Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 5, October 2025, Page. 3419-3429
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4766

Comparing BERTBase, DistilBERT and RoBERTa in Sentiment Analysis for Disaster Response

Hafiz Budi Firmansyah[*1], Aidil Afriansyah[2], Valerio Lorini[3]

[1,2]Department of Informatics, Institut Teknologi Sumatera, Lampung, Indonesia
[3]European Parliament, Luxembourg

Email: [1]hafiz.budi@if.itera.ac.id

Received : May 23, 2025; Revised : Jul 2, 2025; Accepted : Jul 3, 2025; Published : Oct 16, 2025

Abstract

Social media platforms are vital for real-time communication during disasters, providing insights into public emotions and urgent needs. This study evaluates the performance of three transformer-based models—BERTBase, DistilBERT, and RoBERTa—for sentiment analysis on disaster-related social media data. Using a multilingual dataset sourced from the Social Media for Disaster Risk Management (SMDRM) platform, the models were assessed on classification metrics including accuracy, precision, recall, and weighted F1-score. The results show that RoBERTa consistently outperforms the others in classification performance, while DistilBERT offers superior computational efficiency. The analysis highlights the trade-offs between model accuracy and runtime, emphasizing RoBERTa's suitability for scenarios prioritizing accuracy, and DistilBERT's potential in time-sensitive or resource-constrained applications. These findings support the integration of sentiment analysis into disaster response systems to enhance situational awareness and decision-making.

Keywords : *Deep Learning, Disaster Response, Sentiment Analysis, Social Media.*

1. INTRODUCTION

Natural disasters, such as earthquakes and floods, are events that occur naturally and can have a significant impact on society. They can cause physical damage, loss of life, and psychological trauma. In addition, the resulting impact can be so severe that it may trigger subsequent natural disasters. The losses caused by natural disasters can also be considerable [22].

In the event of a disaster, affected individuals often turn to social media to share real-time information about the situation, including reports of damaged infrastructure, requests for aid, and updates on local conditions [6] [7]. The number of information is increasing that reflects the number of users in platform. For instance, Twitter has more than 611 millions active users in 2024 and mostly used by people to share the content [17][25]. This user-generated content can be a valuable source of situational awareness for first responders and emergency management agencies. However, the integration of social media data into operational response workflows remains challenging due to the unstructured and high-volume nature of the information [8] [16].

One of the primary challenges lies in the automated analysis of this data to extract actionable insights. Among various techniques, sentiment analysis has emerged as a promising approach to understanding public emotions, urgency levels, and the severity of reported incidents during disasters [11] [23][24] [26]. By identifying the emotional tone behind social media posts, sentiment analysis can help prioritize resources, detect distress signals, and enhance the responsiveness of emergency services.

In disaster response, some studies using image implement machine learning for classification and geolocation prediction task [14][15]. Using text, several studies have explored the use of sentiment

Vol. 6, No. 5, October 2025, Page. 3419-3429
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4766

analysis in the context of disaster response. A recent sentiment analysis study examined a large-scale dataset of 90,000 COVID-19-related tweets gathered during the initial two months of the pandemic (February–March 2020) [1]. The research compared the performance of traditional machine learning (ML) techniques, including support vector machines, naive Bayes, decision trees, and random forests—with deep learning models such as convolutional neural networks (CNNs) and bidirectional long short-term memory (BiLSTM), using different word embedding techniques like fastText, GloVe, and Word2Vec. The findings indicated that deep learning approaches consistently outperformed conventional ML methods in classifying tweets into negative, positive, and neutral sentiments. A study proposed a hybrid deep learning approach, IDBO-CNN-BiLSTM, which integrates an improved swarm intelligence algorithm to optimize model performance. Applied to tweets from Hurricane Harvey, the model demonstrated superior accuracy compared to other methods, highlighting its potential to support emergency response efforts [2]. Paul et al. (2023) utilized the CrisisLex dataset within an active learning framework to minimize annotation workload while refining transformer-based models for sentiment classification, showcasing the potential of semi-automated labeling in crisis contexts [21].

A study compared machine learning algorithm (Support Vector Machine, Naive Bayes, Random Forest and Logistic Regression) and deep learning algorithm (LSTM)for analysing social media. The result demonstrates that Support Vector Machine reached the highest performance [18]. Other study conducted a study comparing various SVM algorithm kernel to analyse the sentiment in social media for disaster happening in Indonesia. The study shows that linear kernel worked better for social media data [19].

A more recent study compared human-labeled and machine-labeled sentiment annotations on disaster-related social media posts. Their findings show that classifiers trained on human-labeled data consistently outperform those trained on automatically labeled data, highlighting the importance of annotation quality. They also observed that automated tools tend to overestimate positive sentiment, which may distort crisis-related insights in urgent scenarios [4].

Building on this foundation, the present study aims to evaluate and compare the performance of three transformer-based models—BERTBase, DistilBERT, and RoBERTa—for sentiment analysis in the context of disaster response. By systematically analyzing their effectiveness, this research seeks to identify the most suitable model for extracting emotional cues from social media data, thereby enhancing the ability of emergency responders to interpret public sentiment, prioritize interventions, and improve overall situational awareness during crises.

2. METHOD

This section covers the dataset and the analysis pipeline used in this study. The dataset subsection provides a detailed description of the data, while the pipeline subsection outlines the steps taken to analyze sentiment from social media content. Both components are essential to ensure the transparency and reproducibility of the study. A clear understanding of the data and methodology is critical for interpreting the results accurately.

2.1. Dataset

The dataset comprises 5,434 text entries curated by the Joint Research Centre of the European Commission via the Social Media for Disaster Risk Management (SMDRM) platform [3]. The SMDRM platform collects data from Twitter in near real-time, using a combination of keyword-based filtering and location targeting. Data acquisition is triggered manually during unexpected events such as earthquakes.

Once acquired, the data undergo a comprehensive processing pipeline built using Python and orchestrated with Apache Airflow. Textual data are first normalized and annotated using multilingual

Vol. 6, No. 5, October 2025, Page. 3419-3429
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4766

classifiers trained on twelve languages, employing pre-trained language models such as Language-Agnostic Sentence Representations (LASER). The platform supports binary and multi-label classification tasks for detecting event relevance and assessing impacts. Additionally, geolocation information is extracted using named entity recognition (NER) and matched against gazetteers to enrich each data point with spatial attributes. The entire process is modular and scalable, leveraging Docker containers to manage high-volume workloads during crisis peaks.

Although the SMDRM platform does not present formal performance metrics such as precision or recall, it incorporates qualitative evaluation mechanisms to ensure annotation quality and classification reliability. Impact-related messages were annotated through a two-level process by trained volunteers from the European Virtual Operations Support Team (VOST). Furthermore, visual inspections of aggregated classified tweets confirmed the validity of the geolocation and impact detection processes. Future work includes implementing a quantitative evaluation framework to assess classification accuracy and recall across different disaster scenarios.

The dataset covers four major disaster events: the 2021 Catania floods (207 posts), the 2021 European Union floods (1,120 posts), the 2020 Croatia earthquake (869 posts), and the 2010 Haiti earthquake (3,238 posts). Notably, the dataset includes social media content in 44 different languages. Figure 1 shows the data distribution in this study. After an initial pre-processing, the data is shrinking to 902 data. The decreasing size of data points is mainly due to the removal of duplicates or empty posts. The reduction ensures a more balanced and non-redundant dataset for training and evaluation.

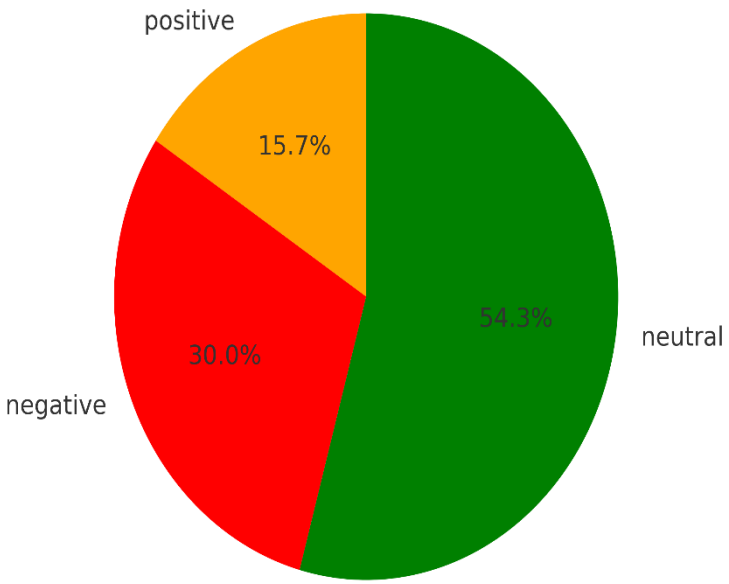

Figure 1. Data distribution used in this study

The labeling process of dataset includes manual method. The manual method involves human to label the data. The task includes asking three different human annotators for giving the label for each text. The annotators are hired from a crowdsourcing platform called projects.co.id. The crowdsourcing approach is commonly used in labeling the data [5] .The annotators selected for this task possessed a good understanding of English and a background in linguistics. To ensure the quality of the annotations, we provided them with detailed guidelines before they began the labeling process. All submitted annotations were manually reviewed prior to acceptance. After collecting the labels, we applied a majority voting approach to establish the ground truth. In rare cases where no agreement was reached, each annotator assigned a different label (positive, negative, and neutral)—we manually assigned the final label to ensure a reliable ground truth.

Vol. 6, No. 5, October 2025, Page. 3419-3429
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4766

Table 1. Sample of dataset

No	Text	Disaster	Label
1	International PAHO deploys experts to support Haiti	Haiti Earthquake	Positive
2	Haiti Earthquake Death toll risen over more than injured	Haiti Earthquake	Negative
3	Croatia First images from the M earthquake in SisakMoslavina County confirm damage	Croatia Earthquake	Negative
4	KLIMAWANDEL IN WUPPERTAL	EU Floods	Neutral
5	Bad weather in Catania Red Cross Unprecedented situation but we are ready with important means	Catania Floods	Neutral

Table 1 presents a sample of the dataset used in this study, which comprises social media posts related to various disaster events. Each entry includes the original text, the type of disaster referenced, and the corresponding sentiment label. The label categorizes as positive, negative, or neutral. The examples span several disaster contexts, including the Haiti Earthquake, Croatia Earthquake, EU Floods, and Catania Floods. These samples illustrate the range of emotional tones expressed in disaster-related communications, from supportive and optimistic messages (e.g., deployment of aid) to reports highlighting damage and loss.

2.2. Pipeline

This subsection presents the analysis pipeline employed in this study, drawing inspiration from the methodology outlined in [4]. The approach is structured to ensure a systematic and reproducible process aligned with the study's objectives.

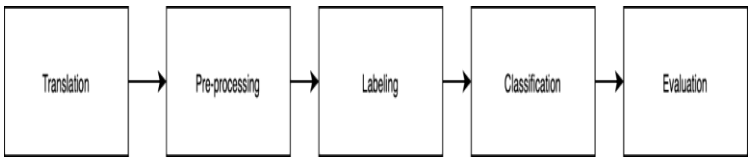

Figure 2. Pipeline used in this study

The approach comprises five main stages: translation, pre-processing, labeling, classification, and evaluation. Figure 2 illustrates the proposed methodology. Each stage is described in detail below:

a. Translation

Given that the dataset includes texts in 44 different languages, this step involves translating all entries into English. The objective is to establish a common linguistic base to facilitate consistent annotation and subsequent sentiment labeling. To translate the text, we rely on deep translator library. The library has some advantages including free access, support features and unlimited translation. It also can detect the misspelling in the text.

b. Pre-processing

This stage focuses on cleaning the textual data by removing stopwords and irrelevant characters. In this step, we also conducted case folding and tokenizing. Effective pre-processing enhances the quality of the input data and contributes to improved model performance [20].

Vol. 6, No. 5, October 2025, Page. 3419-3429
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4766

c. Labeling

Sentiment labels are assigned using a manual-approach strategy. The approach involves manual annotation by experts in linguistics and English to ensure accuracy and reliability in the labeling process. The number of annotators is three to ensure non-bias result.

d. Classification

In this phase, machine learning models are trained to categorize the text data into three sentiment classes: positive, neutral, and negative. The models are trained using three different transformer-based architecture including BERTBase, DistilBERT, and RoBERTa

e. Evaluation

The final step assesses the performance of the classification models using standard evaluation metrics such as accuracy, precision, recall, and weighted F1.

3. RESULT

This section presents result of the experiment. It starts with translation, pre-processing, labeling, classification, model performance evaluation followed by model runtime evaluation. The performance and runtime evaluation measurements aim to provide more holistic result in the experiment.

3.1. Translation

In this step, all data was translated into English to standardize the language, making it easier to label and classify using a transformer-based model. The translation was performed using the Deep Translator library. Standardizing the language also helps reduce inconsistencies that may arise from mixed-language inputs, which can negatively affect model performance. This step ensures that the model receives uniform input, improving the overall reliability of the classification process.

3.2. Pre-processing

After translation, the process continued with data pre-processing. This step involved preparing the text to ensure it was clean, consistent, and suitable for classification. Effective pre-processing is essential for enhancing model performance by reducing noise and standardizing input formats. Specific techniques and procedures used in this phase are detailed below.

a. Data cleansing

Data cleansing aims to remove irrelevant characters, including hashtags, URLs, usernames, question marks, and extra white spaces (see Table 2). This step is crucial to eliminate noise that could interfere with the model's ability to learn meaningful patterns. By refining the textual input, the data becomes more structured and easier to process in subsequent steps. Clean data contributes significantly to the accuracy and reliability of the classification results.

Table 2. Sample of data cleansing

Before	Prayers for Haiti 💔 💔 🙏
After	Prayers for Haiti

b. Case folding

Case folding converts all uppercase letters to lowercase (see Table 3). This normalization step helps ensure that words are treated uniformly, regardless of their original casing. It is particularly important in text classification tasks, as it reduces variability in the data without altering meaning. Consistent text casing improves the model's ability to recognize and learn patterns effectively.

Vol. 6, No. 5, October 2025, Page. 3419-3429
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4766

Table 3. Sample of case folding

Before	Prayers for Haiti
After	prayers for haiti

c. Stopwords removal

This step aims to remove words that carry little or no meaningful information. The process is performed using the NLTK library (see Table 4). Removing the stopwords helps reduce noise and focuses the analysis on more informative terms. This refinement enhances the model's ability to identify relevant features for classification.

Table 4. Sample of stopwords removal

Before	prayers for haiti
After	prayers haiti

d. Tokenizing

This step involves partitioning the text into small chunks called tokens (see Table 5). Tokenization is intended to facilitate further analysis by breaking down the text into manageable units. Each token typically represents a word or punctuation mark, enabling more precise processing in subsequent steps. This foundational process is essential for most natural language processing tasks.

Table 5. Sample of tokenizing

Before	prayers haiti
After	[“prayers”, “haiti]

3.3. Labeling

The labeling process began with data that has already undergone pre-processing. This data was then distributed to three different annotators. The selection criteria for the annotators were defined beforehand. Annotators were recruited through a crowdsourcing platform called Projects.co.id, where individuals could submit bids by showcasing their previous experience and proposed service fees. After a careful selection process, three annotators who met the specified criteria were chosen. Each annotator labeled the data independently. To determine the ground truth, majority voting was applied. In a small number of cases where all three annotators provided different labels, the data was manually labeled to resolve the tie.

3.4. Classification and Evaluation

This subsection presents the classification results along with an evaluation of model performance and runtime. The analysis considers both the effectiveness of each model in producing accurate results and the efficiency in terms of execution time. Differences in performance and runtime are examined to highlight how certain models may be more suitable for time-sensitive applications. Particular attention is given to the trade-off between accuracy and computational cost, which is essential when selecting models for operational environment.

3.4.1. Model performance

This sub-subsection presents a comparison of transformer-based models applied to analyse sentiment of social media data for disaster response. The comparison focuses on BERTBase, DistilBERT, and RoBERTa that were fine-tuned using annotated dataset [10]. The performance metrics include accuracy, precision, recall and weighted F1 to asses the model comprehensively [12]. The choice

Vol. 6, No. 5, October 2025, Page. 3419-3429
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4766

of these metrics ensures a proportional analysis in class imbalance. This is important where positive label is underrepresented and neutral label is overrepresented in data distribution.

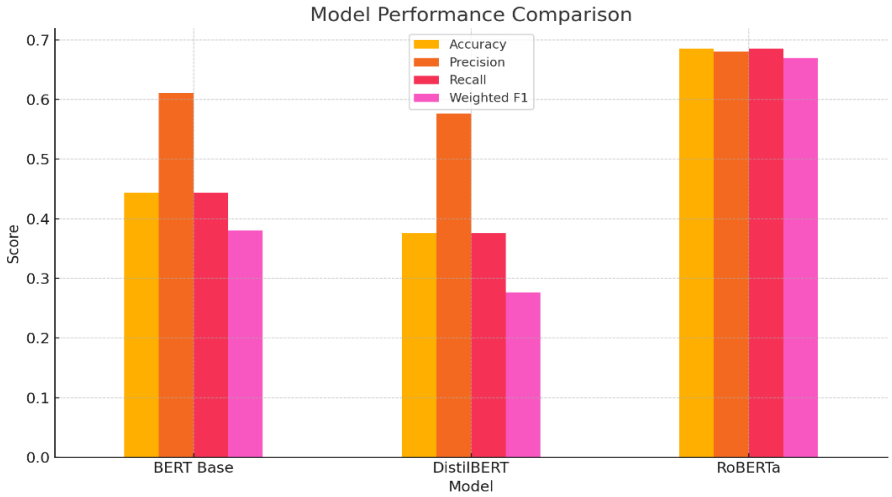

Figure 3. The comparison of model performance

Figure 3 illustrates a comparison across three transformed-based model. The performance metrics used in this comparison include accuracy, precision, recall and weighted F1. The use of weighted F1 is to address the problem of imbalance class in dataset. Among the three models, RoBERTa outperforms the others model. RoBERTa achieves highest score while maintaining minimal variation.

In contrast, both BERTBase and DistilBERT exhibit lower performance. Whole BERTBase demonstrates slight higher accuracy and weighted F1. However, the difference is not substantial. Both models achieve higher precision scores than their recall and F1-scores. This score indicates a tendency to be conservative in their prediction which means potentially minimizing false positives at the cost of increased false negatives. The results highlight RoBERTa's superior capability in handling the classification task in sentiment analysis.

3.4.2. Model runtime

In addition to performance metrics evaluation, it is also critical to assess computational efficiency for each model. In a time-sensitive application, for example disaster response, runtime performance determines how quickly a model can generate sentiment analysis classification. This subsection compares the transformer-based models. The runtime analysis provides insight into trade-offs between model performance and processing speed.

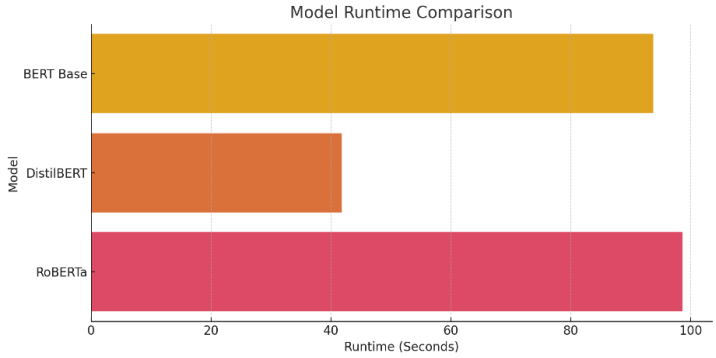

Figure 4. The comparison of model runtime in seconds

Vol. 6, No. 5, October 2025, Page. 3419-3429
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4766

Figure 4 illustrates the runtime performance of three transformer-based models. The comparison highlights the computational efficiency of each model during the evaluation phase. Among the three, DistilBERT demonstrates the shortest runtime, completing the task in just over 40 seconds. This is expected given that DistilBERT is a lightweight, distilled version of BERT designed to reduce computational overhead [13].

In contrast, BERTBase and RoBERTa exhibit significantly longer runtimes. Both models approaches about 90 seconds. Specifically, RoBERTa requires the highest execution time, close to 100 seconds, reflecting its more complex architecture and extensive pre-training. Although BERTBase is marginally faster than RoBERTa, it still requires more than twice the runtime of DistilBERT. These results indicate that the increased accuracy and robustness offered by RoBERTa, as shown in Figure 4, come at the cost of greater computational demands.

3.4.3. Model performance and runtime trade-off

Selecting appropriate for operational in disaster response involves more than evaluating classification prediction. In disaster response, decision-makers may also consider computational cost and speed of model execution. Therefore, understanding the trade-off between model performance and runtime is essential to balance accuracy and deployability. By analyzing these two aspects, the holistic perspective become clearer.

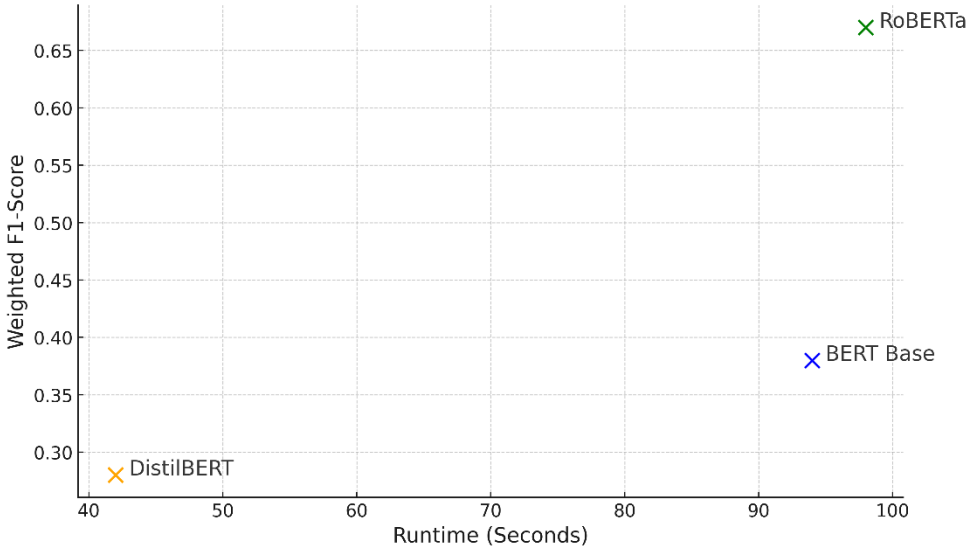

Figure 5. The trade-off between performance and runtime

Figure 5 demonstrates the trade-off between performance and efficiency. While RoBERTa leads in terms of predictive accuracy and balanced metric scores, its higher runtime may pose limitations in time-sensitive applications or environments with constrained computational resources. DistilBERT, on the other hand, offers the best runtime efficiency, making it a suitable option for real-time or large-scale processing scenarios where speed is a priority over marginal performance gains. These findings emphasize the importance of considering both accuracy and efficiency when selecting models for practical deployment. In the operational settings, RoBERTa is recommended to analyse sentiment on social media data while applying more holistic preprocessing technique.

4. DISCUSSIONS

This section discusses the impact of the results. The sentiment analysis of social media using transformer-based approaches has been widely applied in various domains, such as public health, consumer behavior, and disaster management [26][27][28]. The experimental results demonstrated that

Vol. 6, No. 5, October 2025, Page. 3419-3429
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4766

RoBERTa consistently achieved the highest scores across all classification metrics, confirming its robustness in handling disaster-related content. Its balanced performance in accuracy, recall, and weighted F1 shows the model's reliability in classifying social media posts that are often short, informal, and emotionally charged [29].

Compared to previous studies [2][4], our approach demonstrated improved model performance and faster runtime. Moreover, this study places greater emphasis on evaluating transformer-based models in natural language processing for sentiment analysis, using a dataset prepared and curated by the Joint Research Centre (JRC) of the European Commission. The use of a high-quality dataset provides a new perspective and contributes a valuable reference for the crisis informatics research community..

From a practical perspective, sentiment analysis is useful to understand public perception during a disaster. It enables authorities to monitor emotional responses and identify urgent needs in real time. For instance, the analysis can help map public sentiment geographically before distributing aid, improving resource targeting. Accurate classification plays an important role in this process, especially when distinguishing distress signals from neutral updates. Moreover, runtime is also critical when time becomes a decisive factor in disaster response, especially for early warning and triaging actions.

While RoBERTa offers the highest performance, it comes with a computational cost. In contrast, DistilBERT demonstrates superior runtime efficiency, completing inference in less than half the time of the other models. This trade-off highlights the importance of aligning model selection with operational needs. In large-scale or time-sensitive scenarios, DistilBERT could be prioritized, particularly when hardware resources are limited. BERTBase, sitting between both models, may serve as a viable compromise in environments that require both moderate accuracy and acceptable speed.

These results suggest that no single model fits all disaster contexts. The choice depends on the system requirements, including speed, accuracy, or scalability. For more critical use cases where accurate interpretation of public emotion can influence emergency strategies, RoBERTa remains the preferred option. However, the results could be improved by fine-tuning the models and applying more rigorous pre-processing [28]. Future research can explore model deployment in multilingual settings or using real-time data streams to further validate the adaptability of sentiment analysis in disaster response.

5. CONCLUSION

RoBERTa achieved highest performance across all performance metrics. This results indicate that RoBERTa was trained on similar dataset (Twitter sentiment analysis) which was closely related with the experiment dataset. RoBERTa also natively support three different classes (positive, negative, neutral) which was suited for the task. In terms of trade-off between performance metrics and running time, DistilBERT demonstrates the fastest model but has the worst performance. RoBERTa was the slowest but has the best performance.

As practical recommendation, RoBERTa is the best choice for analysing sentiment on social media data. The result could be improved by fine-tuning the model and adding more rigour pre-processing on dataset. If speed is critical, DistilBERT may be used. However, it requires fine-tuning for three class sentiments [9].

These findings highlight the importance of aligning model choice with the specific operational requirements of disaster response systems. In scenarios where accurate emotional interpretation is crucial for prioritizing aid and understanding public distress, RoBERTa provides the most reliable outcomes despite its longer execution time. Meanwhile, BERTBase may serve as an alternative for use cases requiring moderate trade-offs between speed and accuracy.

Future research could focus on extending this work by incorporating multimodal data sources, such as images or videos, to enhance situational awareness. Additionally, exploring multilingual

Vol. 6, No. 5, October 2025, Page. 3419-3429
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4766

sentiment models and domain-adaptive pre-training may further improve performance across diverse linguistic and cultural contexts in real-world disaster events. These directions can support more adaptive and scalable solutions for emergency response systems leveraging social media analysis.

REFERENCES

[1] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, and J. Kim, "COVIDSenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4,pp. 1003–1015, Aug. 2021. doi : 10.1109/TCSS.2021.3051189.

[2] G. Mu, J. Li, X. Li, C. Chen, X. Ju, and J. Dai, "An Enhanced IDBO-CNN-BiLSTM Model for Sentiment Analysis of Natural Disaster Tweets," *Biomimetics*, vol. 9, no. 9, p. 533, 2024, doi: 10.3390/biomimetics9090533.

[3] V. Lorini, E. Panizio, and C. Castillo, "SMDRM: A Platform to Analyze Social Media for Disaster Risk Management in Near Real Time," in *Workshop Proceedings 16th International AAAI Conferences on Web and Social Media (ICWSM)*, 2022. doi: 10.36190/2022.1.

[4] H. B. Firmansyah and A. Afriansyah, "Comparing Human and Machine-Labeled Sentiment Analysis for Disaster Response," in *Proceedings 12th Multidisciplinary International Social Networks Conference (MISNC)*, 2025.

[5] S. Z. Hassan, K. Ahmad, S. Hicks, P. Halvorsen, A. Al-Fuqaha, N. Conci, and M. Riegler, "Visual Sentiment Analysis from Disaster Images in Social Media," *Sensors*, vol. 22, no. 10, p. 3628, May 2022. doi: 10.3390/s22103628.

[6] S. Behl, A. Rao, S. Aggarwal, S. Chadha, and H. S. Pannu, "Twitter for disaster relief through sentiment analysis for COVID-19 and natural hazard crises," *International Journal of Disaster Risk Reduction*., vol. 55, p. 102101, Mar. 2021. doi: 10.1016/j.ijdrr.2021.102101.

[7] P. Y. W. Myint, S. L. Lo, and Y. Zhang, "Unveiling the dynamics of crisis events: Sentiment and emotion analysis via multi-task learning with attention mechanism and subject-based intent prediction," *Information Processing & Management*, vol. 61, no. 4, p. 103695, Jul. 2024. doi: 10.1016/j.ipm.2024.103695.

[8] N. Kankanamge, T. Yigitcanlar, A. Goonetilleke, and M. Kamruzzaman, "Determining disaster severity through social media analysis: Testing the methodology with South East Queensland Flood tweets," *International Journal of Disaster Risk Reduction*, vol. 42, p. 101360, Jan. 2020. doi: 10.1016/j.ijdrr.2019.101360.

[9] A. S. Hashim, N. Moorthy, A. A. Muazu, R. Wijaya, T. Purboyo, R. Latuconsina, C. Setianingsih, and M. F. Ruriawan, "Leveraging Social Media Sentiment Analysis for Enhanced Disaster Management: A Systematic Review and Future Research Agenda," *Journal of System and Management Sciences*, vol. 15, no. 4, pp. 170–191, 2025. doi: 10.33168/JSMS.2025.0412.

[10] A. Areshey and H. Mathkour, "Exploring transformer models for sentiment classification: A comparison of BERT, RoBERTa, ALBERT, DistilBERT, and XLNet," *Expert Systems*, vol. 41, no. 11, e13701, 2024. doi: 10.1111/exsy.13701.

[11] D. Contreras, S. Wilkinson, N. Balan, and P. James, "Assessing post-disaster recovery using sentiment analysis: The case of L'Aquila, Italy," *Earthquake Spectra*, vol. 38, no. 1, pp. 81–108, 2022. doi: 10.1177/87552930211036486.

[12] M. Bouazizi and T. Ohtsuki, "Multi-class sentiment analysis on Twitter: Classification performance and challenges," *Big Data Mining and Analytics*, vol. 2, no. 3, pp. 181–194, Sep. 2019. doi: 10.26599/BDMA.2019.9020002.

[13] A. R. Nair, R. P. Singh, D. Gupta, and P. Kumar, "Evaluating the Impact of Text Data Augmentation on Text Classification Tasks using DistilBERT," *Procedia Computer Science*, vol. 235, pp. 102–111, 2024. doi: 10.1016/j.procs.2024.05.102.

[14] H. B. Firmansyah, J. L. Fernandez-Marquez, and J. Cerquides, "Ensemble learning for the classification of social media data in disaster response," in *Proceedings 19th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, Tarbes, France, May 2022, pp. 710–718.

Vol. 6, No. 5, October 2025, Page. 3419-3429
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4766

[15] H. B. Firmansyah, V. Lorini, M. O. Mulayim, J. Gomes, and J. L. Fernandez-Marquez, "Improving social media geolocation for disaster response by using text from images and ChatGPT," in *Proceedings 11th Multidisciplinary International Social Networks Conf. (MISNC)*, Bali, Indonesia, Aug. 2024, pp. 67–72. doi: 10.1145/3675669.3675696.

[16] K. Seneviratne, M. Nadeeshani, S. Senaratne, and S. Perera, "Use of social media in disaster management: Challenges and strategies," *Sustainability*, vol. 16, no. 11, pp.1-21, Art. no. 4824, 2024. doi: 10.3390/su16114824.

[17] G. A. Ruz, P. A. Henríquez, and A. Mascareño, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers," *Future Generation Computer Systems*, vol. 106, pp.92-104, 2020, doi: 10.1016/j.future.2020.01.005.

[18] G. Airlangga, "Comparative Analysis of Machine Learning Models for Real-Time Disaster Tweet Classification: Enhancing Emergency Response with Social Media Analytics," *Brilliance Research of Artificial Intelligence*, vol. 4, no. 1, pp. 25–31, 2024, doi: 10.47709/brilliance.v4i1.3669.

[19] M. A. Nulhakim, Y. Sibaroni, K. Muhammad, and N. Ku, "Geospatial Sentiment Analysis Using Twitter Data on Natural Disasters in Indonesia with Support Vector Machine," *International Journal on Information and Communication Technology*,vol. 10, no. 2, pp. 242–258, 2024, doi: 10.21108/ijoict.v10i2.1032.

[20] S. Vijayarani, M. J. Ilamathi, and M. Nithya, "Preprocessing Techniques for Text Mining-An Overview Privacy Preserving Data Mining View project," *Journal of Computer Science & Computer Networks*, vol. 5, no. 1, pp. 7-16, 2015.

[21] N. R. Paul, C. Rakesh, and D. Sahoo, "Fine-Tuning Transformer-Based Representations in Active Learning for Labelling Crisis Dataset of Tweets," *SN Computer Science*, vol.4, p. 553,2023, doi : doi.org/10.1007/s42979-023-02061-z

[22] L. Peek, J. Tobin, R. M. Adams, H. Wu, and M. C. Mathews, "A Framework for Convergence Research in the Hazards and Disaster Field: The Natural Hazards Engineering Research Infrastructure Converge Facility," *Frontiers in Built Environment*, vol. 6, p. 110, 2020

[23] R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach," *Journal Informetrics*, vol. 3, no. 2, pp.143-157,2009, doi: 10.1016/j.joi.2009.01.003.

[24] Pang, B., & Lee, L. "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no 1–2, pp. 1–135, 2008

[25] M. Sari, S. Syahrullah, N. T. Lapatta, and R. Ardiansyah, "Twitter (X) Sentiment Analysis of Kampus Merdeka Program Using Support Vector Machine Algorithm and Selection Feature Chi-Square," *Journal Teknik Informatika (JUTIF)*, vol. 5, no. 5, pp. 1249–1256, Oct. 2024, doi: 10.52436/1.jutif.2024.5.5.2037.

[26] A. S. Talaat, "Sentiment analysis classification system using hybrid BERT models," Journal of Big Data, vol. 10, no. 110, pp. 1–18, 2023. doi: 10.1186/s40537-023-00781-w.

[27] R. Catelli, S. Pelosi, and M. Esposito, "Lexicon-Based vs. BERT-Based Sentiment Analysis: A Comparative Study in Italian," Electronics, vol. 11, no. 3, p. 374, Jan. 2022, doi: 10.3390/electronics11030374.

[28] A. Joshy and S. Sundar, "Analyzing the Performance of Sentiment Analysis using BERT, DistilBERT, and RoBERTa," in Proc. 2022 IEEE International Power and Renewable Energy Conference (IPRECON), 2022, pp. 1–6, doi: 10.1109/IPRECON55716.2022.10059542.

[29] A. Irianti, H. Halimah, S. Sutedi, and M. Agariana, "Integration of BERT and SVM in Sentiment Analysis of Twitter/X Regarding Constitutional Court Decision No. 60/PUU-XXII/2024," Journal Teknik Informatika (JUTIF), vol. 6, no. 2, pp. 469–482, Apr. 2025, doi: 10.52436/1.jutif.2025.6.2.4068