Jurnal Pendidikan dan Teknologi Indonesia (JPTI)
Vol. 5, No. 11, November 2025, Hal. 3368-3381

DOI: https://doi.org/10.52436/1.jpti.1182
p-ISSN: 2775-4227
e-ISSN: 2775-4219

Geospatial Sentiment Analysis on Government Flood Response in Jabodetabek Using a Hybrid Deep Learning Method

Argya Mauluvy Senjaya*[1], Yuliant Sibaroni[2]

[1,2]School of Computing, Informatics, Telkom University, Bandung, Jawa Barat, Indonesia
Email: [1]armaluvyyy@student.telkomuniversity.ac.id, [2]yuliant@telkomuniversity.ac.id

Abstract

Floods appear are common natural disasters that affect Indonesia, especially in the Jabodetabek area, and often trigger people to express their opinions and criticisms about the government's performance during flood events through social media platform such as Twitter . Among them, the handling of flood disasters in the Jabodetabek region is a highly discussed topic that causes widespread public reaction.This study aims to classify public sentiment toward government flood response using a hybrid deep learning model. A total of 3,894 Indonesian-language tweets were collected, preprocessed, and labeled. The sentiment classification process used IndoBERT as a semantic feature extractor and a CNN-LSTM architecture to capture both contextual and sequential patterns from the text. Evaluation was carried out using 10-fold cross-validation, with accuracy, precision, recall, and F1-score as performance metrics. IndoBERT achieved an accuracy of 91.76% and an F1-score of 90.66%, while the IndoBERT + CNN-LSTM model showed better performance with 94.92% accuracy and a 95.41% F1-score. To visualize the intensity of public discussions throughout Jabodetabek, geospatial mapping based on tweet location metadata was also carried out separately from sentiment classification using ArcGIS. The combination of semantic modeling and geospatial analysis offers an effective approach to understanding public sentiment in disaster-related contexts and supports better interpretation of regional public responses.

Keywords: *CNN-LSTM, geospatial mapping, IndoBERT, sentiment analysis, social media, twitter (X).*

1. INTRODUCTION

Floods appear as one of the most common and destructive natural disasters that affect Indonesia, especially in the Jabodetabek area [1]. These disasters cause major damage to infrastructure and the economy, and producing strong public reactions through social media platforms. Many people use Twitter as a real-time communication platform to express their opinions and criticisms about the government's performance during flood events [2].

The analysis of Twitter sentiment has proven to be a valuable tool to understand public opinion on different matters including disaster management [3]. The combination of sentiment analysis with geospatial data from tweets enables researchers to understand citizen reactions toward government actions throughout different areas [4]. The method provides a new approach to evaluate public opinion, which proves useful in densely populated urban environments.

Research has investigated sentiment classification within disaster-related contexts. The BERT-MLP model achieved 82% accuracy in classifying disaster-related tweets by combining semantic and spatial features without applying stemming or stopword removal [6]. The CNN-LSTM model achieved sentiment classification accuracy of 94% when analyzing monkeypox virus tweets [7]. IndoBERT demonstrates excellent performance in natural language processing tasks through its pre-trained transformer model, which reaches an F1-score of 95.6% [5]. However, research has not adequately examined the integration of IndoBERT with CNN-LSTM for analyzing disaster-related social media sentiment while accounting for public response spatial patterns.

Sentiment analysis is increasingly being applied not only in the fields of marketing and political research, but also in more critical areas such as disaster response and public communication. In emergency situations such as floods, governments require quick and reliable feedback from citizens to assess whether the measures taken are effective. In this case, sentiment analysis on social media has the potential to provide near real-time insights into public emotions and reactions to government actions [3], [4].

This study proposes a hybrid strategy to close the gap by combining a CNN-LSTM architecture for sentiment classification with IndoBERT as a feature extractor. CNN-LSTM processes both local and global information from each input token [8], while IndoBERT extracts deep semantic representations of Indonesian-language tweets [9]. Additionally, based on the location metadata included in the tweets, ArcGIS is used for geospatial mapping to visualize the sentiment distribution across cities in the Jabodetabek area [4].

p-ISSN: 2775-4227
e-ISSN: 2775-4219

However, although the number of studies in this field continues to grow, literature that specifically addresses the integration of sentiment analysis and geospatial visualization remains limited. In most previous research, sentiment and spatial data are handled separately and are rarely integrated within a single analytical framework [4]. Moreover, only a few studies focus on hybrid models that combine transformer-based feature extraction with sequence-based deep learning architectures, such as CNN and LSTM, to achieve higher classification robustness [7], [24], [26].

The model's performance is evaluated using 10-fold cross-validation and compared to the IndoBERT-only model. To observe spatial sentiment trends, geospatial visualization is applied to the classification results. Evaluation metrics used are accuracy, precision, recall, and F1-score. This study aims to use the IndoBERT + CNN-LSTM model to categorize public opinion regarding the government's handling of flood disasters in the Jabodetabek region.

2. RESEARCH METHOD

2.1. System Design

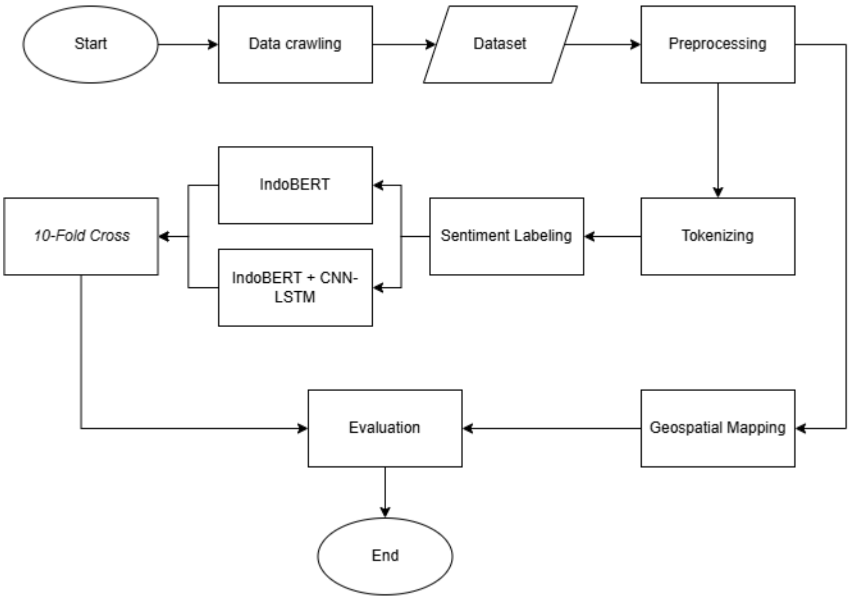

Figure 1. System flow.

The system built for sentiment classification and geospatial analysis, as shown in Figure 1, covers the entire process from data crawling to model evaluation and mapping. The process begins with data collection, which is followed by preprocessing and tokenization. The prepared data is then labeled for sentiment using two sentiment classification models: the IndoBERT model and the IndoBERT + CNN-LSTM hybrid architecture. The performance of the models is then evaluated using 10-fold cross-validation. Finally, geospatial visualization is performed using the preprocessed data before model evaluation.

2.2. Data Crawl

This study collected data from Twitter (X), focusing on user-generated tweets related to flood events in the Jabodetabek area. The crawling process was carried out using a custom Python-based scraper that leverages Playwright to automatically extract tweets from Twitter search results [10],[11]. Tweets were collected using a series of relevant keywords such as "banjir", "banjir" AND "Bekasi", "banjir" AND "pemerintah", and other terms related to flooding and government, within a specified date range based on the major flood disaster in Jabodetabek. In total, around 4,585 tweets were collected through this process and became 3,894 tweets after preprocessing the data that were stored for further analysis. Each tweet included important data such as full text, time stamps, and location-related information. These elements were then used for sentiment classification and geospatial mapping in the analysis.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

2.3. Dataset

This study uses a dataset of 3,894 Indonesian-language tweets that have been collected and preprocessed. The tweets are focused on topics related to floods and how the government responded to the incident in the Jabodetabek area. Tweets that are not related to the Jabodetabek flood, come from outside Jabodetabek, or are duplicates are removed during the cleaning process, to ensure data quality and relevance.

2.4. Preprocessing

The collected data will be preprocessed, the raw data will be transformed into a format that is easier to analyze. This stage will eliminate corrupted and inaccurate data from the available data set [12]. These are several stages in the preprocessing technique.

Case folding: Case Folding is used to standardize text in a dataset, in this study it is used to change all uppercase letters to lowercase letters [13].

Text cleaning: Text cleaning will remove tabs, new lines, back slices, emoticons, non-ASCII, mentions, hashtags, links, URLs, numbers, punctuation and whitespace [14].

Text Normalization: Text normalization used to ensure consistent text format by changing non-standard words or abbreviations into the correct form, using an Indonesian lexicon dictionary to improve the quality of the processed text data [15].

Although no explicit stopword removal, stemming, or lemmatization was applied, the cleaning steps substantially reduced noise in the dataset and facilitated accurate modeling.

2.5. Tokenizing

Tokenization in this study was performed using the *indobenchmark/indobert-base-p1* tokenizer from HuggingFace which converts cleaned and normalized text into a numeric representation suitable for the transformer model. Each sentence was segmented into subword tokens (input_ids) and given an attention_mask to distinguish the original tokens from the padding using the IndoBERT model [16],[17].

2.6. Sentiment Labeling

At this stage, two labeling steps are applied. Random data of 500 tweets is taken from the data set that has been cleaned in the previous stage, then they will be annotated manually with sentiment labels (0 for negative, 1 for neutral, 2 for positive). This manually labeled data will then be used to fine-tune the IndoBERT model for sentiment classification from the remaining unlabeled data. The fine-tuned IndoBERT model will be used to automatically label data that is still unlabeled.

The same experimental configuration was used for both the IndoBERT classification model and the IndoBERT + CNN-LSTM hybrid model. Training was performed using a batch size of 8, a learning rate of 2e-5, and run for 2 epochs. IndoBERT is optimized using the AdamW optimizer, while the Adam optimizer with a CrossEntropyLoss loss function is used to optimize the CNN-LSTM model. The parameters are chosen in line with typical practices of optimizing transformer-based models to the best matching hardware resources for a local run. All experiments were run locally on a personal computer using Python 3.11.1 in Visual Studio Code, with GPU acceleration from NVIDIA GTX 1650 and 16 GB RAM.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

2.7. IndoBERT

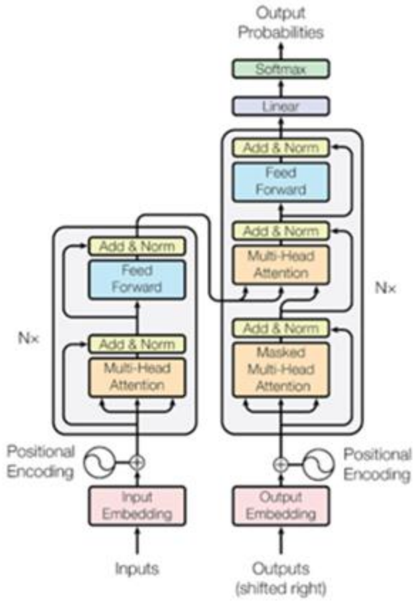

Figure 2. IndoBERT model architecture [27].

IndoBERT is a BERT architecture-based model designed explicitly for Indonesians. The architecture of the IndoBERT model is the same as the BERT model, which consists of 12 encoder layers, where each encoder layer consists of a multi-head self-attention sub-layer and a fully connected sub-layer [18],[19].

In this study, the tweet data collected were Indonesian-language tweets, so the IndoBERT model is the best choice. The advantage of IndoBERT is the ability to detect semantic meaning of word in Indonesian tweet [9].

2.8. IndoBERT + CNN-LSTM

To enhance the effectiveness of sentiment classification, this study utilizes a hybrid architecture that combines IndoBERT as a feature extractor with a CNN-LSTM model [26]. IndoBERT will produce contextual word embeddings that capture the semantic meaning of Indonesian tweets [9].

CNN-LSTM utilies CNN layers as an initial stages to extracts local features and patterns from the text, while also reducing noise. The output is then forwarded to the LSTM layer, which captures temporal dependencies and sequential structures of sentences. CNN-LSTM has been shown to be effective in sentiment analysis tasks on online media platforms [24], [25].

2.9. 10-Fold Cross Valisation

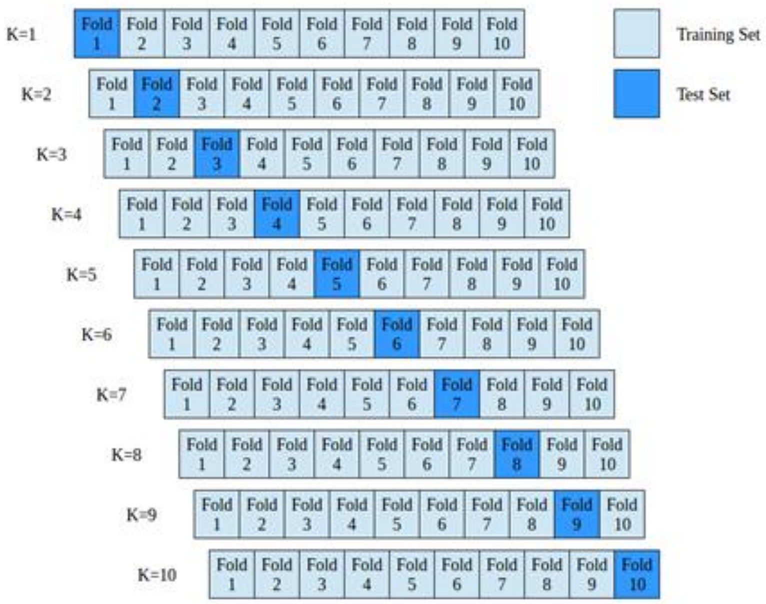

Figure 3. 10-Fold Cross Validation method diagram [21].

10 Fold Cross-Validation is a nasted operator where conceptually this validation divides the training data into 10 equal parts and then learning will be carried out 10 times. Each time another part of the dataset is selected for testing and the remaining data is used for learning. Then the average value of 10 different test results is calculated [20]. The stages of this validation model are presented in the form of images, as shown in Figure 3.

2.10. Geospatial Mapping

Geospatial mapping is a wide definition of all the operations involved in the process of creating maps that use geospatial data. Its primary aim is to show items with geographic coordinates in a geographical framework, providing a representation of the physical world on a map [22], [23].

In this study, geospatial mapping is used to map the distribution of tweets based on the location of the tweets. Geomapping uses the ArcGIS application to perform spatial analysis.

2.11. Evaluation

At this stage, two evaluations will be carried out, evaluation of the cross-validation results of each method, and evaluation of the geospatial distribution.

The performance results of the two sentiment classification models used, IndoBERT and IndoBERT + CNN-LSTM will be evaluated based on four metrics: accuracy, precision, recall, and F1-score, which are calculated from the confusion matrix to determine the overall effectiveness of each approach. Accuracy measures the correct predictions made by the model out of the total predictions as described in Equation 1. Precision evaluates how many of the cases predicted as positive are actually positive in Equation 2. Recall, shown in Equation 3 aims to measure the model's ability to correctly identify all cases that are actually positive. The F1 score combines precision and recall into a single metric to provide a balanced assessment of the classification performance of the model in Equation 4. The following is the calculation of the evaluation value [28].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (3)$$

$$F1 - Score = \frac{2 \times Precision - Recall}{Precision + Recall} \times 100\% \quad (4)$$

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Then an evaluation will be carried out on geospatial mapping using ArcGIS. This stage analyzes the spatial distribution of tweets related to flooding in the Jabodetabek area. The mapping is based on unlabeled raw tweet data and aims to observe tweet density and regional discussion patterns before sentiment classification is applied.

3. RESULT AND DISCUSSION

3.1. Dataset

The dataset consists of 3,894 Indonesian tweets with *conversation_id, created_at, full_text, id_str, lang, location, tweet_url*, and *username*. However, only two columns will be focused on in this study, namely the *full_text* column, because it contains tweets that will be used for sentiment classification, and the *location* column, because it contains the location of the tweets to be used in geospatial. an example of the dataset will be shown in the Table 1 below.

Table 1. Sample dataset.

Full_text	Location
@txtdrbekasi Warga : pemerintah kerja gag becus Pemerintah : banjir karena intensitas hujan yang tinggi Padahal salah 2-2nya gag mau belajar dr kejadian masa lalu dan tidak bisa introspeksi diri.	Bekasi
Pandangan saya tentang banjir Jakarta Banjir di Jakarta bukan semata-mata salah pemerintah coba instrospeksi lagi bagaimana cara kalian masyarakat ibu kota dan sekitarnya berlaku kepada alam sekitar anda tinggal bagaimana cara kalian membuang sampah disekitar tempat anda	DKI Jakarta

3.2. Pre processing

This stage is done to reduce noise and improve model performance. Several preprocessing steps were applied to prepare the data before it was ready to be used for this study. Examples of the steps involved in data preprocessing are presented in Table 2.

Table 2. Preprocessing steps.

Process	Text
Original text	@voicexist Btw maaf tolong jgn menyalahkan pemerintah nya.. bahkan walkot tangerang kota pak arief udah turun ke jalan. Bahkan ngurus Ciledug Indah 1. Sy warga komp Pinang Griya juga banjir segenteng juga sama. Emanf listris pasti di padamkan semua. Act & SAR jg berusaha semaksimal mungkin
Case folding	@voicexist btw maaf tolong jgn menyalahkan pemerintah nya.. bahkan walkot tangerang kota pak arief udah turun ke jalan. bahkan ngurus ciledug indah 1. sy warga komp pinang griya juga banjir segenteng juga sama. emanf listris pasti di padamkan semua. act & sar jg berusaha semaksimal mungkin
Cleanned text	btw maaf tolong jgn menyalahkan pemerintah nya bahkan walkot tangerang kota pak arief udah turun ke jalan bahkan ngurus ciledug indah sy warga komp pinang griya juga banjir segenteng juga sama emanf listris pasti di padamkan semua act amp sar jg berusaha semaksimal mungkin

 p-ISSN: 2775-4227
e-ISSN: 2775-4219

Normalized text	btw maaf tolong jangan menyalahkan pemerintah nya bahkan walkot tangerang kota pak arief sudah turun ke jalan bahkan mengurus ciledug indah saya warga komp pinang griya juga banjir segenteng juga sama emanf listris pasti di padamkan semua act amp sar juga berusaha semaksimal mungkin

3.3. Tokenizing

Tokenization is performed using the IndoBERT tokenizer from the Hugging Face Transformers library to prepare the input model [29]. In this step, each sentence is then converted into two main components: input_ids and attention_mask to ensure the model processes it effectively [30].

Input_ids represent the numerical encoding of each token, based on the IndoBERT vocabulary [31]. Based on research [30], the attention_mask indicates which tokens are actual input (1) and which are padding tokens (0), ensuring that the model only processes relevant tokens and ignores padding. When the sentence length is shorter than the maximum sequence length, a padding token (0) is added at the end of the input_ids, with a corresponding 0 in the attention_mask.

The input length will be limited to a maximum length of 128 tokens to ensure efficient batch processing. this restriction is done to maintain computational efficiency during model training as most tweets have short content. An example of the tokenization result is shown in Table 3.

Table 3. Tokenizing result.

Preprocessed text	Input_ids	Attention_mask
yang selanjutnya berkoordinasi dan bekerjasama dengan serta pemerintah setempat dalam memulihkan kondisi wilayah yang terkena dampak banjir di jakarta utara	[2, 34, 1641, 17694, 41, 8255, 79, 501, 877, 4111, 112, 14802, 1186, 1350, 34, 3841, 3345, 5138, 26, 678, 2024, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

3.4. Sentiment Labeling

The data that has been collected will be labeled first before entering the classification stage, this labeling aims to facilitate the data classification process [32].

Labeling will be done in 2 stages, the first will be manual labeling on 500 random data taken from the dataset. These 500 random data will then be manually annotated with sentiment labels consisting of three classes: 0 for negative, 1 for neutral, and 2 for positive. This labeling is done manually based on the emotional tone and intent of the text content. Here are some examples of manual data labeling on 500 random data in Table 4.

Table 4. The example of manual sentiment labeling.

Cleaned text	Label
kok kecewa sama pak pak bukankah pak sdh menyatakan soal banjir jakarta itu adalah tanggung jawab pemerintah pusat pak basuki jgn ikut on juga lah	0
bogor hujan gede hampir setiap hari banjir longsor banyak kerjaan diluar rumah jadi takut kemana mana	1
grak cepat pemprov dki antisipasi potensi gugatan class action banjir pemerintah provinsi dki jakarta telah mmbahas antisipasi adanya potensi pengajuan gugatan perwakilan kelompok class action lawsuit terkait banjir pada januari oleh masyarakat	2

p-ISSN: 2775-4227
e-ISSN: 2775-4219

After the manual labeling process was complete, the 500 labeled tweets were recombined into the main dataset. The manually labeled data was then used to train the sentiment classification model.

The IndoBERT model was then fine-tuned using the 500 manually labeled tweets. The Hugging Face Trainer API was used to fine-tune the IndoBERT model [33]. The model was trained for 2 epochs with an 80:20 split between the train and validation with a batch size of 4 per device. Checkpoints were saved every 500 steps, and training logs were recorded at the same intervals to monitor progress. The model output was saved to the indobert-finetuned directory.

The remaining unlabeled data will be automatically labeled using the customized IndoBERT model. This automatic labeling procedure is executed incrementally using a custom DataLoader where each tweet is tagged and passed through the model to generate sentiment predictions. As a result, the entire dataset of 3,894 tweets was successfully labeled, the distribution of sentiment labels across the dataset can be seen in Figure 4 and is ready to be used in the next part of this research.

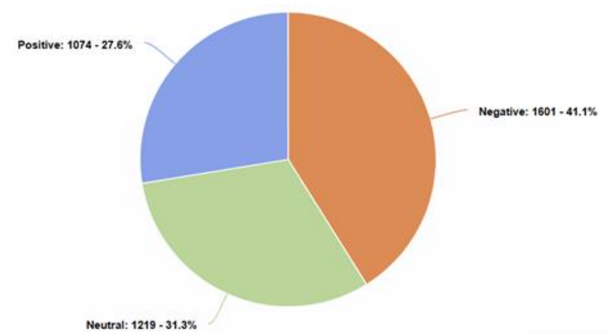

Figure 4. Sentiment distribution.

3.5. Sentiment Classification

In this stage, the sentiment classification process is applied to the labeled dataset. Two different models are used to perform sentiment classification: IndoBERT and a hybrid model that combines IndoBERT with CNN-LSTM. The results will be calculated using four standard metrics (accuracy, precision, recall, and F1-score), each of which is calculated based on the confusion matrix [34].

p-ISSN: 2775-4227
e-ISSN: 2775-4219

3.5.1. IndoBERT Method

At this stage, the classification process using the fine-tuned IndoBERT model was trained for 2 epochs with a batch size of 8 and a learning rate of 2e-5. At each epoch, 90% of the data was used for training and 10% for validation. ScikitLearn's classification_report() function was used to evaluate the performance of accuracy, precision, recall, and F1-score metrics [35].

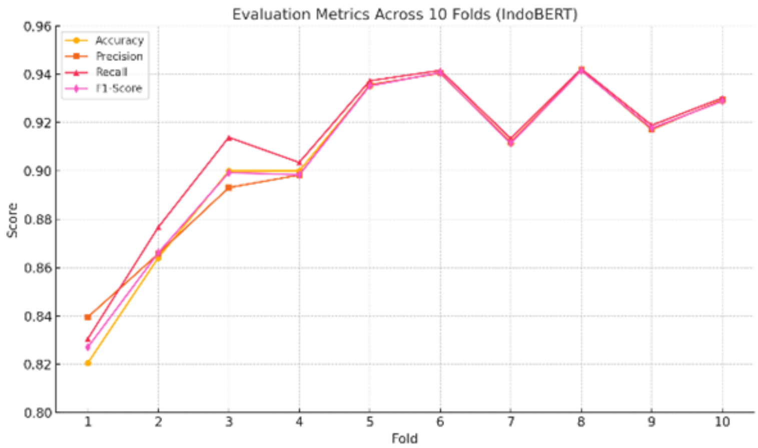

Figure 5. Evaluation metrics across 10-Folds (IndoBERT).

The performance increased consistently from Fold 1 to Fold 6 which can be seen in Figure 5. Fold 6 displayed the best performance with accuracy and F1-score reaching 94.07%. While the lowest performance is seen in Fold 1, with an accuracy of 82.05% and F1-score of 82.70%. The better improvement in early folds can possibly imply the fact that the model performed better with the training data used to train it, making those early folds even better. Performance became more stable on all folds after Fold 5, implying that the model became better at generalizing and its performance was also stable.

Table 5. Average result of IndoBERT model performance.

Metric	Average Score
Accuracy	0.9056
Precision	0.9062
Recall	0.9073
F1-Score	0.9066

Table 5 shows the average results of IndoBERT model performance from all folds. The averages of all metrics are: 90.56% accuracy, 90.62% precision, 90.73% recall, and 90.66% F1-score. These results indicate that IndoBERT is reliable and effective for sentiment classification on IDndonesian tweets.

3.5.2. IndoBERT + CNN-LSTM Method

This stage combines the IndoBERT method with the CNN-LSTM method to improve sentiment classification performance. This method uses the IndoBERT model as a feature extractor to generate contextual embeddings of each token in the tweet. The resulting embedding is then passed through a convolutional layer (CNN) to capture local text patterns, followed by an LSTM layer to extract temporal and sequential dependencies between tokens.

The model was cross-validated 10 times, and trained with 2 epochs, batch size 8, and learning rate 2e-5 at each fold. 6 Performance metrics such as accuracy, precision, recall, and f1-score were evaluated and reported with classification_report() from the Scikit-Learn library.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

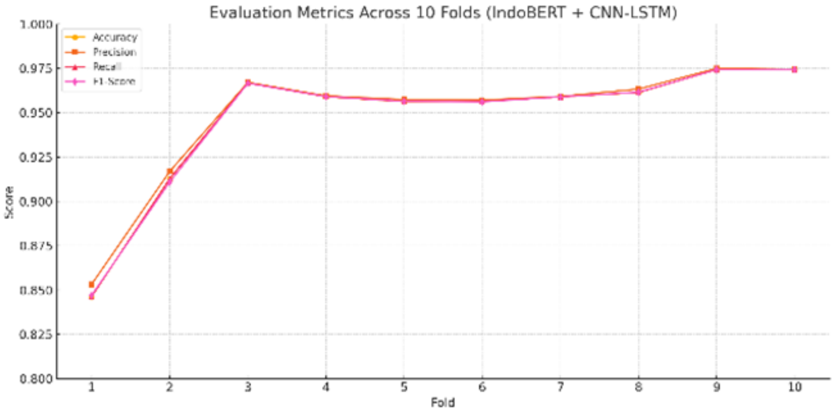

Figure 1. Evaluation metrics across 10-Folds (IndoBERT+CNN-LSTM).

The model performance shows relatively stable and consistent performance as shown in Figure 6 where the maximum value is in Fold 9 and Fold 10 with an accuracy and F1-score of 97.43%. Fold 1 has the worst value, with an accuracy of 84.62% and F1-score of 84.70%. The small difference between folds indicate that the model provided a decent partition of the data, and might be able to generalize better. Given the similarity of results between the majority of the folds it seems likely that the hybrid classification approach provided the model with a more stable and robust set of features for classification.

Table 6. Average result of IndoBERT+CNN-LSTM model performance.

Metric	Average Score
Accuracy	0.9492
Precision	0.9537
Recall	0.9553
F1-Score	0.9541

The average performance was 94.92% accuracy, 95.37% precision, 95.53% recall and 95.41% F1-score as presented in Table 6. These findings show that the combination of CNN and LSTM is effective for modeling local and sequential information in sentiment classification.

3.6. Geospatial MappingD

Geospatial mapping was conducted to see the spatial distribution of tweets related to responses to government performance on flooding in the Greater Jakarta area. The location of the tweets was extracted from the geographic coordinates (latitude and longitude) available in the metadata of each tweet. Visualization was done using the ArcGIS platform, which displays the number of tweets per city/district in the form of proportionally sized dotted symbols.

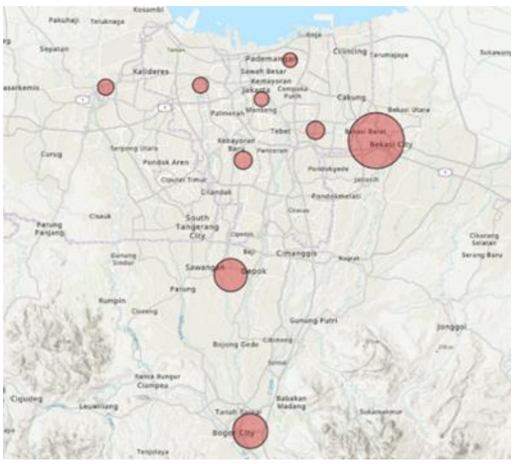

Figure 2. Geospatial distribution map.

The spatial distribution representing the entire dataset used in the training and evaluation of the model is illustrated in Figure 7. This mapping was done prior to sentiment classification, to provide an initial overview of the concentration of public discussion on responses to government performance in dealing with floods across regions. High tweet activity was observed in areas such as Bekasi, Depok, and East Jakarta, which are historically known as flood-prone areas. The clustering of tweets in these cities may not only indicate higher flood exposure but also more active public response and reporting behavior from residents in those areas. This distribution provides valuable insights into which areas received the most public attention during the flood event.

3.7. Evaluation

A comparative evaluation was conducted to measure the performance of two sentiment classification models: the fine-tuned IndoBERT model and the hybrid IndoBERT + CNN-LSTM architecture. Both models were tested using 10-fold cross-validation, and their performance was measured using standard evaluation metrics, which calculate accuracy, precision, recall, and F1-score. This evaluation aims to determine which model is more effective and consistent in classifying public sentiment in Indonesian tweets relating to the government's response to the flood disaster in Jabodetabek.

The IndoBERT model achieved good results with an average accuracy of 91.76% and F1-score of 90.66%. While the model showed effective classification overall, its performance varied more across folds, with Fold 1 resulting in the lowest score, while Fold 6 resulted in the highest score. The model maintained relatively balanced precision and recall values across all classes. As a transformer-based model customized specifically for Indonesian, IndoBERT demonstrates that contextual embedding alone is sufficient for sentiment analysis.

On the other hand, IndoBERT + CNN-LSTM has better performance than IndoBERT based on all indicators. It uses IndoBERT as a frozen feature extractor, followed by a CNN layer to capture local patterns and an LSTM layer to capture dependencies between tokens. The result is a more robust and consistent classifier, with an average accuracy of 94.92% and an F1 value of 95.41%. The strongest performance is in Fold 9 and Fold 10 where both metrics have a value of 97.43%, indicating strong generalization and low fold variance.

Figure 3. Model performance comparison.

Table 7. Comparison of average evaluation metrics between models.

Model	Accuracy	Precision	Recall	F1-Score
IndoBERT	91.76%	91.26%	91.39%	90.66%
IndoBERT+CNN-LSTM	94.92%	95.37%	95.53%	95.41%

The performance comparison of the two models is shown in Table 7 and Figure 8. An improvement was achieved by IndoBERT + CNN-LSTM hybrid model with a +3.16% increase in accuracy and a +4.75% gain in F1-score compared to the IndoBERT model. With clear differences in accuracy, precision, recall and F1-score, it shows that the IndoBERT + CNN-LSTM model outperforms the IndoBERT model on all metrics. This progress demonstrates the advantages of using semantic, spatial and sequential modeling in sentiment classification on social media tweets in Indonesian.

Based on the performance analysis results, IndoBERT + CNN-LSTM is the best model to perform sentiment classification in this study. Although the geospatial mapping presented uses raw tweet locations without sentiment

p-ISSN: 2775-4227
e-ISSN: 2775-4219

classification, the results of this experiment provide a strong foundation for selecting the most appropriate model when performing sentiment-aware spatial analysis in the future.

3.8. Comparison of Classification Models

Table 8. Comparison of accuracy between previous study.

Model	Accuracy
BERT + MLP	82%
CNN-LSTM	94%
IndoBERT	91.76%
IndoBERT + CNN-LSTM	94.92%

As shown in Table 8. BERT is combined with MLP to improve sentiment classification performance. The model was trained to classify disaster-related tweets using a public dataset from Twitter which showed an accuracy of 82% [6].

There are other studies that try different method to improve accuracy in classification tasks, one of these studies used Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM). By classifying sentiment using tweets related to monkeypox, this method achieved an accuracy rate of 94% [7].

In this study being discussed, IndoBERT is being fine-tuned and use to classifying sentiment using tweets related to public response regarding the government’s handling of flood disasters in Jabodetabek. This model showed an accuracy of 91%.

To improve IndoBERT’s performance, a hybrid model combining IndoBERT with CNN-LSTM was performed for sentiment classification. This combination achieved a higher accuracy score of 94.92%, which is higher than other method.

From the comparison of previous research, all the various deep learning method such as BERT-MLP, CNN-LSTM, IndoBERT, and IndoBERT + CNN-LSTM shows different strengths and varying level of accuacy, reflecting the diverse effectiveness of sentiment classification approaches. Each method has different advantages and accuracy results, which shows that there are various effective approaches to classify sentiment on social media especially Twitter (X). This comparison also shows that using the right combination of models and techniques can significantly improve the accuracy of sentiment classification.

3.9. Discussion

This research presents several contributions that distinguish it from previous studies. Although previous studies using BERT-MLP and CNN-LSTM methods showed good performance in sentiment classification tasks, most of them used transformation-based or sequence-based models separately. In this study, IndoBERT is used as a feature extractor and combined with CNN-LSTM, which allows the model to utilize contextual understanding and sequential learning. This combination helped to improve classification performance and maintain stable results across all folds in 10-fold cross-validation.

Another important aspect of this research is the integration with spatial analysis. Although the spatial mapping has not been based on sentiment classification, the visualization of tweet distribution across the region provides a starting point for understanding the public focus in flood-prone areas. This approach has rarely been explored in similar research and has the potential to support more specific policy evaluations in the future.

To improve sentiment classification, it is recommended to integrate sentiment classification results into geospatial mapping, so that the distribution of sentiment across regions can be visualized more clearly. Explore other models based on transformers or ensemble methods to improve classification accuracy and performance.

4. CONCLUSION

This research uses sentiment classification approaches such as IndoBERT and a combination of IndoBERT + CNN-LSTM on Indonesian tweets related to public opinion on the government's response to flood disaster in Jabodetabek. The dataset used is indonesian-language tweets that have been collected and preprocessed, totaling 3984 tweets. Data preprocessing includes case folding, data cleaning, normalization and tokenizing. Then confusion matrix is used to measure the performance of the model. The tuned IndoBERT model shows an average accuracy of 91.76%, precision of 91.26%, recall of 91.39%, and F1-score of 90.66%. While the IndoBERT + CNN-LSTM model shows a significant increase in results, with the confusion matrix this model shows an accuracy

p-ISSN: 2775-4227
e-ISSN: 2775-4219

of 94.92%, precision of 95.37%, recall of 95.53%, and F1-Score of 95.41%. By combining semantic and sequential modeling, the performance of sentiment classification can be improved.

It is recommended that future research look into transformer-based architectures or other ensemble methods that can give even better classification results than the ones that exist now. Adding sentiment classification results to geospatial mapping will also help to look at the spatial pattern of public opinion in more detail. This will make it easier for the government to respond to what people respond based on how their sentiment.

REFERENCES

[1] Dino, "Banjir: Pengertian, Penyebab, dan Dampaknya," *BPBD Provinsi Jawa Timur*, Oct. 19, 2023. https://web.bpbd.jatimprov.go.id/2023/10/19/banjir-pengertian-penyebab-dan-dampaknya/.

[2] M. W. Widjaya, "Penyalahgunaan Kekuatan Sosial Media dalam Membentuk Opini Publik Halaman 1 - Kompasiana.com," *KOMPASIANA*, Jan. 10, 2024. [Online]. Available: https://www.kompasiana.com/malikwisnuwidjaya/659e6331c57afb31676c9a52/penyalahgunaan-kekuatan-sosial-media-dalam-membentuk-opini-publik/.

[3] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P.-M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," *Expert Systems With Applications*, vol. 223, p. 119862, Mar. 2023, doi: 10.1016/j.eswa.2023.119862.

[4] M. A. Nulhakim, Y. Sibaroni, and K. M. N. K. Khalif, "Geospatial Sentiment Analysis Using Twitter Data on Natural Disasters in Indonesia with Support Vector Machine (SVM) Algorithm," *socjs.telkomuniversity.ac.id*, 2024, doi: 10.21108/ijoict.v10i2.1032.

[5] M. F. Mubaraq and W. Maharani, "Sentiment Analysis on Twitter Social Media towards Climate Change on Indonesia Using IndoBERT Model," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 4, p. 2426, Oct. 2022, doi: 10.30865/mib.v6i4.4570.

[6] I. Maulana and W. Maharani, "Disaster Tweet Classification Based On Geospatial Data Using the BERT-MLP Method," *2021 9th International Conference on Information and Communication Technology (ICoICT)*, Yogyakarta, Indonesia, 2021, pp. 76-81, doi: 10.1109/ICoICT52021.2021.9527513.

[7] K. K. Mohbey, G. Meena, S. Kumar, and K. Lokesh, "A CNN-LSTM-Based hybrid deep learning approach for sentiment analysis on Monkeypox tweets," *New Generation Computing*, vol. 42, no. 1, pp. 89–107, Aug. 2023, doi: 10.1007/s00354-023-00227-0.

[8] M. Cho, J. Ha, C. Park, and S. Park, "Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition," *Journal of Biomedical Informatics*, vol. 103, p. 103381, Jan. 2020, doi: 10.1016/j.jbi.2020.103381.

[9] A. A. Hafiza and E. B. Setiawan, "Enhancing cyberbullying detection on platform 'X' using IndoBERT and hybrid CNN-LSTM model," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 2, pp. 655–672, Apr. 2025, doi: 10.52436/1.jutif.2025.6.2.4321.

[10] B. Ališauskas, "How to Scrape X.com (Twitter) using Python (2025 Update)," *ScrapFly Blog*, Jan. 21, 2025. https://scrapfly.io/blog/how-to-scrape-twitter/.

[11] H. Satria, "Cara mendapatkan data (Crawl) Twitter X - Maret 2024," *Helmi Satria Writings: Stories and Learnings*, Jul. 02, 2024. https://helmisatria.com/blog/updated-crawl-data-twitter-x-maret-2024/.

[12] A. F. Sidabutar, R. Habibi, and W. I. Rahayu, "PERBANDINGAN METODE KLASIFIKASI UNTUK PENGELOMPOKAN RISIKO MAGANG MAHASISWA," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 3, pp. 2071–2076, Dec. 2023, doi: 10.36040/jati.v7i3.7026\.

[13] H. Imaduddin, L. A. Kusumaningtias, and F. Y. A'la, "Application of LSTM and GLOVE word embedding for hate speech detection in Indonesian Twitter data," *Ingénierie Des Systèmes D Information*, vol. 28, no. 4, pp. 1107–1112, Aug. 2023, doi: 10.18280/isi.280430..

[14] A. L. S. A.-Z. Gunawan, J. Jondri, and K. M. Lhaksamana, "Analisis sentimen pada media sosial Twitter terhadap penanganan bencana banjir di Jawa Barat dengan metode jaringan saraf tiruan," *eProceedings of Engineering,* no. Vol. 8 No. 2, Apr. 2021, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/14695/.

[15] M. R. Haziq, Y. Sibaroni, and S. S. Prasetyowati, "WORD EMBEDDING OPTIMIZATION IN SENTIMENT ANALYSIS OF REVIEWS ON MYTELKOMSEL APP USING LONG SHORT-TERM MEMORY AND SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 6, pp. 1581–1589, Dec. 2024, doi: 10.52436/1.jutif.2024.5.6.2498.

[16] C. Toraman, E. H. Yilmaz, F. ŞahiNuç, and O. Ozcelik, "Impact of tokenization on language Models: An analysis for Turkish," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 4, pp. 1–21, Mar. 2023, doi: 10.1145/3578707\.

[17] "indobenchmark/indobert-base-p1 · Hugging Face," *Hugging Face*.

https://huggingface.co/indobenchmark/indobert-base-p1/.
[18] K. Kusnawi and K. Anam, “Comparison ff Sentiment Labeling Using Textblob, Vader, and Flair in Public Opinion Analysis Post-2024 Presidential Inauguration with IndoBERT,” *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 2, pp. 803–818, Apr. 2025, doi: 10.52436/1.jutif.2025.6.2.4015.
[19] E. Yulianti and N. K. Nissa, “ABSA of Indonesian customer reviews using IndoBERT: single- sentence and sentence-pair classification approaches,” *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 5, pp. 3579–3589, Jul. 2024, doi: 10.11591/eei.v13i5.8032.
[20] I. Tri Julianto, D. Kurniadi, M. Rikza Nashrulloh, and A. Mulyani, “Comparison of data mining Algorithm for forecasting Bitcoin Crypto currency trends,” *Jurnal Teknik Informatika (Jutif)*, vol. Vol. 3 No. 3, Jun. 2022, [Online]. Available: https://jutif.if.unsoed.ac.id/index.php/jurnal/article/view/343.
[21] J. J. Andrade, L. G. Da Fonseca, M. Farage, and G. L. De Oliveira Marques, “PREDICTION OF THE PERFORMANCE OF BITUMINOUS MIXES USING ADAPTIVE NEURO-FUZZY INFERENCE SYSTEMS,” *Revista Mundi Engenharia Tecnologia E Gestão* (ISSN 2525-4782), vol. 5, no. 6, Aug. 2020, doi: 10.21575/25254782rmetg2020vol5n61367.
[22] M. Jedrzejowska, “What is geospatial mapping, and how does it work?,” *Spyrosoft*, Oct. 15, 2024. https://spyro-soft.com/blog/geospatial/what-is-geospatial-mapping-and-how-does-it-work/.
[23] I. Mwanja, “Research Guides: An Introduction to Geospatial Mapping: Understanding Geospatial Mapping,” *Virginia Tech*, May 16, 2025. https://guides.lib.vt.edu/an-introduction-to-geospatial-mapping/.
[24] I. E. Livieris, E. Pintelas, and P. Pintelas, “A CNN–LSTM model for gold price time-series forecasting,” *Neural Computing and Applications*, vol. 32, no. 23, pp. 17351–17360, Apr. 2020, doi: 10.1007/s00521-020-04867-x.
[25] D. T. Hermanto, A. Setyanto, and E. T. Luthfi, “Algoritma LSTM-CNN untuk Binary Klasifikasi dengan Word2vec pada Media Online,” *Creative Information Technology Journal*, vol. 8, no. 1, p. 64, Mar. 2021, doi: 10.24076/citec.2021v8i1.264.
[26] W. Anggraeni, M. F. Arrizal Kusuma, E. Riksakomara, R. P. Wibowo, Pujiadi, and S. Sumpeno, “Combination of BERT and hybrid CNN-LSTM models for Indonesia Dengue tweets classification,” *International Journal of Intelligent Engineering and Systems*, vol. 17, no. 1, pp. 813–826, Dec. 2023, doi: 10.22266/ijies2024.0229.68.
[27] B. Juarto and Yulianto, “Indonesian news classification using IndoBert,” *International Journal of Intelligent Systems and Applications in Engineering (IJISAE)*, vol. Vol 11 No.2, Feb. 2023, [Online]. Available: https://ijisae.org/index.php/IJISAE/article/view/2654/.
[28] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, “Evaluating trust prediction and confusion matrix measures for web services ranking,” *IEEE Access*, vol. 8, pp. 90847–90861, Jan. 2020, doi: 10.1109/access.2020.2994222.
[29] “Tokenizer,” *Hugging Face*. https://huggingface.co/docs/transformers/main_classes/tokenizer/.
[30] A. Musa, F. M. Adam, U. Ibrahim, and A. Y. Zandam, “HauBERT: A Transformer Model for Aspect-Based Sentiment Analysis of Hausa-Language Movie Reviews,” *MDPI*, p. 43, Apr. 2025, doi: 10.3390/engproc2025087043.
[31] G. Afham Asnawi, “Perbandingan Model IndoBERT dan Model Hybrid pada Analisis Sentimen Opini Masyarakat Terhadap Judi Online,” *UIN Ar-Raniry Banda Aceh*, 2025, [Online]. Available: https://repository.ar-raniry.ac.id/id/eprint/41795.
[32] A. Zhdanovskaya, D. Baidakova, and D. Ustalov, “Data labeling for Machine learning Engineers: Project-Based curriculum and Data-Centric Competitions,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 13, pp. 15886–15893, Jun. 2023, doi: 10.1609/aaai.v37i13.26886.
[33] “Fine-tuning a model with the Trainer API - Hugging Face LLM Course,” *Hugging Face*. https://huggingface.co/learn/llm-course/chapter3/3.
[34] E. Helmud, E. Helmud, F. Fitriyani, and P. Romadiana, “Classification comparison performance of supervised machine learning random forest and decision tree algorithms using confusion matrix,” *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, vol. 13, no. 1, pp. 92–97, Feb. 2024, doi: 10.32736/sisfokom.v13i1.1985.
[35] “classification_report,” *Scikit-learn.* https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html.