Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6, 169-183
2024
Publisher: Learning Gate
DOI: 10.55214/25768484.v8i6.2033

Behavioural Indonesian disaster data classification in social media using KNN, random forest, and RNN in machine learning

Tito Waluyo Purboyo[1]*, Rifki Wijaya[2], Roswan Latuconsina[3], Casi Setianingsih[3], Faris Ruriawan[3]
[1]Biomedical Engineering, Telkom University, Bandung, Indonesia; titowaluyo@telkomuniversity.ac.id (T.W.P.)
[2]Informatics Engineering, Telkom University, Bandung, Indonesia.
[3]Computer Engineering, Telkom University, Bandung, Indonesia.

Abstract: This research focuses on the classification of Indonesian disaster data extracted from social media using three distinct machine learning models: K-Nearest Neighbors (KNN), Random Forest, and Recurrent Neural Network (RNN). The study evaluates and compares the performance of these models based on accuracy metrics. The KNN model demonstrates competency with an 82% accuracy, showcasing its ability to handle data based on nearest neighbors. However, there are indications of potential limitations in handling complex patterns or high-dimensional datasets. In contrast, Random Forest achieves a notable 94% accuracy, highlighting its effectiveness in combining decision trees to enhance performance and mitigate overfitting. The RNN model exhibits the highest performance, achieving 96% accuracy, attributed to its proficiency in understanding temporal relationships and sequential patterns in data, making it a robust choice for sequential datasets such as text or time series. The findings underscore the importance of selecting models tailored to specific dataset characteristics and desired classification analyses. Furthermore, future research directions involve exploring adaptations or optimizations of these models for diverse disaster types and social media data, aiming to develop more sophisticated and responsive models for disaster-related analysis in the Indonesian context.

Keywords: *Disaster data, Disaster response, Model architecture, Social media behaviour.*

1. Introduction

In the face of the complexity of disaster-related challenges, the role of social media platforms, such as Twitter/X, becomes crucial as a communication channel during crisis situations [1]. The digital age witnessed an unprecedented influx of information during the crisis, with more than 8.7% of Indonesia's total population, approximately 24 million, actively using such platforms by early 2023. This rapid growth, totaling 5.6 million users between 2022 and early 2023, reflects the significance of social media platforms in providing communication channels during disasters [2].

While social media platforms provide great benefits, this growth brings new challenges, particularly with regard to information overload and huge volumes of data [3]. The management and classification of disaster information requires a careful approach, focusing on identifying and filtering relevant information to provide a clearer and more accurate view of the disaster situation [4].

Information spread on social media during disasters covers vital needs such as casualties, food, water, electricity, medical emergencies, shelter, evacuation routes and clothing [5]. However, the signal-to-noise ratio poses a severe challenge, where significant noise on social media hinders the extraction of actionable insights for informed decision-making. This challenge is faced not only by government agencies but also by non-governmental organizations (NGOs) and individuals invested in disaster response efforts [6].

* Correspondence: titowaluyo@telkomuniversity.ac.id

To address these challenges, this research introduces models that use Machine Learning algorithms such as K-Nearest Neighbors (KNN), Random Forest, and Recurrent Neural Network (RNN). The model is designed to cope with information overload and huge volumes of data, with the aim of improving situational awareness and facilitating faster and more informed decision-making from relevant parties. The model is expected to empower government agencies, NGOs, and individuals with the ability to quickly access and act on critical information, contributing to a more effective and responsive disaster management strategy [7].

This research also includes a comparison of the accuracy of the three machine learning algorithms used (KNN, Random Forest, and RNN), with the aim of determining the best algorithm for categorizing disaster data such as water, food, medical emergencies, shelter, and electricity. Thus, this research is expected to make a valuable contribution in optimizing disaster information management through social media to support more effective countermeasures.

In addition, there are several previous studies that have conducted research related to the classification of tweets related to natural disasters on Twitter, such as in paper Classification of Tweets Related to Natural Disasters Using Machine Learning Algorithms. This research focuses on identifying and analyzing tweets related to natural disasters on Twitter using Machine Learning algorithms such as Bernoulli Naive Bayes (BNB), Multinomial Naive Bayes (MNB), Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), and Random Forest (RF). A dataset of 122 thousand tweets with geographic locations is processed with stemming and lemmatization techniques. Training and testing results show that BNB, MNB, and LR models have the best performance with an accuracy rate of 87%, while KNN, DT, and RF reach 82%, 75%, and 86% [8].

In the journal Sentiment analysis and classification of COVID-19 tweets using machine learning classifier In March 2020, WHO identified COVID-19 as a new pandemic and issued a statement in connection with it. This research aims to classify pandemic-related sentiment using feature extraction and machine learning techniques on social media datasets. There are four phases in the system: data collection, pre-processing and normalization, feature extraction and feature selection, and classification. Data is collected from social media sources such as Twitter using the Twitter API. Tweets are then divided into three categories: positive, neutral, and negative during pre-processing. In the feature extraction phase, methods such as Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, and FastText are used to collect the feature dataset. In the last phase, various machine learning classification algorithms are applied to detect sentiment. In the experimental analysis, BoW with Modified Support Vector Machine (mSVM) gave better results than other machine learning algorithms, achieving an accuracy of 98.15%. mSVM also achieved 93% accuracy for positive sentiment, surpassing the performance of other machine learning classifiers [9].

In the journal A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification this paper discusses the construction of a BBC news text classification model using machine learning algorithms, focusing on logistic regression, random forest, and K-nearest neighbor. The study compared the performance of the three algorithms based on precision, accuracy, F1-score, support, and confusion matrix parameters. Logistic regression with TF-IDF Vectorizer feature achieved the highest accuracy (97%), making it a stable choice for small data sets. Random forest ranked second with 93% accuracy, while K-nearest neighbor achieved 92% accuracy. Logistic regression showed the expected performance, making it the best choice [10].

Based on a number of previous studies, it is evident that models such as Random Forest, K-Nearest Neighbors (KNN), and Recurrent Neural Network (RNN) provide good classification results in various situations. By applying the models that have been developed, this research makes a significant contribution to improving the effectiveness of disaster management strategies in Indonesia. The performance evaluation in this research is conducted using Confusion Matrix, ROC, Precision, and Recall methods. The results of this evaluation will be used to compare the performance between models with the aim of finding the most optimal model.

2. Method

The initiation of this study involves the utilization of Twitter/X data, motivated by the observable surge in Indonesia's user base over the preceding year. The adoption of a Python-based crawling approach is deemed imperative for the execution of this research. It is noteworthy that Twitter/X imposes constraints on data accessibility, leading to the deferment of research-related data acquisition activities.

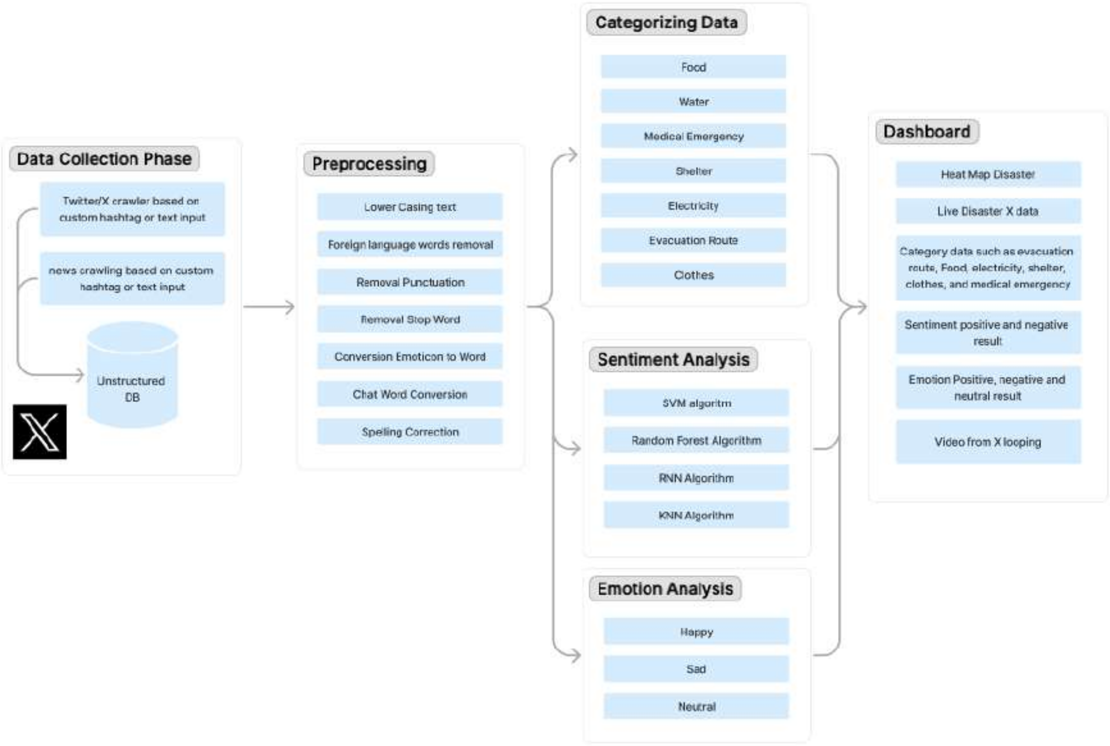


Figure 1.
Flow diagram to enhance disaster response.

We have developed a comprehensive system for disaster-related data collection and analysis, encompassing both news articles and Twitter data. Our data crawling system efficiently gathers information on various aspects, such as food, water, medical emergencies, shelter availability, electricity status, evacuation routes, and clothing needs. This collected data is then categorized systematically to facilitate a clear understanding of the diverse dimensions of disaster-related challenges. Additionally, our system incorporates sentiment analysis to gauge public opinion and emotion analysis to understand the emotional tone of the collected data. To enhance user accessibility and comprehension, we have implemented a user-friendly dashboard system for data visualization. This dashboard serves as a centralized platform, presenting insightful visual representations of the categorized data, sentiment trends, and emotional dynamics, thereby providing valuable insights for efficient disaster response and management.

Each of system has connected through database, database can be expanded as data will grow overtime. The main idea of this system is to prevent false data and making best decision based on fact data. There is system to validate data from crowd sourcing, government and other party can contribute data and validate them in one station. Architecture of the system are important to expand the system into bigger data/information.

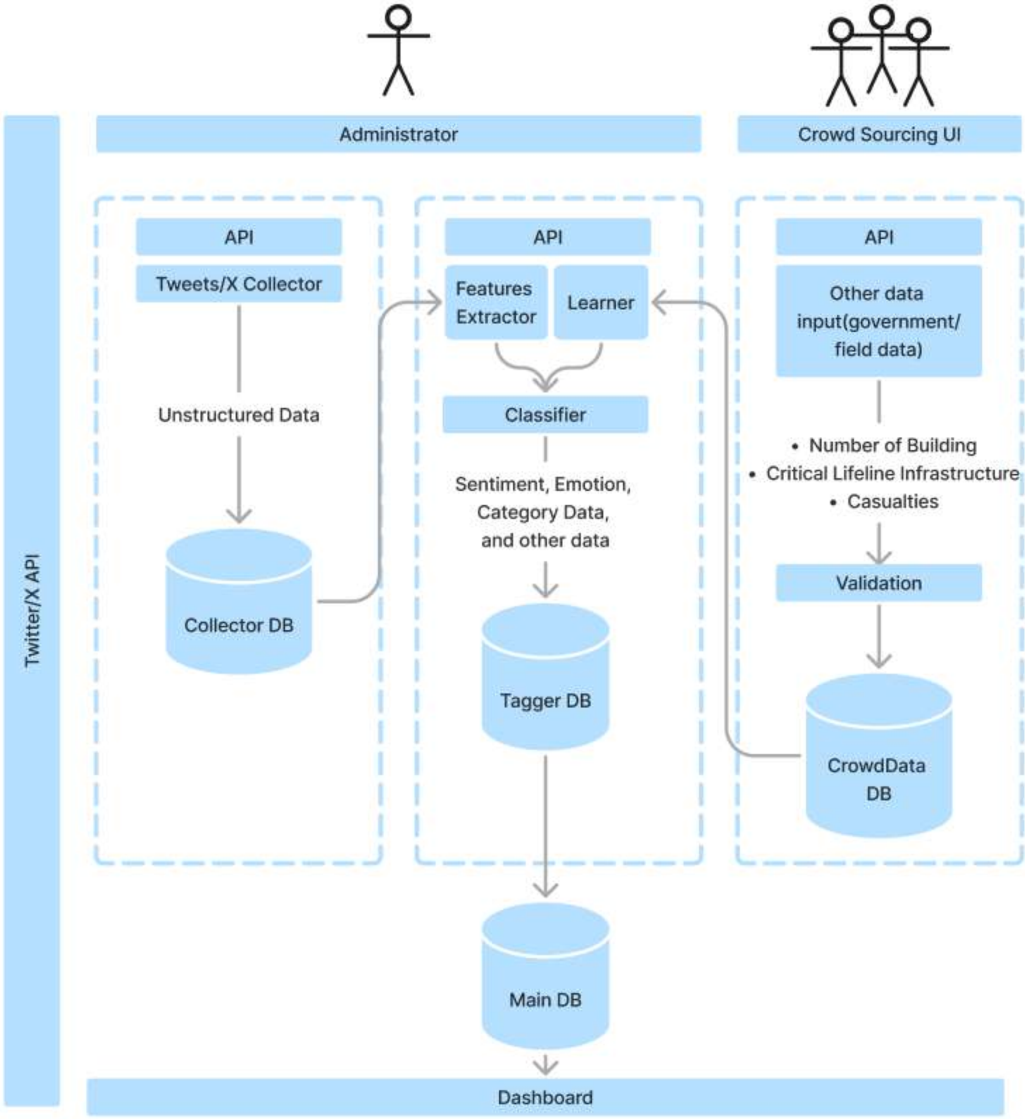


Figure 2.
Architecture model for disaster response.

2.1. Data Collection

In our research, we have collected an extensive dataset comprising over 9,000 Twitter posts focusing on four distinct disasters that frequently occur in Indonesia due to its geographical location on the Pacific "Ring of Fire." [11]. This initial dataset serves as a foundation for our investigation into the dynamics of disasters, including tsunamis, earthquakes, floods, landslides, droughts, and forest fires. Recognizing the evolving nature of disaster events, our research aims to expand its scope by incorporating data from crowd-sourcing platforms and other social media channels. Indonesia's vulnerability to a spectrum of natural disasters necessitates a comprehensive approach to understanding and mitigating their impact, and the integration of real-time information from diverse sources will contribute to the development of a robust and adaptable disaster response system.

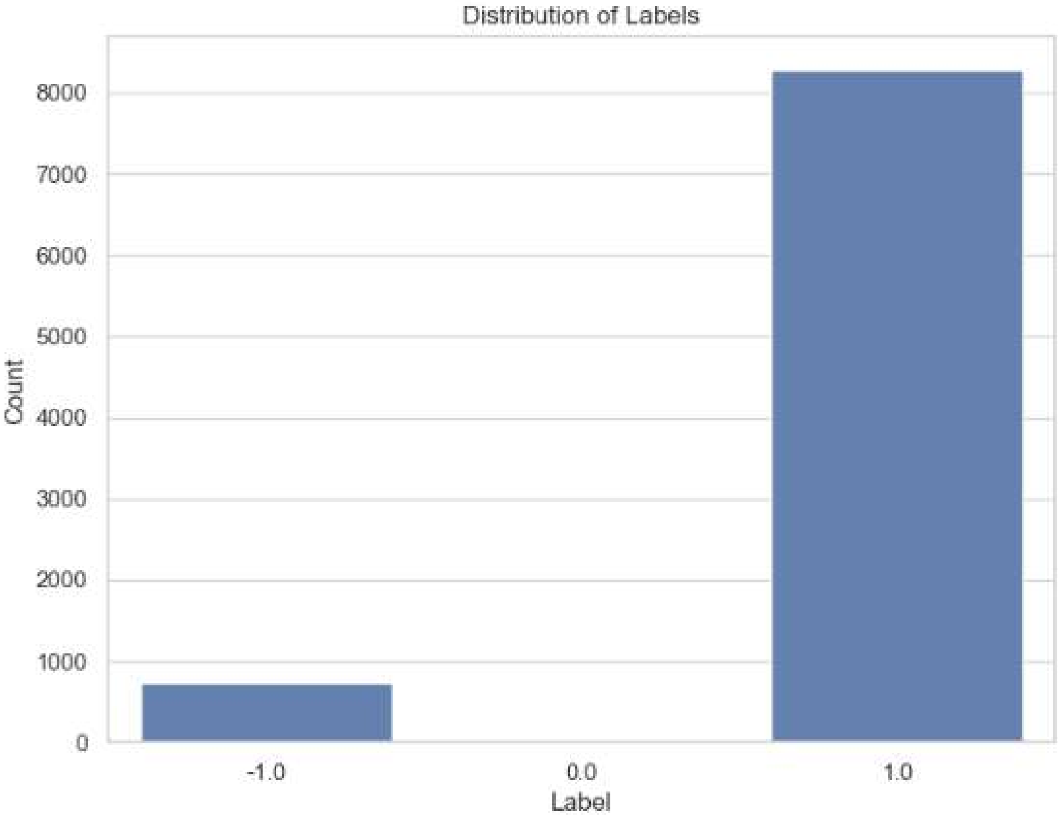


Figure 3.
Distribution of labels 1 for positive and -1 for negative.

2.2. Pre-Processing

The preprocessing pipeline for textual data encompasses a series of essential steps aimed at refining raw text into a format conducive to effective natural language processing (NLP) [12]. Case folding initiates this process by converting all characters to lowercase, ensuring uniformity and eliminating discrepancies arising from differing letter cases. Subsequently, the cleansing phase involves the removal of extraneous characters, symbols, or HTML tags, further enhancing the cleanliness and coherence of the text. Stopword removal follows, targeting common words that contribute little semantic meaning, streamlining the dataset for more meaningful analysis [13].

Tokenization plays a pivotal role in breaking down sentences into individual words or tokens, facilitating subsequent feature extraction and analysis. Object stemming is then employed to reduce words to their root form, aiding in the consolidation of related terms. Label encoding is utilized to convert categorical labels into numerical representations, a crucial step in preparing data for machine learning models [14].

Finally, the TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer methodology is applied to transform the tokenized and stemmed data into numerical vectors, emphasizing the importance of words within the corpus [15]. This comprehensive methodology, incorporating case folding, cleansing, stopword removal, tokenization, object stemming, label encoding, and TF-IDF vectorization, lays the foundation for robust and meaningful analysis in various NLP applications. The paper navigates through the intricacies of each step, elucidating their individual contributions and collective impact on optimizing textual data for subsequent processing and analysis [16].

DOI: 10.55214/25768484.v8i6.2033

In this case, Term Frequency Inverse Document Frequency (TF-IDF) is used to convert text into a value [1] So that it can be calculated using K-Means clustering. Following are the equations for TF-IDF:

$$Wdt = tfdt * Idft \qquad (1)$$

In this context, Wdt represents the weight of the d document against the t-word, $tfdt$ denotes the number of times the word is searched for in a document, and $Idft$ is the Inverse Document Frequency calculated as $(\log(\frac{N}{df}))$, where N is the total number of documents and df is the number of documents containing the searched word.

2.3. Sentiment Analysis

Sentiment analysis, also known as sentiment analysis or opinion mining, is a discipline in natural language processing that aims to identify, analyze, and evaluate sentiments or opinions contained in text [18]. The focus is on determining whether the text contains positive, negative, or neutral sentiments related to a particular subject or topic. Using techniques such as machine learning and statistical analysis, sentiment analysis is widely used in a variety of contexts, including in understanding customer opinions on products, assessing social responses to news, and providing valuable insights for text-based decision-making [19].

2.4. KNN

K-Nearest Neighbors (KNN) is one of the simple and commonly used classification algorithms in machine learning [20]. The fundamental concept of KNN involves classifying data based on the majority class of a certain number of its nearest neighbors [21]. In the given implementation context, the KNN model is configured with specific parameters. Firstly, `n_neighbors` is a hyperparameter that determines the number of neighbors to consider when making predictions for new data. Values [40, 50, 60, 70, 80, 90] are used to search for the best combinations. Additionally, `metric` is a hyperparameter that specifies the distance metric used to measure the proximity between data points. In this case, the distance metric used is 'manhattan'.

The hyperparameter tuning process is conducted using `RandomizedSearchCV`, a method that performs a random search over a predefined set of parameters. In the given example, the search is conducted over combinations of values for `n_neighbors` and `metric` to find a combination that yields the best model performance. Other settings include utilizing all CPU cores (`n_jobs=-1`) for parallel processing and displaying progress messages (`verbose=1`). Subsequently, the KNN model utilizes `KNeighborsClassifier` from scikit-learn to be trained on the training data (`x_train_bow` and `y_train`). After undergoing the tuning process, the best results regarding hyperparameter combinations and model performance can be accessed through the `best_params_` and `best_score_` attributes of the `RandomizedSearchCV` object.

2.5. Random Forest

Random Forest is a machine learning algorithm that falls under the category of ensemble learning, where multiple learning models are combined to improve performance and prediction accuracy [22]. Specifically, Random Forest uses a large number of decision trees created randomly on different subsets of data. Each tree provides predictions, and their results are combined through majority voting to generate the final prediction. The main advantage of Random Forest lies in its ability to overcome overfitting and provide stable and accurate prediction results for various types of datasets. This algorithm can be applied to both classification and regression tasks [23].

In implementing the Random Forest Classifier using the scikit-learn library in Python, the model is initialized with the n_estimators parameter set to 200 to indicate the number of decision trees in the ensemble. Additionally, the random_state parameter is set to 0 for reproducibility purposes. The training process then occurs using the provided training data, where X_train represents the features and y_train represents the labels. RandomForestClassifier is an ensemble learning method that constructs multiple decision trees during training, and its output is the mode of the classes (for classification) or the mean prediction (for regression) of the individual trees. In this context, the model is configured to build an ensemble with 200 decision trees, an approach that often yields a stable and accurate classification model suitable for various machine learning tasks.

2.8. RNN

Implement an RNN (Recurrent Neural Network) model using Keras in TensorFlow for a binary classification task [24]. *The model* structure consists of several key layers. First, an input layer is created to accept sequences of words with a maximum length of max_len. Next, an embedding layer is used to convert the words into a vector representation with a dimension of 50. An LSTM (Long Short-Term Memory) layer is then added to understand the temporal patterns in the sequences of words.

After that, further processing is done with a 256-unit dense layer and ReLU activation function, followed by an additional ReLU activation layer. The addition of a dropout layer with a dropout rate of 0.5 aims to reduce the risk of overfitting. An output layer with one unit and sigmoid activation function is used to perform binary classification according to the given task. The model is further compiled using binary crossentropy as a loss function, RMSprop as an optimizer, and evaluation metrics involving accuracy and precision. The training process is performed using training data (sequences_matrix) and labels (y_train) in batches of 128 for 10 epochs, with 20% of the data used as validation data. The training results, such as accuracy and loss, are recorded in the history variable. The ultimate goal of this implementation is to generate an RNN model capable of classifying sequential data with high accuracy in a binary classification task.

2.9. Confusion matrix

Confusion matrix is a commonly used evaluation tool in classification to evaluate the performance of machine learning models [25]. In this study, the confusion matrix helps to understand the extent to which the model is able to classify disaster-related data on Indonesian social media using the K-Nearest Neighbors (KNN), Random Forest, and Recurrent Neural Network (RNN) methods. The confusion matrix contains four main components that make up the matrix as follows [26]:

Table 1.
Confusion matrix.

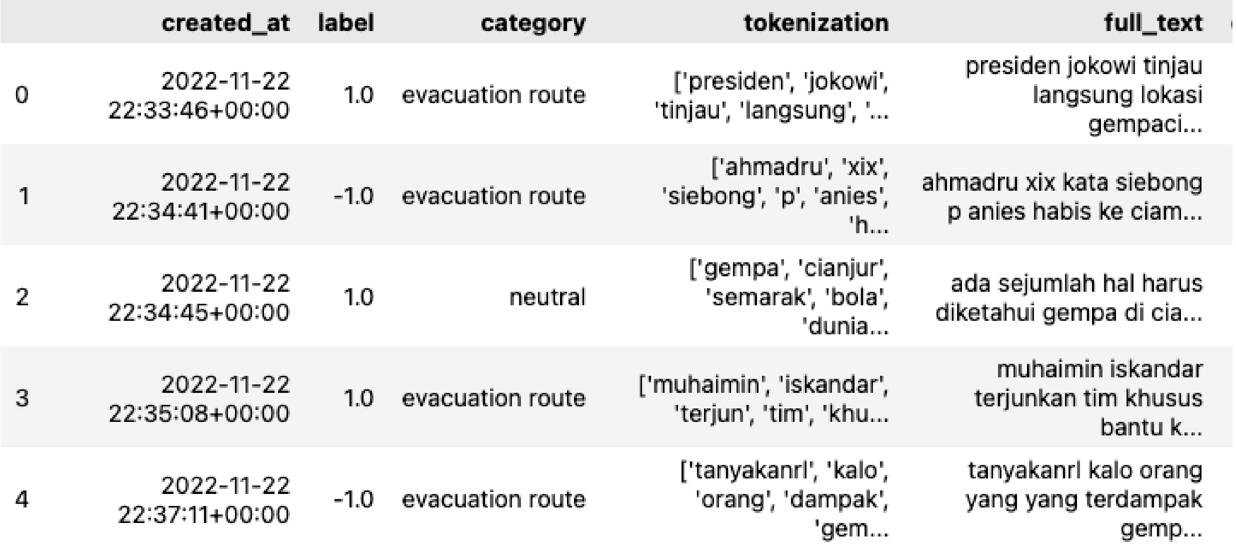
	Predicted positive	**Predicted negative**
Actual positive	True positive (TP)	False negative (FP)
Actual negative	False positive (FP)	True negative (FN)

From the table above, the four components can be explained as follows:

- True Positive (TP):
  This is the amount of data that actually falls into the positive category (disaster) and is successfully predicted correctly by the model.
- True Negative (TN):
  Represents the amount of data that actually falls into the negative category (non-disaster) and is correctly predicted by the model.
- False Positive (FP):
  Represents the amount of data that actually belongs to the negative (non-disaster) category, but was incorrectly predicted as positive (disaster) by the model.
- False Negative (FN):
  Represents the amount of data that actually belongs to the positive (disaster) category, but is incorrectly predicted as negative (non-disaster) by the model.

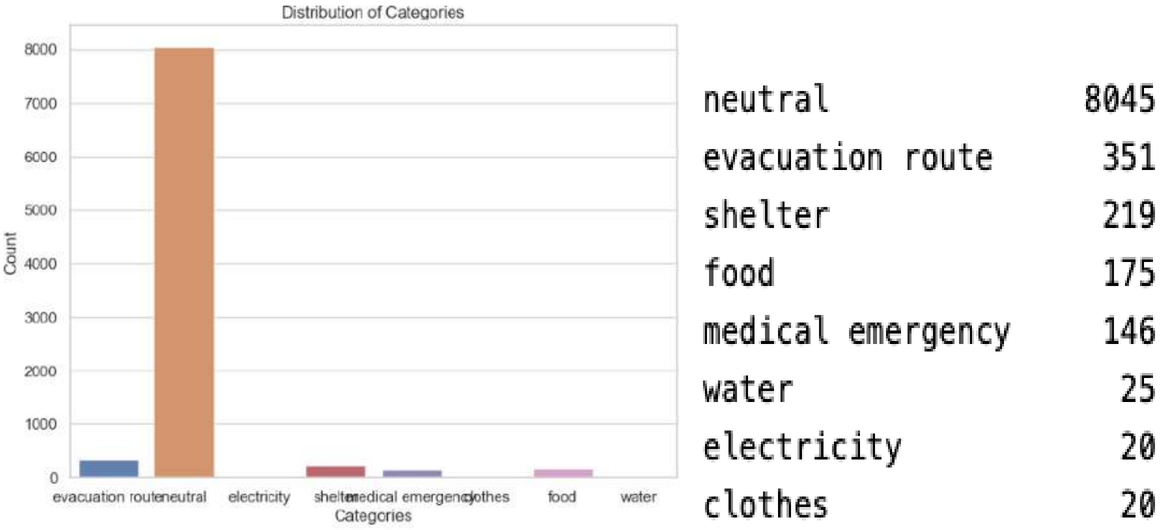
3. Results and Discussion

In an effort to provide a more in-depth understanding of the label distribution within the proposed classification system, this information is visually presented through a bar chart. This chart not only offers a clear visual representation but also provides a comprehensive overview of the prevalence of each behavioral disaster data category. By leveraging graphical representation, researchers and stakeholders can easily discern the comparisons between the quantities of data associated with specific behaviors in the context of disasters. This approach aims to enhance comprehension of data distribution and lay a robust foundation for further decision-making in the development of an effective classification system.



	created_at	label	category	tokenization	full_text
0	2022-11-22 22:33:46+00:00	1.0	evacuation route	['presiden', 'jokowi', 'tinjau', 'langsung', '...	presiden jokowi tinjau langsung lokasi gempaci...
1	2022-11-22 22:34:41+00:00	-1.0	evacuation route	['ahmadru', 'xix', 'siebong', 'p', 'anies', 'h...	ahmadru xix kata siebong p anies habis ke ciam...
2	2022-11-22 22:34:45+00:00	1.0	neutral	['gempa', 'cianjur', 'semarak', 'bola', 'dunia...	ada sejumlah hal harus diketahui gempa di cia...
3	2022-11-22 22:35:08+00:00	1.0	evacuation route	['muhaimin', 'iskandar', 'terjun', 'tim', 'khu...	muhaimin iskandar terjunkan tim khusus bantu k...
4	2022-11-22 22:37:11+00:00	-1.0	evacuation route	['tanyakanrl', 'kalo', 'orang', 'dampak', 'gem...	tanyakanrl kalo orang yang yang terdampak gemp...

Figure 4.
Data category of Evacuation route, shelter, food, medical emergency, water, electricity, and clothes.



```
neutral              8045
evacuation route      351
shelter               219
food                  175
medical emergency     146
water                  25
electricity            20
clothes                20
```

Figure 5.
Distribution of category.

The initial phase of data preprocessing is a crucial step that encompasses several essential tasks aimed at refining and structuring the raw text data for subsequent analysis. Among these tasks, the removal of mentions, URLs, numbers, and punctuations plays a pivotal role in streamlining the text corpus. By eliminating extraneous elements such as social media mentions, web links, numeric characters, and punctuation marks, the processed dataset becomes cleaner and more focused. This meticulous cleaning process contributes to the creation of a high-quality dataset, setting the foundation for enhanced accuracy in downstream natural language processing (NLP) tasks. The refined dataset not only aids in mitigating noise and irrelevant information but also facilitates a more effective and precise analysis of textual content, ultimately improving the overall performance of NLP models and algorithms applied to the data.

Table 1.
Removal of mentions, URLs, numbers, and punctuations in Indonesian.

Text	**Text with removal mentions, emoji, URLS, numbers and punctuations**
udah pada baca ini? 👇🏻 Kapolri Akan Kerahkan Tenaga Medis Tambahan #PenangananPascaGempa Untuk Peduli Warga Cianjur Kapolri meninjau langsung kondisi masyarakat terdampak gempa bumi yang dirawat di Rumah Sakit Bhayangkara Cianjur, Jawa Barat, Selasa (22/11).	udah pada baca ini Kapolri Akan Kerahkan Tenaga Medis Tambahan PenangananPascaGempa Untuk Peduli Warga Cianjur Kapolri meninjau langsung kondisi masyarakat terdampak gempa bumi yang dirawat di Rumah Sakit Bhayangkara Cianjur Jawa Barat Selasa

In an effort to preserve contextual nuances, stop words are retained during text analysis. This decision aims to provide a deeper and more nuanced understanding of the language while preventing potential information loss that may occur if stop words are removed. Although stop words may be considered common or less meaningful individually, their presence can contribute to the overall meaning in the context of sentences and documents. By retaining stop words, text analysis can more accurately represent the complexity of language and context, enabling a more holistic understanding of the meaning contained in the text.

Table 2.
Stop removal text analysis in Indonesian.

Text	**Text with stop removal**
Bener-bener ya yuk udah yuk gempa susulannya kasian teman saudara dsana mana kemarin ujan skrg pun matahari sulit Dan entah kenapa aku selalu ikut rasain getaran gempa susulannya walau kecil kecil aja udah setakut itu aplgi org Cianjur yaak	Bener-bener yayuk yuk gempa susulannya kasian teman saudara dsana kemarin ujan matahari sulit. aku ikut rasain getaran gempa susulannya kecil kecil setakut org Cianjur

Before inputting data into the algorithm, the dataset was split into training and testing sets with a ratio of 67-33 for KNN, 80-20 for random forest, and RNN models. This step was taken to ensure consistency in the execution of the experiments. Next, the three different algorithms, namely KNN, random forest, and RNN, were applied to analyze the data with the aim of gaining insights. This approach provides a deeper understanding of the performance of the three algorithms in the context of the dataset used. The results of this analysis can provide a more comprehensive perspective on the ability and suitability of each algorithm in processing data, which can be used as a basis for further decision-making in the development of models or systems related to the dataset.

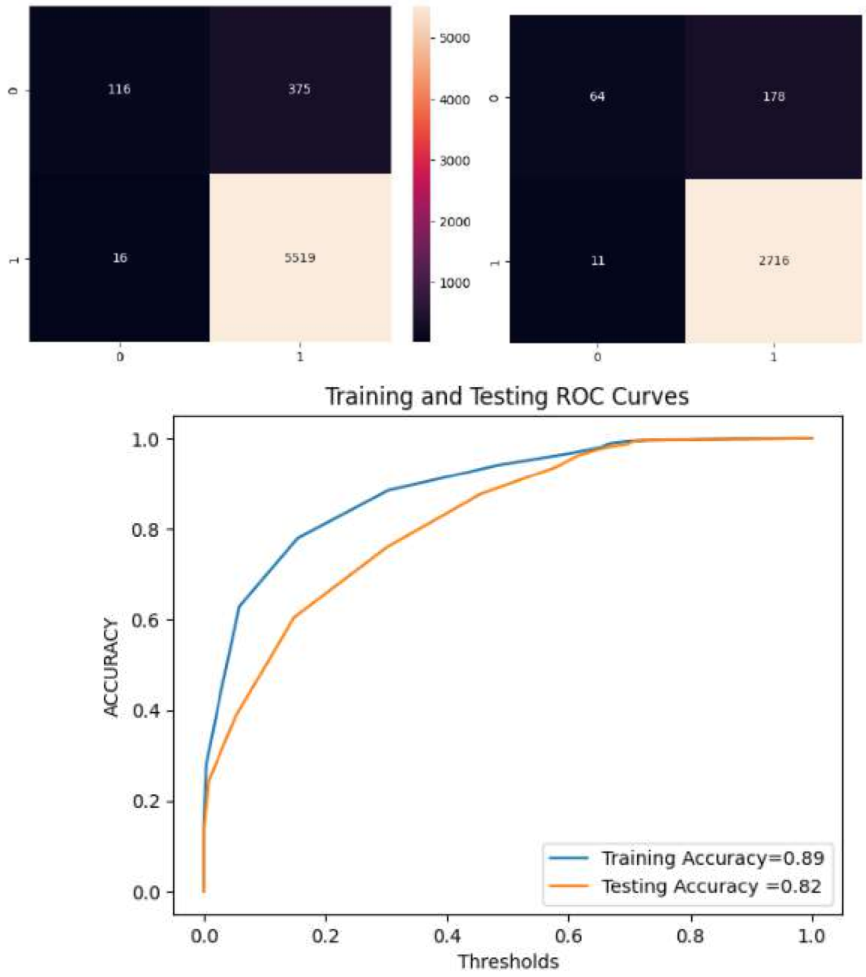


Figure 6.
Result of KNN confusion matrix data train, confusion matrix data test and ROC-Curve.

From the confusion matrix in the training stage, the KNN model successfully classified 5,000 samples correctly as natural disasters and 116 samples correctly as non-natural disasters. However, there were 375 errors where the model classified samples that should not have been natural disasters as natural disasters, and 16 errors where samples that were actually natural disasters were classified as non-natural disasters. The overall evaluation shows that the KNN model performed satisfactorily in classifying natural disaster data on Twitter social media in the training stage.

In the test stage, the confusion matrix shows that the KNN model successfully classified 2,500 samples correctly as natural disasters and 1 sample correctly as non-natural disasters. However, there were 64 errors where the model classified samples that should not have been natural disasters as natural disasters, and 11 errors where samples that should have been natural disasters were classified as non-natural disasters. The overall evaluation shows that it performs relatively well but needs improvement in classifying test data related to natural disasters on social media Twitter.

The ROC curves provide further insight into the performance of the KNN model. The curve for the training data shows good performance, achieving an accuracy of 0.89 with a TPR of about 0.89 at an FPR of about 0.11. However, on the test data, there is a slight drop in performance with an accuracy of 0.82 and a TPR of about 0.82 at an FPR of about 0.18. Although the model was able to identify many true positives, there was an increase in false positives at lower confidence thresholds. This evaluation provides important insights for decision-making, highlighting the challenges of KNN models in generalization to test data and the need to pay attention to false positive management. In conclusion, the KNN model performed well, but improvements are needed to improve generalization and manage false positives on test data related to natural disasters on Twitter social media.

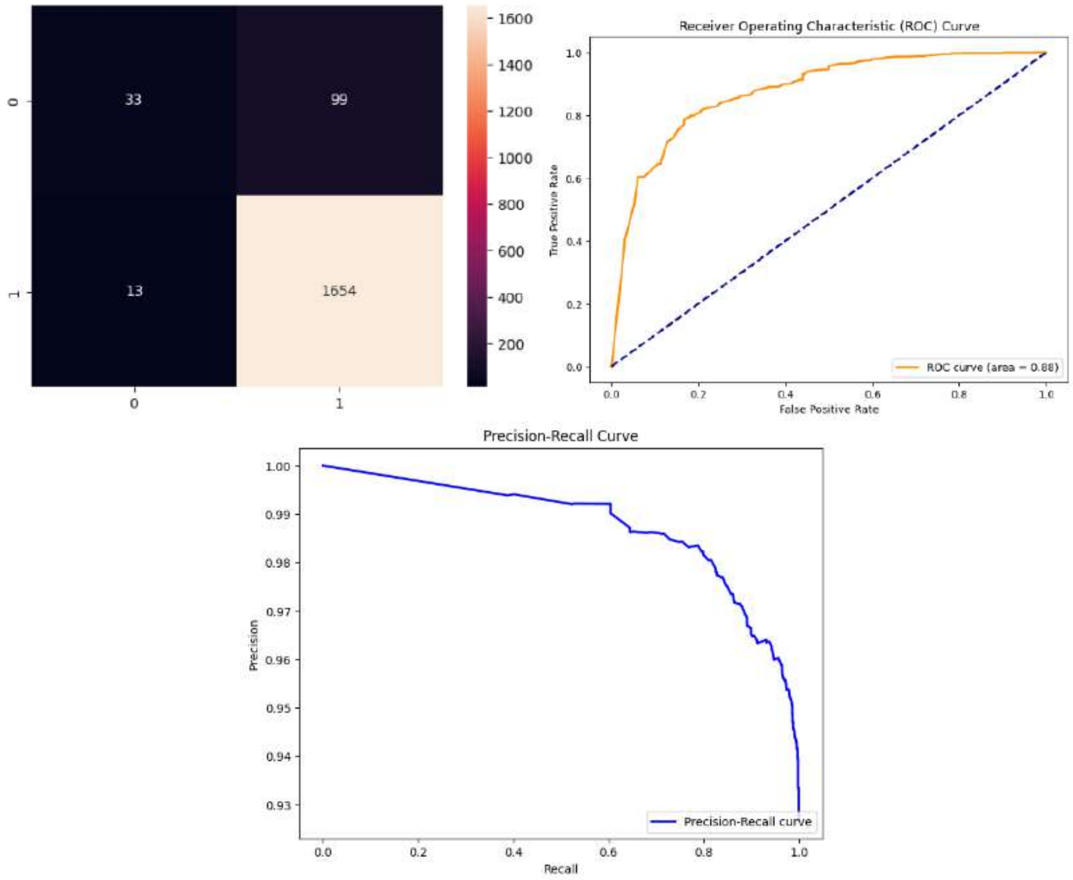


Figure 7.
Result of random forest with confusion matrix, ROC-Curve, and precision-recall curve.

Based on the Figure 7 analysis results involving the confusion matrix, Receiver Operating Characteristic (ROC) curve, and precision-recall curve of the RandomForest model, it can be concluded that the model shows excellent performance in classifying data between disaster and non-disaster situations. The model's ability to distinguish the context of twitts from disaster-related twitter/X makes it an effective tool in predicting the relevance of a conversation to disasters.

Specifically, the model has an accuracy of 94%, an AUC ROC of 0.88, and an AUC precision-recall of 0.95.

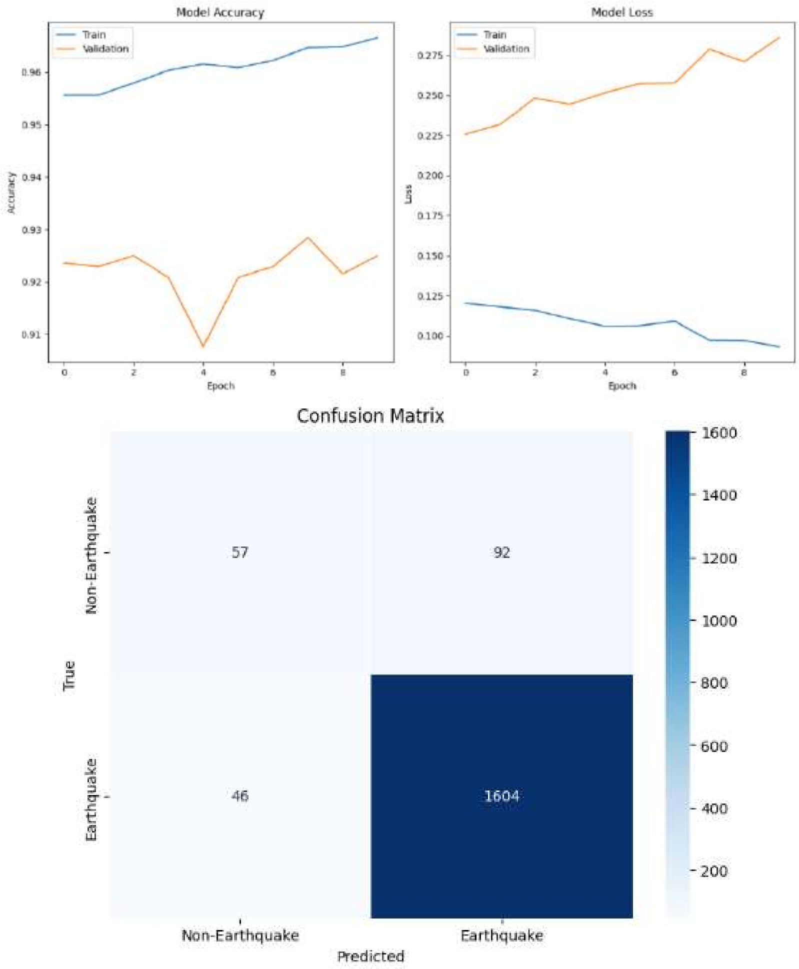


Figure 8.
Result of RNN with model accuracy, model loss, and confusion matrix.

The RNN model in Figure 8 exhibits excellent performance, with an accuracy of approximately 96% for the training data and 95% for the validation data. This high accuracy indicates the RNN model's ability to predict outputs accurately, both on the training data and on previously unseen data. The analysis of the model's loss also indicates that the RNN model effectively learns from the training data, with a significant decrease in loss from the beginning of training until the 8th epoch. This demonstrates the model's capability to reduce errors with an increasing number of epochs. Overall, the accuracy and loss results of the RNN model signify its strong performance and reliability for various tasks such as classification, prediction, and pattern recognition.

Based on the confusion matrix in Figure 8, the image illustrates the results of the classification report of the RNN model used to classify earthquake data. The classification report is a table that shows the model's predictions on the test data, consisting of four main cells: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). According to the Figure 8, it can be concluded that the RNN model exhibits good performance in classifying earthquake data. The details include the model accurately predicting 1604 out of 1609 earthquake instances (TP), correctly identifying 92 out of 91 non-earthquake instances (TN), making incorrect predictions for 46 non-earthquake instances as earthquakes (FP), and erroneously classifying 9 earthquake instances as non-earthquakes (FN).

Table 3.

Performance sentiment analysis of Indonesian language on disaster response.

Method	Accuracy
KNN	82%
Random forest	94%
RNN	96%

From the table above, three different models were evaluated based on accuracy. The K-Nearest Neighbors (KNN) model achieved an accuracy of 82%, signifying its ability to handle data based on nearest neighbors. While this result is satisfactory, there are indications that the KNN model may be less effective in handling complex patterns or high-dimensional datasets. On the other hand, Random Forest achieved 94% accuracy, demonstrating its ability to combine decision trees to improve performance and reduce overfitting. These results illustrate that Random Forest can be a more optimal choice for handling different types of datasets. Finally, the Recurrent Neural Network (RNN) model showed the highest performance with 96% accuracy. The advantage of RNN lies in its ability to understand temporal relationships and sequential patterns in data, making it a strong choice for sequential data such as text or time sequences. Therefore, the selection of an appropriate model needs to consider the specific characteristics of the dataset and the needs of the desired classification analysis.

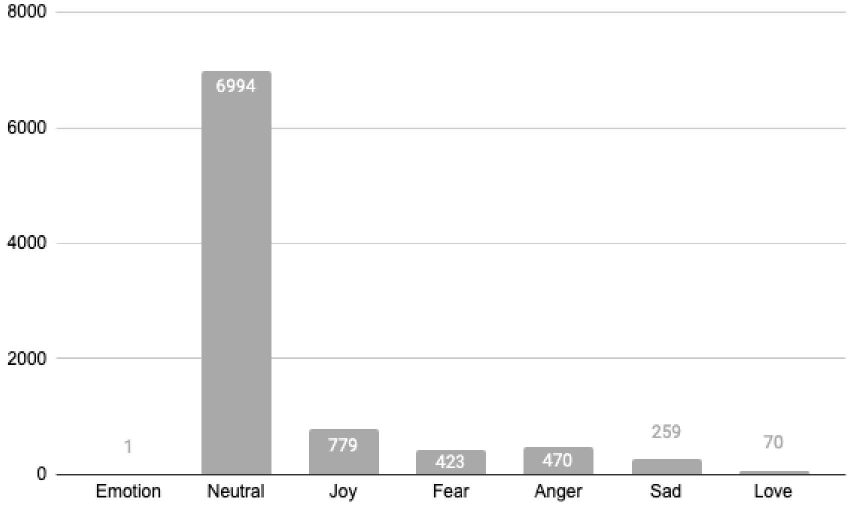


Figure 9.
Graph result emotion detected in dataset.

The bar chart above provides an overview of the distribution of emotional responses in the Twitter dataset focusing on disaster-related content. Out of a total of 6,994 tweets, the 'Neutral' category dominates, indicating that most disaster-related tweets are informative or factual without strong emotional overtones. This could include updates on the progress of the disaster, safety information, or factual statements about its impact. In the context of disasters, Twitter serves as a hub for real-time information dissemination, with many users seeking objective updates.

While the mood is generally serious in the context of a disaster, there were 779 tweets categorized as 'Joy'. These may include messages of relief, stories of successful rescues, or thanks for support and assistance. Furthermore, 423 tweets expressed 'Fear', reflecting concerns for safety, uncertainty of the situation, or anxiety about the future.

'Anger' was present in 470 tweets, reflecting frustration at the perceived inadequacy of the disaster response or anger that human factors were to blame. 'Sad' tweets (259 tweets) were more likely to express sadness, loss or sympathy, especially in relation to loss of life, loss of property or loss of livelihood.

The emotion 'Love' was the least frequently detected (70 tweets), representing messages of support, solidarity and compassion. This could include tweets from individuals expressing love and concern for those affected or gratitude to first responders and volunteers. This analysis reveals the diversity of emotional responses in the context of disasters on Twitter, covering a spectrum from neutral to more intense emotions.

4. Conclusion

From the evaluation results of three different models in the study "Behavioural Indonesian Disaster Data Classification in Social Media Using KNN, Random Forest, and RNN in Machine Learning," it can be concluded that each model has its own strengths and characteristics. The K-Nearest Neighbors (KNN) model achieved an accuracy of 82%, indicating its ability to handle data based on nearest neighbors. While the results are satisfactory, there is an indication that KNN may be less effective in handling complex patterns or high-dimensional datasets.

On the other hand, Random Forest achieved an accuracy of 94%, showcasing its ability to combine decision trees to enhance performance and reduce overfitting. These results illustrate that Random Forest could be a more optimal choice for handling various types of datasets.

Ultimately, the Recurrent Neural Network (RNN) model demonstrated the highest performance with an accuracy of 96%. RNN's primary strength lies in its ability to understand temporal relationships and sequential patterns in data, making it a robust choice for sequential data like text or time series.

In model selection, it is crucial to consider the specific characteristics of the dataset and the desired classification analysis. Thus, these findings provide valuable insights to guide the selection of the most suitable model for the situation and data type at hand. For future research, further exploration is needed on how these models can be adapted or optimized for various types of disasters and social media data. A deeper understanding of the factors influencing model performance in the context of disasters in Indonesia can serve as a foundation for developing more sophisticated and responsive models.

Copyright:

References

[1] "ROLE OF SOCIAL MEDIA IN DISASTER MANAGEMENT," *International Journal of Public Health and Clinical Sciences*, vol. 6, no. 2, 2019, doi: 10.32827/ijphcs.6.2.77.

[2] H. Zhang *et al.*, "Spatiotemporal Information Mining for Emergency Response of Urban Flood Based on Social Media and Remote Sensing Data," *Remote Sens (Basel)*, vol. 15, no. 17, 2023, doi: 10.3390/rs15174301.

[3] N. Sheng, C. Yang, L. Han, and M. Jou, "Too much overload and concerns: Antecedents of social media fatigue and the mediating role of emotional exhaustion," *Comput Human Behav*, vol. 139, 2023, doi: 10.1016/j.chb.2022.107500.

[4] S. M. Khan *et al.*, "A Systematic Review of Disaster Management Systems: Approaches, Challenges, and Future Directions," *Land*, vol. 12, no. 8. 2023. doi: 10.3390/land12081514.

[5] J. Jethwaney, "Introduction: Taking Media and Communication on Board in Disaster Management," in *International Handbook of Disaster Research*, 2023. doi: 10.1007/978-981-16-8800-3_207-1.

[6] S. Sohrabizadeh *et al.*, "Challenges and Barriers to the Participation of Non-Governmental Organizations (NGOs) in Flood Management: A Qualitative Study from Iran," *Disaster Med Public Health Prep*, vol. 17, no. 2, 2023, doi: 10.1017/dmp.2022.123.

[7] A. A. Shah, A. Ullah, G. T. Mudimu, N. A. Khan, A. Khan, and C. Xu, "Reconnoitering NGOs strategies to strengthen disaster risk communication (DRC) in Pakistan: A conventional content analysis approach," *Heliyon*, vol. 9, no. 7, 2023, doi: 10.1016/j.heliyon.2023.e17928.

[8] O. Iparraguirre-Villanueva *et al.*, "Classification of Tweets Related to Natural Disasters Using Machine Learning Algorithms," *International Journal of Interactive Mobile Technologies*, vol. 17, no. 14, 2023, doi: 10.3991/ijim.v17i14.39907.

[9] C. S. Lakshmi, S. Saxena, and B. S. Kumar, "Sentiment analysis and classification of COVID-19 tweets using machine learning classifier," *Journal of Autonomous Intelligence*, vol. 7, no. 2, pp. 1–13, 2024, doi: 10.32629/jai.v7i2.801.

[10] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," *Augmented Human Research*, vol. 5, no. 1, 2020, doi: 10.1007/s41133-020-00032-0.

[11] M. Masum and M. Ali Akbar, "The Pacific Ring of Fire is Working as a Home Country of Geothermal Resources in the World," in *IOP Conference Series: Earth and Environmental Science*, 2019. doi: 10.1088/1755-1315/249/1/012020.

[12] L. Li, J. Geissinger, W. A. Ingram, and E. A. Fox, "Teaching Natural Language Processing through Big Data Text Summarization with Problem-Based Learning," *Data Inf Manag*, vol. 4, no. 1, 2020, doi: 10.2478/dim-2020-0003.

[13] F. Ali *et al.*, "Transportation sentiment analysis using word embedding and ontology-based topic modeling," *Knowl Based Syst*, vol. 174, 2019, doi: 10.1016/j.knosys.2019.02.033.

[14] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical data for neural networks," *J Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00305-w.

[15] Z. Labd, S. Bahassine, K. Housni, F. Z. A. Hamou Aadi, and K. Benabbes, "Text classification supervised algorithms with term frequency inverse document frequency and global vectors for word representation: a comparative study," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 1, p. 589, Feb. 2024, doi: 10.11591/ijece.v14i1.pp589-599.

[16] X. Tan, Y. Wang, C. Liu, X. Yuan, and K. Wang, "Unlocking operational excellence: A deep dive into a communication-driven multi-strategy state transition algorithm for industrial process optimization," *Chemometrics and Intelligent Laboratory Systems*, vol. 240, 2023, doi: 10.1016/j.chemolab.2023.104934.

[17] S. A. P. Raj and Vidyaathulasiraman, "Determining Optimal Number of K for e-Learning Groups Clustered using K-Medoid," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 400–407, 2021, doi: 10.14569/IJACSA.2021.0120644.

[18] T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, and L. Galligan, "Sentiment analysis and opinion mining on educational data: A survey," *Natural Language Processing Journal*, vol. 2, 2023, doi: 10.1016/j.nlp.2022.100003.

[19] P. William, A. Shrivastava, P. S. Chauhan, M. Raja, S. B. Ojha, and K. Kumar, "Natural Language Processing Implementation for Sentiment Analysis on Tweets," in *Lecture Notes in Networks and Systems*, 2023. doi: 10.1007/978-981-19-7982-8_26.

[20] M. Arunadevi, M. Rani, R. Sibinraj, M. K. Chandru, and C. Durga Prasad, "Comparison of k-nearest Neighbor & Artificial Neural Network prediction in the mechanical properties of aluminum alloys," *Mater Today Proc*, 2023, doi: 10.1016/j.matpr.2023.09.111.

[21] M. A. Araaf, K. Nugroho, and D. R. I. M. Setiadi, "Comprehensive Analysis and Classification of Skin Diseases based on Image Texture Features using K-Nearest Neighbors Algorithm," *Journal of Computing Theories and Applications*, vol. 1, no. 1, 2023, doi: 10.33633/jcta.v1i1.9185.

[22] S. González, S. García, J. Del Ser, L. Rokach, and F. Herrera, "A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities," *Information Fusion*, vol. 64, 2020, doi: 10.1016/j.inffus.2020.07.007.

[23] Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, and X. Liang, "An improved random forest based on the classification accuracy and correlation measurement of decision trees," *Expert Syst Appl*, vol. 237, 2024, doi: 10.1016/j.eswa.2023.121549.

[24] D. Sarkar, R. Bali, and T. Ghosh, "Hands-On Transfer Learning with Python: Implement advanced deep learning and neural network models using TensorFlow and Keras," *Journal of Materials Processing Technology*, vol. 1, no. 1. 2018.

[25] D. Valero-Carreras, J. Alcaraz, and M. Landete, "Comparing two SVM models through different metrics based on the confusion matrix," *Comput Oper Res*, vol. 152, 2023, doi: 10.1016/j.cor.2022.106131.

[26] J. Li, H. Sun, and J. Li, "Beyond confusion matrix: learning from multiple annotators with awareness of instance features," *Mach Learn*, vol. 112, no. 3, 2023, doi: 10.1007/s10994-022-06211-x.