained accuracies of 69.59% and 69.83%, respectively, while the accuracy in the 90:10 configuration was 65.38%. These results indicate that *BiLSTM* provides competitive performance across different data partitioning scenarios, although its performance remains lower than that of the proposed CNN model in the best-performing configuration.

As a machine learning-based comparison model, *Random Forest* demonstrated relatively stable performance across the four data-sharing configurations. Its highest accuracy was obtained in the 80:20 configuration at 69.88%, while the 90:10, 70:30, and 60:40 configurations resulted in accuracies of 68.46%, 67.53%, and 68.47%, respectively. Although Random Forest showed relatively consistent performance, its accuracy remained lower than those of both deep learning-based models, particularly CNN.

Overall, the comparison results show that the proposed CNN achieved the highest performance in the 80:20 configuration, with an accuracy of 74.52%, compared with 71.43% for *BiLSTM* and 69.88% for Random Forest. The higher performance of CNN suggests that its convolution-based feature extraction is effective in identifying local patterns in the tweet representations used in this study. In particular, *Word2Vec* provides dense semantic representations that serve as the initial weights of the CNN embedding layer, allowing the convolutional filters to process contextual relationships among words rather than relying solely on discrete word representations.

Discussion and Error Identification

Based on the evaluation results, the CNN model achieved the best performance in the 80:20 data sharing configuration with an accuracy of 74.52%, precision of 74.58%, recall of 74.52%, and an F1-score of 74.41%. These results show that the combination of word representation using *Word2Vec* and local feature extraction through convolutional operations is able to produce good classification performance on the sentiment dataset regarding Tapera policies.

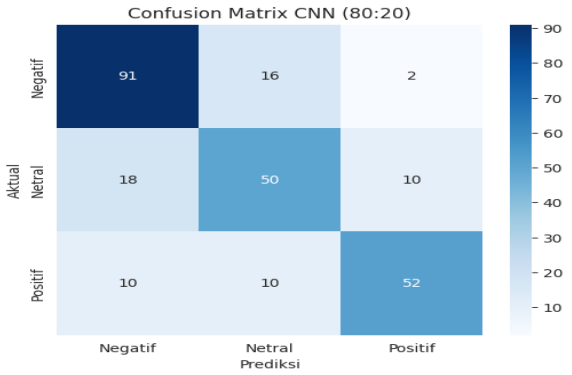

Figure 6. Confusion matrix CNN

To analyze the behavior of the model in more detail, an evaluation was carried out using a confusion matrix, as shown in Figure 6. Based on these results, CNN managed to correctly classify 91 negative tweets, 50 neutral tweets, and 52 positive tweets. The negative class had the highest number of correct predictions, indicating that the model was able to recognize linguistic patterns that contained negative opinions consistently. This is because most negative tweets use vocabulary that has a strong polarity, such as words related to rejection, objection, or criticism of Tapera's policies.

In contrast, misclassification occurs most in neutral classes. A total of 18 neutral tweets were predicted to be negative, while another 10 tweets were predicted to be positive. This condition suggests that neutral tweets have more ambiguous characteristics than the other two classes. Many neutral tweets only convey information or opinions that do not contain strong emotional expressions, so the line between neutral sentiment and positive or negative sentiment becomes less clear. As a result, the features CNN learned from *the Word2Vec* representation still found similarities in context to other classes.

In the positive class, the model managed to correctly classify 52 tweets, while some positive tweets were still predicted as negative or neutral. The error indicates that some positive tweets use word choices that don't explicitly indicate positive sentiment, but rather are hopes, suggestions, or indirect support. This language pattern causes the representation of features between classes to overlap, increasing the likelihood of misclassification.

Word2Vec may have contributed to the performance of the proposed CNN by representing semantically related words in nearby vector spaces. The model was trained using a 100-dimensional embedding with a *Skip-Gram architecture*, enabling words to be represented according to their contextual relationships within the tweet corpus. These representations provide the CNN with richer input features for extracting local semantic patterns. The *Word2Vec* visualization also showed that the word 'tapera' was located in a similar semantic context to words such as 'rumah', 'beli', 'kredit', 'pensiun', 'kelola', 'cair', 'jangkau', and 'sulit', reflecting several themes discussed in the Tapera-related tweets. Consequently, the combination of *Word2Vec* and CNN provides a semantically informed representation and convolution-based feature extraction mechanism that may explain the competitive performance of the proposed model. However, because this study did not include an ablation experiment comparing CNN with and without *Word2Vec*, the specific contribution of *Word2Vec* to the accuracy improvement cannot be isolated. Future studies should therefore conduct an ablation analysis to quantify the individual contribution of the embedding representation.

However, a number of misclassifications were still found that were influenced by the characteristics of the data and labeling methods. This study uses *the Indonesian Sentiment Lexicon* to automatically generate sentiment labels. This approach provides a fast and consistent labeling process, but is not yet able to understand complex language contexts, such as negation, irony, sarcasm, and implicit meaning. As a result, some tweets have the potential to acquire labels that do not match their true meaning, thus influencing the learning process of the CNN model.

As a validation of the proposed model, CNN's performance was compared to *BiLSTM* and *Random Forest*. The test results showed that CNN obtained the highest accuracy of 74.52%, higher than *BiLSTM* (71.43%) and *Random Forest* (69.88%) in the best data sharing configuration.

In addition, *the confusion matrix* of the two comparison models also shows a lower number of correct predictions than CNN. These findings indicate that the combination of *Word2Vec* and CNN is more effective in capturing local patterns and semantic relationships between words in tweet data than the comparison model used.

From a practical perspective, the findings demonstrate that sentiment analysis can serve as a complementary tool for monitoring public responses to government policies. The predominance of negative sentiment in the observed dataset indicates that issues related to policy communication, contribution mechanisms, transparency, and perceived economic burden require greater attention. Continuous monitoring of public opinion on social media could help policymakers identify emerging concerns, evaluate the effectiveness of policy communication, and formulate more responsive policy improvements. Sentiment analysis should therefore be used as a supporting source of evidence alongside formal policy evaluation and other stakeholder inputs.

Overall, the results of the study show that CNN is able to provide a good balance between feature extraction capabilities and generalization capabilities on sentiment classification tasks. Although it still faces difficulties in distinguishing tweets with neutral characteristics, the proposed model managed to provide the best performance among all the models evaluated and is therefore suitable for use as an approach for sentiment analysis of public opinion regarding Tapera policies.

4. CONCLUSION

This study succeeded in developing a sentiment classification model for public opinion regarding the Tabungan Perumahan Rakyat (Tapera) policy on social media X by combining the *Indonesian Sentiment Lexicon, Word2Vec*, and *Convolutional Neural Network* (CNN). The Indonesian Sentiment Lexicon was used to generate automatic sentiment labels, while *Word2Vec* built semantic representations of words that were used as the initial weights of the CNN embedding layer to support feature extraction. The main contribution of this study is the development of an integrated framework that combines automatic lexicon-based labeling, semantic word representation, and CNN-based classification for Indonesian public policy sentiment analysis.

The experimental results showed that variations in data sharing affected model performance. Among the four scenarios tested, the 80:20 configuration produced the best performance with an accuracy of 74.52%, surpassing the 90:10, 70:30, and 60:40 configurations. The comparison with *BiLSTM* and Random Forest also showed that CNN achieved better classification performance on the evaluated dataset. From an implementation perspective, policymakers can use social media sentiment monitoring as a complementary source of evidence to identify public concerns and improve communication regarding contribution mechanisms, fund management transparency, and the expected benefits of the Tapera program.

Despite these findings, the model still suffers from misclassification, particularly for neutral tweets with ambiguous or implicit meanings such as irony and sarcasm. These limitations are also influenced by *lexicon-based labeling*, which may not fully capture contextual meaning. Future research should explore context-aware language representations such as *IndoBERT*, *RoBERTa*, or other transformer-based architectures, as well as hybrid deep learning models and larger datasets. Combining automatic lexicon-based labeling with manual expert validation may also improve label quality. In addition, ablation experiments comparing CNN with and without *Word2Vec* are recommended to quantify the contribution of the embedding representation. These developments can strengthen the contribution of the proposed framework to Indonesian sentiment analysis by improving contextual understanding, label reliability, and model generalization.

Despite this, the model still suffers from misclassification, especially on tweets with neutral sentiments that have ambiguous language characteristics or contain implicit meanings, such as irony and sarcasm. These limitations are also influenced by the use of *lexicon-based labeling* that is not yet fully able to capture the context of the language in depth. Therefore, further research can explore more contextual language representations, such as *IndoBERT* or other transformer-based models, as well as combine automatic labeling with manual validation to improve label quality and classification accuracy.

REFERENCES

Asril, A. A., Rifai, A., & Shebubakar, A. N. (2022). Penyelenggaraan Tabungan Perumahan Menurut Undang-Undang Nomor 4 Tahun 2016 Ditinjau Dari Perspektif Perlindungan Hukum. *Jurnal Magister Ilmu Hukum*, *7*(1), 1. https://doi.org/10.36722/jmih.v7i1.1185

Atandoh, P., Zhang, F., Adu-gyamfi, D., Atandoh, P. H., & Elimeli, R. (2023). Integrated deep learning paradigm for document-based sentiment analysis. *Journal of King Saud University - Computer and Information Sciences*, *35*(7), 101578. https://doi.org/10.1016/j.jksuci.2023.101578

Çano, E., & Morisio, M. (2019). *Word Embeddings for Sentiment Analysis: A Comprehensive Empirical Survey*. 1–20.

Cheng, Y., Yao, L., Xiang, G., Zhang, G., Tang, T., & Zhong, L. (2020). Text Sentiment Orientation Analysis Based on Multi-Channel CNN and Bidirectional GRU With Attention Mechanism. *IEEE Access*, *8*, 134964–134975. https://doi.org/10.1109/ACCESS.2020.3005823

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Fahry, F., Utami, E., & Sudarmawan, S. (2023). OPTIMIZING SENTIMENT ANALYSIS OF PRODUCT REVIEWS ON MARKETPLACE USING A COMBINATION OF PREPROCESSING TECHNIQUES, WORD2VEC, AND CONVOLUTIONAL NEURAL NETWORK. *Jurnal Teknik Informatika (Jutif)*, *4*(1 SE-Articles), 101–107. https://doi.org/10.52436/1.jutif.2023.4.1.815

Fauzy, S. R. (2022). Prediksi Dan Analisis Jumlah Penduduk Desa Kecamatan Cidolog Tahun 2022 Menggunakan Artificial Neural Network Pada Aplikasi Matlab. *Naratif : Jurnal Nasional Riset, Aplikasi Dan Teknik Informatika*, *4*(2), 155–160. https://doi.org/10.53580/naratif.v4i2.163

Firdaus, M. P., & Trisnawarman, D. (2025). Analisis Sentimen Publik terhadap Program Tabungan Perumahan Rakyat Menggunakan Model IndoBERT Lite pada Komentar YouTube. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *5*(1), 359–368. https://doi.org/10.57152/malcom.v5i1.1744

Gandhi, U. D., Malarvizhi Kumar, P., Chandra Babu, G., & Karthick, G. (2021). Sentiment Analysis on Twitter Data by Using Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM). *Wireless Personal Communications*. https://doi.org/10.1007/s11277-021-08580-3

Gupta, V., Jain, N., Shubham, S., Madan, A., Chaudhary, A., & Xin, Q. (2021). Toward Integrated CNN-based Sentiment Analysis of Tweets for Scarce-resource Language—Hindi. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *20*(5), 1–23. https://doi.org/10.1145/3450447

Kurniasari, I., & Fatta, H. Al. (2021). Analisis Sentimen Opini Publik pada Instagram mengenai Covid-19 dengan SVM. *JTECS : Jurnal Sistem Telekomunikasi Elektronika Sistem Kontrol Power Sistem& Komputer*, *1*(1), 67–74.

Leidiyana, H., Misriati, T., & Aryanti, R. (2024). Klasifikasi Sentimen Terhadap Kebijakan Tapera Menggunakan Komparasi Machine Learning dan SMOTE. *Jurnal Komtika (Komputasi Dan Informatika)*, *8*(2), 125–135. https://doi.org/10.31603/komtika.v8i2.12595

Musthafa, A., Harmini, T., Rafiq, A., & Marantika, N. (2025). Pemanfaatan Machine Learning dalam Menganalisis Sentimen Terhadap Program TAPERA di Platform Digital X. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *5*(2), 587–597. https://doi.org/10.57152/malcom.v5i2.1801

Purwitasari, D., Devi, A., & Ariyanto, P. (2023). *Transformer-based Information Extraction from Twitter Text on Complaint Monitoring System*.

Sabitha, A. F. (2022). Analisis Pengaruh Tingkat Urbanisasi Terhadap Ketersediaan Lahan Lahan Permukiman Perumahan Di Kota Surabaya. *Jurnal Lemhannas RI*, *10*(1), 19–26. https://doi.org/10.55960/jlri.v10i1.268

Setiawan, B. (2024). A Review of Sentiment Analysis Applications in Indonesia Between 2023-2024. *JIEET: Journal Information Engineering and Educational Technology*, *08*, 71–83.

Sugiyono, S. (2019). Metodologi Penelitian Kualitatif Kuantitatif Dan R&D. *Bandung: Cv. Alfabeta*.

Taboada, M., Brooke, J., & Voll, K. (2011). *Lexicon-Based Methods for Sentiment Analysis*. *August 2010*.

Tang, H., Zhang, N., Yu, X., Mao, T., & Wang, L. (2023). Enhancing Sentiment Analysis with Word2Vec and LSTM : A Comparative Study. *Journal of Basic and Applied Research International*, *29*(3), 1–10. https://doi.org/10.56557/JOBARI/2023/v29i38342

Tania, N., Novienco, J., & Sanjaya, D. (2021). Kajian Teori Hukum Progresif Terhadap Implementasi Produk Tabungan Perumahan Rakyat. *Perspektif*, *26*(2), 73–87. https://doi.org/10.30742/perspektif.v26i2.800

Yin, R. K. (2017). *Case study research and applications: Design and methods*. Sage publications.