Natural Disaster Detection Through Crisis Mapping and Social Media Data

Mohamed Moubarak Mohamed Misbahou Mkouboi, Dr. Sabrina Tiun
Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 UKM Bangi
Selangor, Malaysia
p139575@siswa.ukm.edu.my, sabrinatiun@ukm.edu.my

***Abstract*:** *Human lives, infrastructure, and the global economies face natural disasters continuously which are the most significant risks to the world. It is hard to get information and local contextual information, with traditional disaster detection methods that mostly rely on official reports and sensors. For disaster events identification and understanding, social media platforms have multiplied, and user-generated content has become a significant supplementary data source. Improving disaster mapping techniques and natural disaster detection by using CrisisNLP dataset is the goal of this research. The system uses NLP techniques to classify and extract useful information from a large amount of unstructured social media data. Using labelled tweets related to disasters, the study builds and evaluates a variety of machine learning models, including Logistic Regression, Support Vector Machine (SVM), Random Forest, and Naive Bayes, as well as deep learning models, including BERT and LSTM. The performance of each model was evaluated using classification accuracy, with accuracy levels of 94% for BERT, 93% for LSTM, 88% for Logistic Regression, 92% for SVM, 95% for Random Forest, and 81% for Naive Bayes. The suggested framework shows the integration of social media and data-driven methods to improve disaster detection systems. These models help improve crisis management initiatives by providing insights to emergency responders and decision-makers.*

Keywords: *NLP, CrisisNLP, LSTM, BERT, Logistic Regression, SVM, Random Forest, Naïve Bayes, Disaster Tweet Classification*

I. INTRODUCTION

Recent increases in the frequency and intensity of natural disasters highlight the critical need for accurate and efficient disaster response systems. Social media platforms are now vital resources for information sharing and assistance coordination since they provide user-generated data. For example, studies have shown how Twitter data may be used to detect and track natural disasters, providing emergency response teams with useful information [1], [2].

Innovative approaches for evaluating data connected to disasters, including sentiment analysis and topic modelling, have been inspired by the vast amount and diversity of content found on social media [3], [4]. These methods use natural language processing (NLP) to efficiently and successfully find information related to disasters. Furthermore, researchers are now able to integrate temporal and spatial information into disaster response systems because of advances in social media analysis. For instance, locating disaster spots and monitoring their development have been accomplished by combining machine learning algorithms with social media data [5], [6].

Despite these developments, it is still difficult to separate useful data from the noisy social media landscape, especially in situations with multiple languages and different locations [7], [8]. To overcome these challenges, scalable machine learning models that can analyze massive datasets are needed in addition to strong data preparation techniques. According to recent research, crisis-specific datasets such as CrisisNLP [9] are crucial for training models that identify and categorize content linked to disasters [10], [11].

Furthermore, the availability of social media data during disasters has been greatly improved by the growing popularity of mobile devices and the expansion of internet connectivity. Instead of being passive recipients of information or victims, citizens now actively participate in reporting disasters and exchanging information. On social media platforms such as Facebook, Instagram, and Twitter, users can post updates, images, and videos. This generates a large amount of unstructured data that can provide deep insights on crises if handled appropriately. This development emphasizes the necessity of strong analytical frameworks for efficiently processing and utilizing this massive quantity of user-generated content.

When combined with machine learning methods, disaster mapping has the potential to completely transform disaster management. Responders can use it to better allocate resources, prioritize places that need immediate assistance. The transition from reactive to proactive disaster management shows how data science and technology can greatly improve community resilience and reduce the damage caused by disasters.

The objective of this project is to build a context-aware system that uses social media data for crisis mapping and natural disaster detection. This study adds to the expanding body of knowledge in disaster informatics and NLP by improving upon current approaches and resolving existing constraints, opening the door to more efficient disaster response systems.

II. LITERATURE REVIEW

This section reviews the literature on disaster detection, applications, and challenges of using social media data in crisis management, emphasizing significant developments, approaches, and gaps.

A. Disaster Detection Using Social Media

Social media websites, especially Twitter, have developed into vital resources for disaster detection since they offer real-time information on developing emergencies. Researchers have investigated a range of machine learning and deep learning techniques to efficiently classify tweets related to disasters. In their evaluation of neural network models using word embeddings, for example, ALRashdi and O'Keefe showed the effectiveness of Bi-LSTM on the CrisisNLP dataset, obtaining an F1 score of 62.04 [10]. Similarly, to cluster tweets relevant to disasters, Krishna et al. used a hybrid fusion model that combined SVM, Naive Bayes, and Random Forest [2].

Tweet classification has been further enhanced by advanced methods that integrate NLP and entity masking. Seeberger and Riedhammer proposed a multi-task learning approach using entity-masked language modelling, which resulted in a significant F1 score improvement for actionable tweet types [11]. Shetty et al. demonstrated that multimodal approaches combine text and image data to provide comprehensive disaster assessments [12]. Adili and Chen discussed active learning methods that improve detection frameworks by addressing issues such as resource constraints and unstructured data [13].

Another crucial component of disaster detection is sentiment analysis. An automated method that combines NER, sentiment analysis, and anomaly detection was proposed by Sufi and Khalil to monitor worldwide disasters with an accuracy of 97% [14]. Behl et al. evaluated sentiment trends during the COVID-19 and natural hazard crises using MLP models with TF-IDF [3]. Likewise, Ruz et al. used Bayesian networks to categorize sentiments from tweets on earthquakes, and they were able to achieve excellent accuracy using SVM [15]. In their additional investigation of visual sentiment analysis from disaster images, Hassan et al. demonstrated the efficacy of a sentiment analyzer based on deep learning [16].

Paul et al. introduced a deep learning system combining CNN and GRU for crisis-related data classification, providing strong situational awareness [17]. Madichetty and Sridevi introduced a neural-based technique using the RoBERTa model to recognize situational information from tweets during disasters, surpassing classic CNN and LSTM models [8]. Kosugi et al. used Twitter to create an early disaster information-sharing system that allowed for real-time engagement [1]. Park and Cho improved domain-specific spam tweets by adding company-relevant factual knowledge to pre-trained language models [18]. This method can be used to improve tweet filtering related to disasters.

These papers demonstrate data-related problems including sarcasm detection and annotation biases, as well as NLP techniques such as TF-IDF, Bayesian networks, and RoBERTa that are relevant to the study of social media data for crisis mapping from a methodological approach.

B. Social Media in Disaster Response

Social media data is essential for giving information about disasters in a context. Using topic modelling and crisis categorization techniques, Kumar et al. examined social media posts to address issues with the healthcare supply chain during the COVID-19 pandemic [19]. Milusheva et al. improved urban planning by using geolocation algorithms, verifying traffic crashes from Twitter data with 92% accuracy [6].

There has been many research done on the geographical behaviour of disaster impacted populations. González used Twitter data to study hurricane evacuation compliance, providing dynamic insights into population movements [5]. Karimiziarani et al. created a Hazard Risk Awareness (HRA) index using Twitter data during Hurricane Harvey to map regional awareness levels, increasing situational analysis [7]. Integration of GIS with AI has also gained attention. The significance of GIS in disaster management was emphasized by Abid et al., who demonstrated its use in earthquake and flood scenarios [20]. Similarly, Cao et al. investigated the integration of BIM and GIS for urban disaster management, focusing on interoperability issues and potential future directions [21].

There are still difficulties despite these developments. The limits of geotagged data in disaster classification and location prediction, especially for underrepresented regions, were highlighted [22]. In their analysis of multilingual social media data for natural disaster management, Nurdin et al. used deep learning algorithms, highlighting the need for noise reduction during tweet preprocessing [4]. These studies show how crucial it is to improve disaster response capabilities by fusing innovative AI approaches with social media data.

These studies offer insightful information about methods and challenges related to using geotagged data for disaster detection.

C. Applications for Crisis Management

For disaster management, systems are essential for handling massive amounts of dynamic data. Using Twitter, Kosugi et al. created an easy-to-use disaster information-sharing system that makes data sharing during emergencies easier [1]. A big data analytics approach for Twitter trend analysis was proposed by Rodrigues et al., which successfully identified new disaster trends [23].

Applications monitoring has been further improved with interactive dashboards and streaming APIs. Alabdulaali et al. created a multimodal dashboard that combines collaboration boards and sentiment analysis to offer a thorough understanding of crisis scenarios [24]. By providing dynamic visualizations of public sentiment during crises, Bahrawi highlighted the value of Twitter's Streaming API for real-time sentiment analysis [25]. Nielsen et al. proposed a disaster social media ontology, recognizing the value of user feedback in improving real-time disaster systems [26].

One important development in crisis management is the use of conversational dashboards. A dashboard with natural language queries was established by Ruoff et al., increasing user efficiency in accessing disaster information [27]. In their assessment of machine learning applications for emergency response, Dwarakanath et al. pointed out that issues such as processing overhead and scalability still exist [28].

This studies focuses on using social media data for disaster information sharing and analysis. Along with its benefits and drawbacks, including data overload and challenges with multilingual support, it emphasizes techniques like topic modeling, multimodal visualizations, and dashboards based on natural language.

D. Challenges in Using Social Media for Disaster Detection

There are advantages and disadvantages to using social media for disaster management, especially when it comes to handling ethical and privacy issues. Karimiziarani researched the ethical implications of social media analytics, focusing particular attention to privacy concerns and the possibility of spreading misleading information [29]. Similarly, Lovari and Bowen emphasized the necessity of ethical standards to battle false information in their research on methods of communication during flood disasters [30]. Paul et al. emphasized how social media-based disaster monitoring systems can be made more dependable by using ethical methods of communication [17].

Applications of crowdsourcing in disaster management also confront several obstacles. Nielsen et al. discovered gaps in the literature on disaster risk management, especially with reference to the Global South's sociocultural factors [26]. Zhang et al. emphasized in their roadmap for intelligent public warning systems the significance of ethical and open data norms [31].

Even though there have been significant developments in disaster detection, applications, scalability, ethical issues, and data quality continue to be problems. Addressing these problems is essential to creating creative solutions that successfully incorporate social media into frameworks for disaster detection. The knowledge gained from this evaluation offers a strong basis for addressing in these gaps in future research.

The papers mentioned in this part focused on social media analytics for disaster management. It showcases approaches such as case study analysis [30], hybrid deep learning [17], and machine learning technique reviews [29]. Enhanced situational awareness and the recognition of ethical issues are important results, but there are drawbacks as well, such as gaps in low-resource applications and a lack of integration tools.

III. RESEARCH OBJECTIVES

1. To develop and identify models for dynamic crisis mapping that uses social media data.
2. To design and develop a scalable system for dynamic crisis mapping that uses social media data.
3. To evaluate the performance of the proposed methodology in improving disaster detection.

IV. RESEARCH QUESTIONS

- How can social media data be effectively leveraged for disaster detection?
- What are the key challenges in integrating social media data into crisis mapping framework?
- How does the proposed system for social media data compare with traditional methods in terms of accuracy?

V. METHODOLOGY

This section begins with a summary of the research concept and then goes into detail about the datasets used, including how they were prepared and annotated. It introduces the planned machine learning components; the specific training and evaluation will be covered in the result and discussion section.

The development and optimization of the classification model are the main objectives of the machine learning (ML) and deep learning (DL) training phase. Models such as BERT and LSTM are chosen because of their demonstrated ability to process domain-specific, context-rich text [32], [33], [34], once the CrisisNLP dataset is split into testing, validation, and training. Furthermore, Logistic Regression, SVM, Random Forest and Naïve Bayes are used because of its computational effectiveness and robustness in handling the informal and noisy text that is frequently present in social media, offering a supplementary method for disaster tweet classification [35]. The goal of the training method is to classify tweets about disasters while defining desired results, like identifying relevant content as shown in Fig. 1.

Lastly, the evaluation process determines how well the model works and how applicable it is to real-world scenarios. The model's performance is evaluated using metrics such as accuracy, precision, recall, and F1-score, and its practicality is ensured by testing with half of the data from the CrisisNLP dataset.

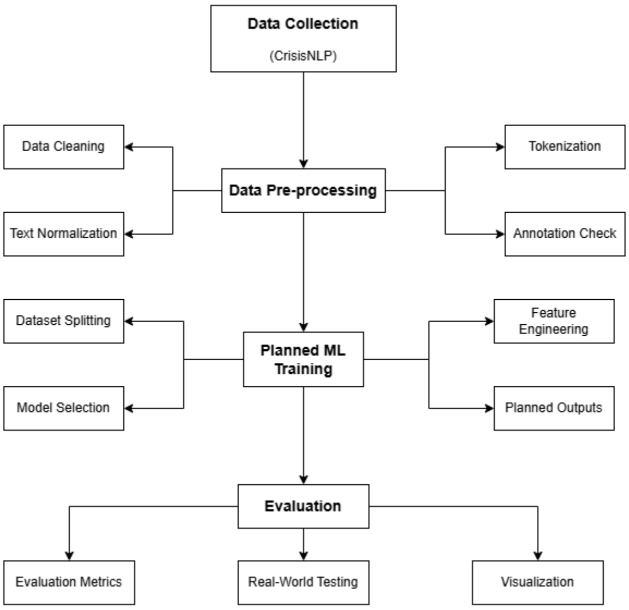


Fig. 1. *Flowchart of Overall Methodology*

A. *Dataset Collection*

The CrisisNLP dataset is the source of data used in this research [36]. Since this project focuses on natural disasters, which includes annotated social media data regarding the previously mentioned incidents, our primary focus will be earthquakes, typhoons, volcanoes, floods, and landslides. This dataset provides tagged instances for information classification and disaster detection tasks. The half of this dataset will be used for evaluation and analysis of the models.

CrisisNLP Dataset includes data with labels from several crisis events. Every subdirectory stands for a distinct disaster and the details are shown in Table I.

TABLE I. DATASETS DETAILS

Crisis Type	Crisis Name	Country	Language	Number of Tweets	Year
Earthquake	Nepal Earthquake	Nepal	English	4,223,937	2015
Earthquake	Chile Earthquake	Chile	English	368,630	2014
Earthquake	California Earthquake	USA	English	254,525	2014
Earthquake	Pakistan Earthquake	Pakistan	English	156,905	2013

Typhoon	Cyclone PAM	Vanuatu	English	490,402	2015
Typhoon	Typhoon Hagupit	Philippines	English	625,976	2014
Typhoon	Hurricane Odile	Mexico	English	62,058	2014
Volcano	Iceland Volcano	Iceland	English	83,470	2014
Floods	Pakistan Floods	Pakistan	English	1,236,610	2014
Floods	India Floods	India	English	5,259,681	2014
Landslide	Landslides worldwide	Worldwide	English	382,626	2014

B. *Preprocessing*

Preprocessing ensures that the dataset is clean and ready to be used in subsequent processes. The procedures consist of:

- Lowercasing and Cleaning: Converted text to lowercase; removed URLs, punctuation, special characters, and emojis.
- Stopword Removal: Filtered out common English stopwords to reduce noise.
- CrisisNLP Normalization: Replaced slang and abbreviations using the CrisisNLP OOV dictionary.
- Tokenization:
  - Traditional Models/LSTM: Used word-level tokenization.
  - BERT: Used subword tokenization via HuggingFace's AutoTokenizer.
- Class Balancing: Applied oversampling and undersampling and class weights to handle label imbalance.
- Label Encoding: Converted string labels to integers for model compatibility.
- Vectorization:
  - TF-IDF: For traditional models like SVM, Naive Bayes, etc.
  - Sequence Padding: For LSTM models using Keras preprocessing.
  - BERT Input Formatting: Token IDs, masks, and padding handled via transformer tokenizer.

C. *Annotation*

The pre-annotated data in the CrisisNLP dataset has detailed labels that are very important to disaster detection. These labels consist of:

- Injured or dead people: Reports of casualties and/or injured individuals.
- Missing, trapped, or found people: Information about missing or located individuals.
- Displaced people and evacuations: Posts related to evacuations and temporary relocations.
- Infrastructure and utilities damage: Reports of damage to buildings, roads, and essential services.
- Donation needs, offers, or volunteering services: Requests or offers for aid, such as food, water, shelter, and medical supplies.
- Caution and advice: Warnings, guidance, and safety recommendations.
- Sympathy and emotional support: Messages expressing prayers, condolences, and emotional support.
- Other useful information: Additional relevant details that provide insight into the disaster situation.
- Not related or irrelevant: Posts that do not contain relevant disaster-related information.

These annotations will be essential for training and validating the suggested models in subsequent chapters.

To further analyze the dataset structure, Fig. 2 presents a word frequency distribution of the annotated labels, highlighting their prevalence within the dataset.

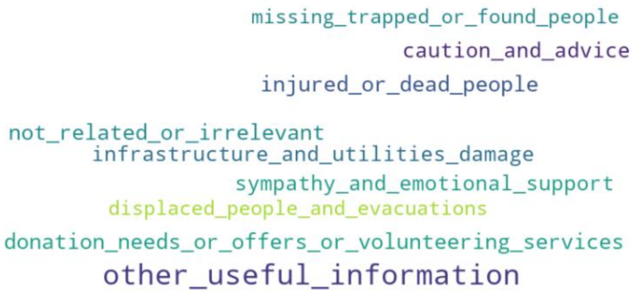


Fig. 2. *Word Cloud Across All Labelled Crisis Data*

D. *Model Development And Inference*

This section offers a summary of the planned approach, while subsequent sections will address the specific implementation of machine learning models.

1) *Model Architecture:*

To provide a reliable method for tweet classification linked to disasters, the suggested architecture makes use of transformer-based models, such as LSTM, BERT, SVM, Logistic Regression, Random Forest and Naïve Bayes. To comprehend the complex language frequently seen in social media posts on disasters, BERT was selected because of their demonstrated capacity to capture deep, bidirectional contextual connection to text [32], [34]. They can efficiently model complex semantic patterns and fine-tune to domain-specific tasks because of their considerable pre-training on big datasets. LSTM, on the other hand, provides an accessible and effective alternative that performs exceptionally well when processing noisy and informal language data, which is typical in social media [33]. Combining those models allows the system to achieve the best possible balance between computational efficiency and performance by using both the speed and convenience of use of LSTM and its deep contextual understanding of transformers.

2) Integration & Evaluation of the Crisis Mapping Framework:

This project implements a multi-model ensemble system for text classification on disaster-related tweets. The system supports various machine learning and deep learning architectures and allows users to input custom texts or sample randomly from a dataset.

a) Web-Based Platform Integration:

The model will be incorporated into a web-based platform to process new tweets, classify them, and display the outcomes as shown in Fig. 3.

The framework will:

- Data Ingestion: Collect tweets using the half of the CrisisNLP dataset.
- Processing: Use the trained models to automatically classify tweets.
- Visualization: Give users the ability to quickly evaluate crisis disasters by displaying the models' accuracies score and confidence on disaster-related in a framework.

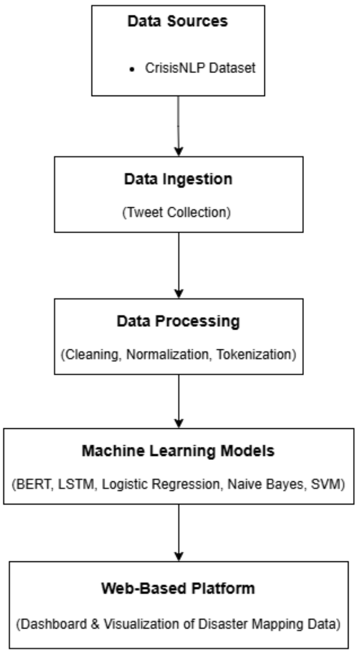

Data Sources
• CrisisNLP Dataset
Data Ingestion
(Tweet Collection)
Data Processing
(Cleaning, Normalization, Tokenization)
Machine Learning Models
(BERT, LSTM, Logistic Regression, Naive Bayes, SVM)
Web-Based Platform
(Dashboard & Visualization of Disaster Mapping Data)

Fig. 3. *Integration Schema for the Web-Based Platform*

b) Models Used:

- LSTM (Keras): Using pre-processed input padded to a fixed length of 100 tokens, a deep learning model was trained on tokenized tweet sequences.
- BERT (Transformer): Optimized for sequence classification, a fine-tuned BERT model was loaded via Hugging Face's transformers library.
- Traditional ML Models trained on TF-IDF features:
  - Support Vector Machine (SVM)
  - Random Forest (RF)
  - Logistic Regression (LR)
  - Naive Bayes (NB)

Using joblib and native model-saving formats in keras and Hugging Face directories, each model was trained and serialized.

c) Text Input & Preprocessing:

- Input can be:
  - Typed or pasted text directly.
  - A randomly sampled tweet from the test dataset (data/disaster_tweets.csv).
- Inputs are fed into each model using the appropriate preprocessing:
  - LSTM: Tokenization using Padded sequence saved in Keras tokenizer.
  - Traditional models: A pre-saved TfidfVectorizer was used to vectorize.
  - BERT: Padded and truncated to max 128 tokens using tokenized with AutoTokenizer method.

d) Prediction Pipeline:

Each model returns a Predicted class label and Confidence score. However, SVM uses softmax(decision_function) for approximate confidence scores because it does not originally output probabilities.

e) User Interface:

- Built using Streamlit.
- A button click launches the prediction.
- Output includes:
  - Model-by-model prediction summary.
  - Bar chart displaying class prediction consensus.
  - Bar chart displaying model confidence scores.

E. EVALUATION METRICS

The evaluations will concentrate on metrics such as:

- Accuracy: The percentage of tweets that are accurately classified.
- Precision and Recall: To assess how complete and relevant the classifications are.
- F1-Score: To find a balance between recall and precision.

Table II shows the utility of the metrics in this project.

TABLE II. SUMMARIZATION OF THE METRICS SPECIFICATIONS

Metric	Definition	Relevance to the Study
Accuracy	The percentage of cases that were correctly classified to all instances.	Shows how well the algorithm performed overall in classifying tweets about disasters.
Precision	The percentage of predicted positives that are actually	Reduces false positives in the model, which is essential for preventing irrelevant data from

	positive.	being misclassified.
Recall (Sensitivity)	The percentage of actual positives to true positive predictions.	Evaluates the model's capacity to find all relevant tweets about disasters, making sure that important information is not overlooked.
F1-Score	The harmonic means of recall and precision.	Combines recall and precision to offer one metric to evaluate the model's overall classification reliability.

VI. RESULTS & DISCUSSION

This section presents and analyzes the results of classifying social media posts into disaster-relevant categories. By automatically classifying tweets based on the type of information they convey ranging from infrastructure damage and missing persons reports to calls for donations and expressions of sympathy, this research aims to improve situational awareness during crises. A collection of annotated tweets from multiple historical disaster events from CrisisNLP dataset were trained and evaluated on several machine learning and deep learning models to achieve this.

Additionally, the visual exploration of the dataset through techniques such as word clouds, label distribution plots, and tweet length distributions, both before and after data balancing took account of the characteristics of the dataset. The assessment of each model's learning dynamics and generalization capabilities are analyzed using the training and validation curves visualizations.

Through this comparative evaluation, it highlights the strengths and limitations of each model technique, offering valuable guidance for selecting the most suitable model in real-world crisis response scenarios. Ultimately, contributing to more effective decision-making in disaster management will bridge the gap between technical implementation and practical application.

A. Dataset Overview

The dataset consists of labelled tweets collected from the CrisisNLP dataset, having several types of natural disasters such as earthquakes, floods, typhoons, and more. The dataset contained approximately 25,000 tweets labelled into the nine classes that have been mentioned in the Methodology section.

B. Traditional Models

Accuracy, Precision (macro average), Recall (macro average), and F1-Score (macro average) are the evaluation metrics that represent each model performance.

1) Logistic Regression (LR):

LR is a simple model used in NLP for classifying text, such as determining which label belongs to which class in a classification problem [35]. For many NLP tasks it makes a good starting point because it predicts the probability of a text belonging to a class.

2) Support Vector Machine (SVM):

Finding the best boundary to separate different text categories such as positive and negative, SVM comes in handy as it is a machine learning algorithm used in NLP for tasks including text classification [35]. It's particularly good at handling high-dimensional text, and complex data.

3) Naïve Bayes (NB):

For text classification tasks such as spam detection NB is often used as it is a fast, probabilistic model for NLP [35]. Even with limited data, it assumes words in text are independent which making it efficient and simplifies the process.

4) Random Forest (RF):

RF is an ensemble method that improves accuracy and reduces errors in NLP, and it uses multiple decision trees to classify text [35]. It is robust and effective at tasks similar to news article classification.

The different performances of the traditional models are shown in Table III.

TABLE III. PERFORMANCE OF TRADITIONAL MODELS

Model	Accuracy	Precision (macro avg)	Recall (macro avg)	F1-Score (macro avg)
LR	0.88	0.87	0.88	0.87
SVM	0.92	0.91	0.92	0.91
NB	0.81	0.80	0.81	0.80
RF	0.95	0.94	0.95	0.94

Random Forest demonstrates the highest performance across all metrics among the models, achieving an accuracy of 0.95, precision of 0.94, recall of 0.95, and an F1-score of 0.94. With an accuracy of 0.92, precision of 0.91, recall of 0.92, and F1-score of 0.91, SVM follows closely behind indicating strong and consistent performance. With an accuracy of 0.88, precision of 0.87, recall of 0.88, and F1-score of 0.87, Logistic Regression shows moderate results, while Naïve Bayes yields the lowest scores, with accuracy at 0.81, precision at 0.80, recall at 0.81, and F1-score at 0.80. While Naïve Bayes appears to be the least effective for this task, the results highlight Random Forest as the top-performing model followed by SVM.

C. LSTM Model

LSTM is a type of neural network used for tasks where the order of words matters in NLP. It remembers long-term patterns in text, making it suitable for understanding sequences [8]. It was trained on tokenized tweet sequences. It used embedding layers and handled sequential dependencies effectively.

In Table IV, the performance of the LSTM model is good enough to continue with the testing.

TABLE IV. PERFORMANCE OF LSTM MODEL

Model	Accuracy	Precision (macro avg)	Recall (macro avg)	F1-Score (macro avg)
LSTM	0.93	0.93	0.93	0.93

With identical scores across precision, recall, and F1-score, the model achieves an accuracy of 0.93, indicating a well-balanced and consistent performance. These results suggest that the LSTM model effectively classifies data with high reliability and minimal bias toward any specific class, making it a strong candidate for tasks requiring accurate sequence modeling.

D. BERT Model

BERT is an advanced model in NLP called a “transformer-based model” that excelling in tasks such as question answering and sentiment analysis because it understands text by considering both left and right context [34]. It's pre-trained on large text data, making it powerful for complex NLP tasks. It was trained separately using the HuggingFace framework. It was fine-tuned for multi-class classification.

The model did a good performance which gives a promising prediction rate on unseen data. Table V shows the results of the BERT model.

TABLE V. PERFORMANCE OF BERT MODEL

Model	Accuracy	Precision (macro avg)	Recall (macro avg)	F1-Score (macro avg)
BERT	0.94	0.93	0.94	0.94

With nearly identical scores of 0.93 in precision, 0.94 of recall, and 0.94 of F1-score, the model achieves an accuracy of 0.94, demonstrating a highly balanced and consistent performance across all metrics. Making it a reliable choice for text-based machine learning applications, these results indicate that the BERT model effectively handles classification tasks with strong generalization and minimal class bias.

E. Comparative Evaluation

Their performances were summarized using two key evaluation metrics, accuracy and F1-score, to systematically compare the effectiveness of the models. Due to their relevance in multi-class classification tasks, these metrics were selected. While the macro-averaged F1-score captures how well the model performs across all categories, accuracy offers a general measure of overall correctness, as shown in Table VI.

TABLE VI. MODEL COMPARISON

Model	Accuracy	F1-Score (macro avg)	Notes
LR	0.88	0.87	Weak on nuanced or noisy classes
SVM	0.92	0.91	Strong baseline
NB	0.81	0.80	Weakest traditional model
RF	0.95	0.94	Top-performing non-deep learning model
LSTM	0.93	0.87	Good sequential learning
BERT	0.94	0.91	Best contextual understanding

Random Forest achieves the highest accuracy of 0.95 and F1-score 0.94 among the models, earning its title as the top-performing non-deep learning model. With an accuracy of 0.94 and F1-score of 0.91 BERT follows closely, known for its superior contextual understanding in text classification. With 0.92 accuracy and 0.91 F1-score SVM performs well as a strong baseline, while with 0.93 accuracy but a notably lower F1-score of 0.87 LSTM shows solid learning capabilities suggesting inconsistent performance across classes. Logistic Regression demonstrates moderate performance, and Naïve Bayes underperforms compared to other traditional models. Therefore, Naïve Bayes are the weakest models among the traditional and all the models in this research.

F. Data and Model Analysis

To better understand the challenges and dynamics of disaster-related tweet classification, this section provides a visual and statistical exploration of the dataset and model behavior.

1) Word Cloud of Most Common Words in Tweets:

To visualize the most frequent terms appearing in the tweets, a word cloud was generated, as shown in Fig. 4.

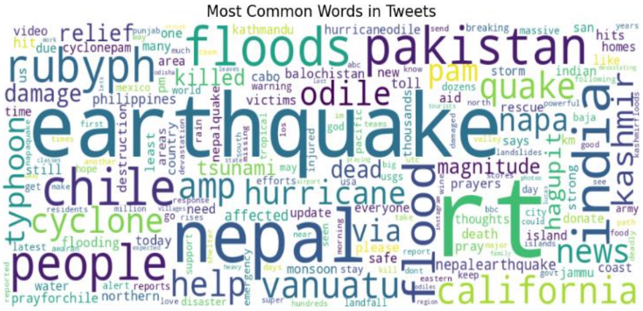


Fig. 4. *Most Common Words in Tweets*

This word cloud shows a strong emphasis on natural disasters, their human impact, and associated humanitarian responses and offers a quick and effective overview of the themes and concerns present in disaster-related tweets. It serves as a useful tool for understanding public discourse during crisis events.

2) Label Distribution Before and After Data Balancing:

Fig. 5 displays the frequency of different labels in the dataset before any data balancing is applied. Here, the counts vary significantly across categories. For instance, while others such as “Missing trapped or found people”, and “Displaced people and evacuation” have lower than 500 counts, some labels like "Other useful information" and "Not related or irrelevant" have very high counts which is around 2500 counts. This unequal distribution shows the imbalance in the original dataset, where certain categories dominate while others are underrepresented. If not addressed through data balancing techniques, such differences can lead to biased model performance.

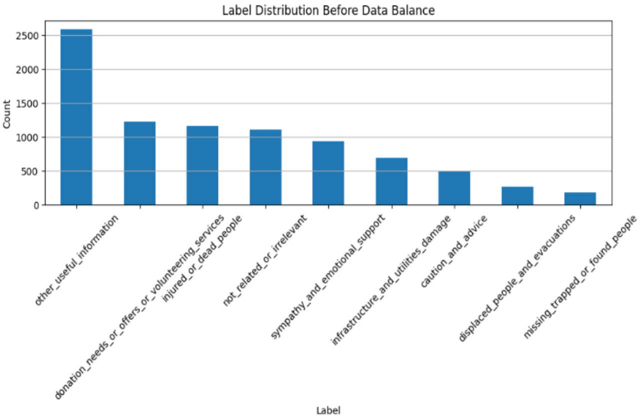


Fig. 5. *Label Distribution Before Data Balance*

The bar chart (Fig. 6) shows the frequency of different labels in a dataset after applying data balancing techniques. Each label represents a specific category, such as "Displaced people and evacuations," "Not related or irrelevant," and others. All labels have nearly equal counts, with most hovering around 2500 instances after balancing. Ensuring that each category is represented equally, this uniform distribution indicates that data balancing has successfully addressed any imbalances in the original dataset. To avoid bias towards more frequent classes, the consistent heights of the bars show a well-balanced dataset which is important for training machine learning models.

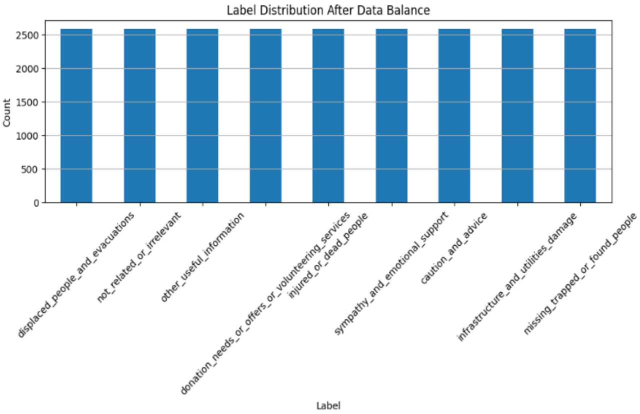


Fig. 6. *Label Distribution After Data Balance*

3) Tweet Length Distribution (Token Count) Before and After Data Balancing:

Fig. 7 shows the original distribution of tweet lengths. Most tweets cluster around 10 tokens, but the distribution is inconsistent, with noticeable peaks and dips. The smooth curve indicates some skewness or imbalance, and extreme values are more common. This shows the data may be biased toward certain tweet lengths before balancing.

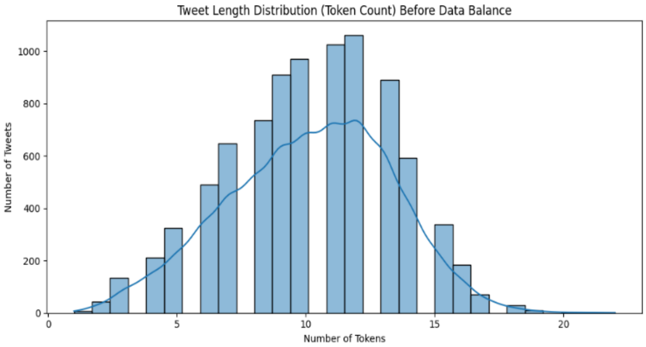


Fig. 7. *Tweet Length Distribution Before Data Balancing*

Fig. 8 shows smoother and more uniform distribution. The overall shape is more symmetrical with fewer extremes, while the peak remains around 10 tokens. Data balancing has reduced variability, leading to a more even representation of tweet lengths. For models such as LSTM and BERT that are sensitive to sequence length, this balanced distribution is preferable for fair and accurate analysis and model training.

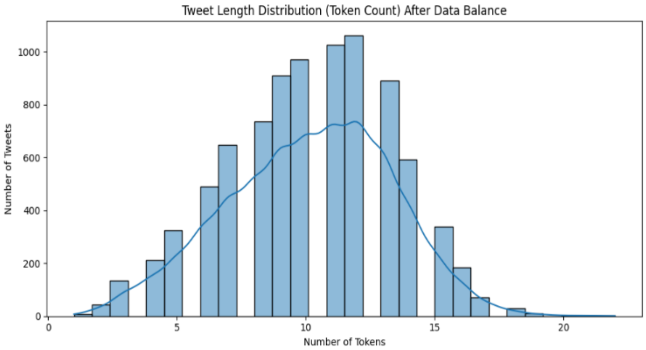


Fig. 8. *Tweet Length Distribution After Data Balancing*

4) Model Accuracy Comparison (Traditional Models):

Fig. 9 compares the performance of traditional models.

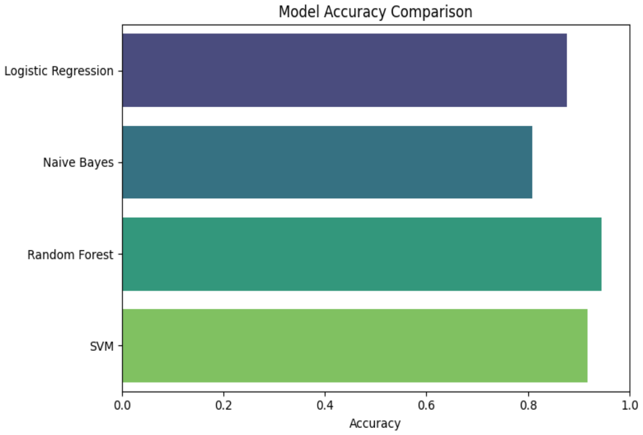


Fig. 9. *Model Accuracy Comparison (Traditional Models)*

This bar chart provides a comparison of the accuracy of four machine learning models. Random Forest emerges as the top-performing model, followed by SVM, Logistic Regression, and Naïve Bayes. Visual representation makes it easy to identify which models are more suitable for the given task based on their accuracy scores.

5) LSTM Training & Validation Loss / Accuracy:

The loss graph (Fig. 10) shows a decrease in prediction errors for both training and validation data during training. Training loss drops quickly from around 1.2 to nearly 0.2, showing strong initial learning. Validation loss also decreases but stops earlier and slightly increases near the end, ranging around 0.3. This pattern shows that the model may begin to overfit slightly as training progresses, while the model is learning well from the training data, since the validation loss does not improve further and even worsens a bit.

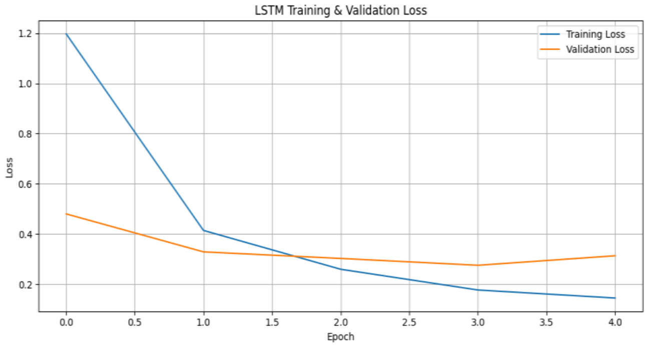


Fig. 10. *LSTM Training & Validation Loss*

The accuracy graph (Fig. 11) shows how the LSTM model's performance improves over epochs on both the training and validation datasets. While validation accuracy begins at about 0.85 and reaches a peak of approximately 0.92 before stabilizing, the training accuracy starts around 0.60 and increases steadily to nearly 0.95. The consistent rise in both curves indicates that the model is learning effectively. Although training accuracy remains slightly higher than validation accuracy, the gap is small, suggesting minimal overfitting.

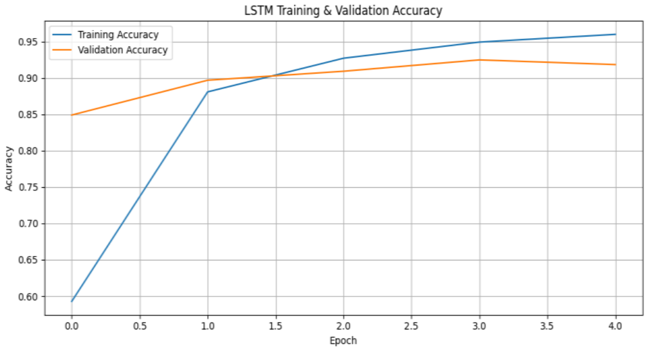


Fig. 11. *LSTM Training & Validation Accuracy*

6) BERT Training & Validation Loss / Validation Accuracy:

Fig. 12 illustrates both the training loss and validation loss of the BERT model across epochs. The validation loss starts slightly higher around 0.4 and remains relatively stable, ranging between 0.3 and 0.4 throughout training, the training loss begins at about 0.9 and decreases sharply in the early epoch, eventually stabilizing near 0.1 by the third epoch. In contrast. The significant gap between training and validation losses shows that the model may be overfitting to some extent, as the training loss drops much more than the validation loss.

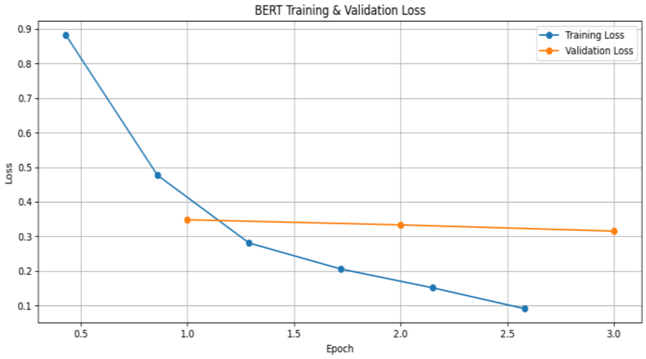


Fig. 12. *BERT Training & Validation Loss*

Fig. 13 shows the validation accuracy of a BERT model over epochs. As training progresses, the validation accuracy starts at approximately 0.90 and increases steadily. It reaches a peak value of around 0.935 by the end of the third epoch. The smooth rise trend indicates that the model is consistently improving its performance on unseen data, showing effective learning without signs of overfitting during this training.

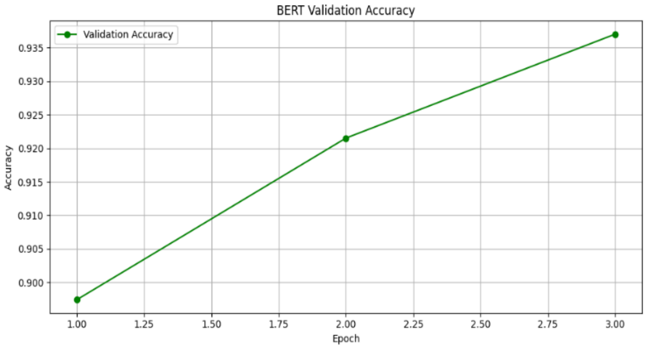
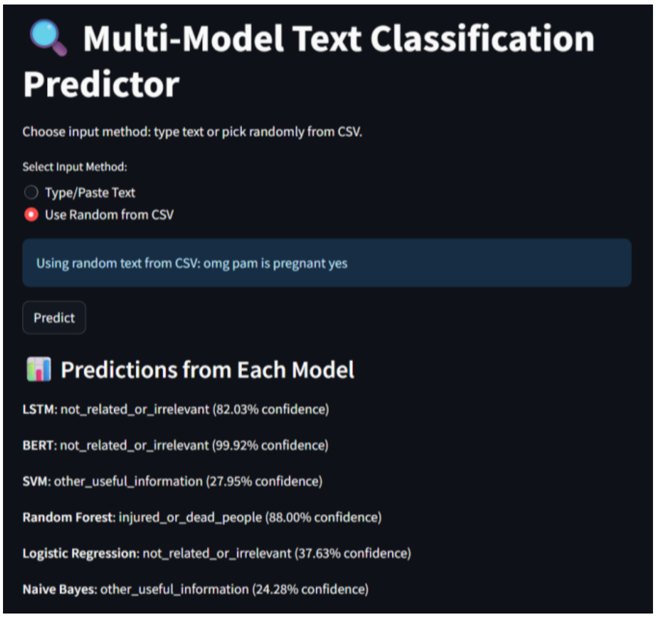


Fig. 13. *BERT Validation Accuracy*

G. Prediction Comparison using the Interface

Each model outputs its predicted label along with a confidence score. This setup allows a side-by-side evaluation of how different models interpret the same input. It often indicates high confidence when predictions align across models. In contrast, disagreement can indicate ambiguous or borderline cases, especially in noisy disaster-related texts (Fig. 14).



Multi-Model Text Classification Predictor

Choose input method: type text or pick randomly from CSV.

Select Input Method:

○ Type/Paste Text

● Use Random from CSV

Using random text from CSV: omg pam is pregnant yes

Predict

Predictions from Each Model

LSTM: not_related_or_irrelevant (82.03% confidence)

BERT: not_related_or_irrelevant (99.92% confidence)

SVM: other_useful_information (27.95% confidence)

Random Forest: injured_or_dead_people (88.00% confidence)

Logistic Regression: not_related_or_irrelevant (37.63% confidence)

Naive Bayes: other_useful_information (24.28% confidence)

Fig. 14. *Comparison of Model Prediction and Confidence*

H. Sample Output Example using the Interface

Two primary visualizations provide quick interpretability:

- Consensus Chart: Highlights how many models predicted each class, making it easy to spot dominant predictions, as shown in Fig. 15.

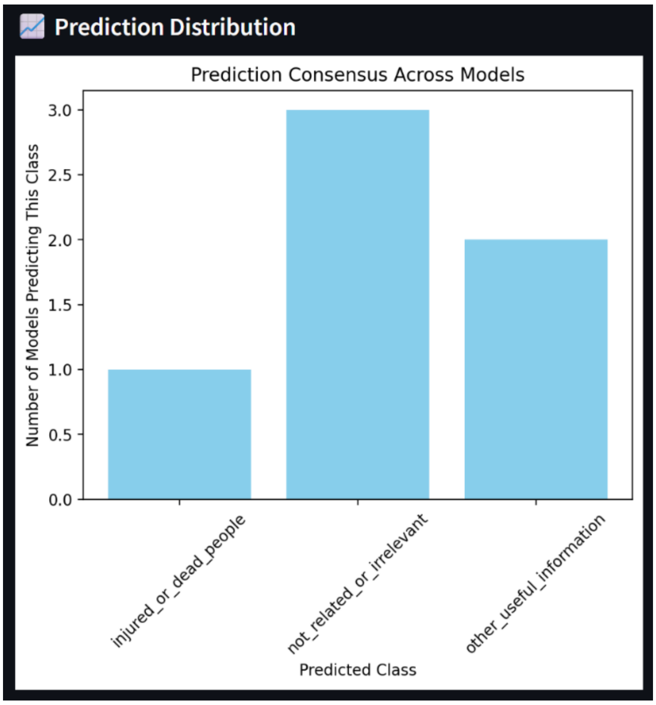


Fig. 15. *Visual Example of Model Consensus*

- Confidence Chart: Shows how confident each model was in its output, which helps evaluate reliability immediately even when predictions differ, as shown in Fig. 16.

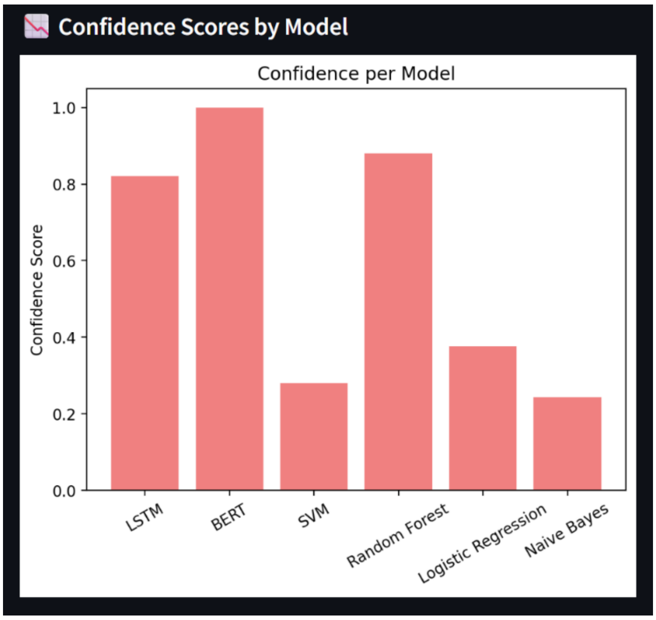


Fig. 16. *Visual Example of Model Confidence*

VII. CONCLUSION

In conclusion, this study successfully demonstrated the potential of machine learning and NLP techniques in automating the classification of disaster-related tweets. It demonstrates Random Forest and BERT as highly effective solutions for classifying social media content into actionable information during crises through a comparative evaluation of multiple models.

As an interactive web tool that can support rapid decision-making in disaster scenarios, the Streamlit-based deployment validated the system's usability [37]. It sets the stage for further development toward a scalable interface even though it is not yet real-time.

These models contribute to more informed and decision-making in disaster response by improving the speed and accuracy of information extraction from social media. However, to improve existing limitations and ensure reliable, scalable, and inclusive deployment of these technologies in real-world crisis scenarios, continued research and development are essential.

This research supports safer and more resilient communities worldwide by providing a solid foundation for future improvements in the intersection of artificial intelligence and humanitarian response.

ACKNOWLEDGMENT

The author would like to express this sincere gratitude to everyone who supported him during this project. Special thanks to his supervisor Assoc. Prof. Dr. Sabrina Tiun for her guidance, feedback, and encouragement throughout the process.

He also acknowledges the support and resources provided by Faculty of Information Science and Technology at Universiti Kebangsaan Malaysia, which made his work possible. Finally, thanks to his peers, friends, and family for their patience and motivation during this project

REFERENCES

[1] M. Kosugi *et al.*, “A twitter-based disaster information sharing system,” *2019 IEEE 4th International Conference on Computer and Communication Systems, ICCCS 2019*, pp. 395–399, Feb. 2019, doi: 10.1109/CCOMS.2019.8821719.

[2] D. S. Krishna, G. Srinivas, and P. V. G. D. Prasad Reddy, “A Deep Parallel Hybrid Fusion Model for disaster tweet classification on Twitter data,” *Decision Analytics Journal*, vol. 11, p. 100453, Jun. 2024, doi: 10.1016/J.DAJOUR.2024.100453.

[3] 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.

[4] N. Nurdin, K. Kluza, M. Fitria, K. Saddami, and R. S. Utami, “Analysis of Social Media Data Using Deep Learning and NLP Method for potential use as Natural Disaster Management in Indonesia,” *Proceeding - 2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering: Sustainable Development for Smart Innovation System, COSITE 2023*, pp. 143–148, 2023, doi: 10.1109/COSITE60233.2023.10249849.

[5] Y. González, “Leveraging Geotagged Social Media to Monitor Spatial Behavior During Population Movements Triggered by Hurricanes,” *Theses and Dissertations*, Jul. 2019, Accessed: Dec. 06, 2024. [Online]. Available: https://scholarcommons.sc.edu/etd/5367

[6] S. Milusheva, R. Marty, G. Bedoya, S. Williams, E. Resor, and A. Legovini, “Applying machine learning and geolocation techniques to social media data (Twitter) to develop a resource for urban planning,” *PLoS One*, vol. 16, no. 2, p. e0244317, Feb. 2021, doi: 10.1371/JOURNAL.PONE.0244317.

[7] M. Karimiziarani, K. Jafarzadegan, P. Abbaszadeh, W. Shao, and H. Moradkhani, “Hazard risk awareness and disaster management: Extracting the information content of twitter data,” *Sustain Cities Soc*, vol. 77, p. 103577, Feb. 2022, doi: 10.1016/J.SCS.2021.103577.

[8] S. Madichetty and M. Sridevi, “A Neural-Based Approach for Detecting the Situational Information from Twitter during Disaster,” *IEEE Trans Comput Soc Syst*, vol. 8, no. 4, pp. 870–880, Aug. 2021, doi: 10.1109/TCSS.2021.3064299.

[9] M. Imran, P. Mitra, and C. Castillo, “Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages,” *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pp. 1638–1643, May 2016, Accessed: Nov. 21, 2024. [Online]. Available: https://arxiv.org/abs/1605.05894v2

[10] R. ALRashdi and S. O’Keefe, “Deep Learning and Word Embeddings for Tweet Classification for Crisis Response,” Mar. 2019, Accessed: Nov. 21, 2024. [Online]. Available: https://arxiv.org/abs/1903.11024v1

[11] P. Seeberger and K. Riedhammer, “Enhancing Crisis-Related Tweet Classification with Entity-Masked Language Modeling and Multi-Task Learning,” Nov. 2022, Accessed: Nov. 21, 2024. [Online]. Available: https://arxiv.org/abs/2211.11468v1

[12] N. P. Shetty, Y. Bijalwan, P. Chaudhari, J. Shetty, and B. Muniyal, “Disaster assessment from social media using multimodal deep learning,” *Multimed Tools Appl*, pp. 1–26, Jul. 2024, doi: 10.1007/S11042-024-19818-0/FIGURES/11.

[13] P. Adili and Y. Chen, “Fast Disaster Event Detection from Social Media: An Active Learning Method,” *INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL*, vol. 19, no. 2, Mar. 2024, doi: 10.15837/IJCCC.2024.2.6460.

[14] F. K. Sufi and I. Khalil, “Automated Disaster Monitoring from Social Media Posts Using AI-Based Location Intelligence and Sentiment Analysis,” *IEEE Trans Comput Soc Syst*, vol. 11, no. 4, pp. 4614–4624, 2024, doi: 10.1109/TCSS.2022.3157142.

[15] 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, May 2020, doi: 10.1016/J.FUTURE.2020.01.005.

[16] S. Z. Hassan *et al.*, “Visual Sentiment Analysis from Disaster Images in Social Media,” *Sensors 2022, Vol. 22, Page 3628*, vol. 22, no. 10, p. 3628, May 2022, doi: 10.3390/S22103628.

[17] N. R. Paul, D. Sahoo, and R. C. Balabantaray, “Classification of crisis-related data on Twitter using a deep learning-based framework,” *Multimed Tools Appl*, vol. 82, no. 6, pp. 8921–8941, Mar. 2023, doi: 10.1007/S11042-022-12183-W/METRICS.

[18] J. Park and S. Cho, “Incorporation of company-related factual knowledge into pre-trained language models for stock-related spam tweet filtering,” *Expert Syst Appl*, vol. 234, p. 121021, Dec. 2023, doi: 10.1016/J.ESWA.2023.121021.

[19] V. V. Kumar, A. Sahoo, S. K. Balasubramanian, and S. Gholston, “Mitigating healthcare supply chain challenges under disaster conditions: a holistic AI-based analysis of social media data,” *Int J Prod Res*, Feb. 2024, doi: 10.1080/00207543.2024.2316884.

[20] S. K. Abid *et al.*, “Toward an Integrated Disaster Management Approach: How Artificial Intelligence Can Boost Disaster Management,” *Sustainability 2021, Vol. 13, Page 12560*, vol. 13, no. 22, p. 12560, Nov. 2021, doi: 10.3390/SU132212560.

[21] Y. Cao, C. Xu, N. M. Aziz, and S. N. Kamaruzzaman, “BIM–GIS Integrated Utilization in Urban Disaster Management: The Contributions, Challenges, and Future Directions,” *Remote Sensing 2023, Vol. 15, Page 1331*, vol. 15, no. 5, p. 1331, Feb. 2023, doi: 10.3390/RS15051331.

[22] J. P. Singh, Y. K. Dwivedi, N. P. Rana, A. Kumar, and K. K. Kapoor, “Event classification and location prediction from tweets during disasters,” *Ann Oper Res*, vol. 283, no. 1–2, pp. 737–757, Dec. 2019, doi: 10.1007/S10479-017-2522-3/FIGURES/9.

[23] A. P. Rodrigues, R. Fernandes, A. Bhandary, A. C. Shenoy, A. Shetty, and M. Anisha, “Real-Time Twitter Trend Analysis Using Big Data Analytics and Machine Learning Techniques,” *Wirel Commun Mob Comput*, vol. 2021, no. 1, p. 3920325, Jan. 2021, doi: 10.1155/2021/3920325.

[24] A. Alabdulaali, A. Asif, S. Khatoon, and M. Alshamari, “Designing Multimodal Interactive Dashboard of Disaster Management Systems,” *Sensors 2022, Vol. 22, Page 4292*, vol. 22, no. 11, p. 4292, Jun. 2022, doi: 10.3390/S22114292.

[25] N. Bahrawi, “Online Realtime Sentiment Analysis Tweets by Utilizing Streaming API Features From Twitter,” *Jurnal Penelitian Pos dan Informatika*, vol. 9, no. 1, pp. 53–62, Oct. 2019, doi: 10.17933/JPPI.V9I1.271.

[26] A. B. Nielsen, D. Landwehr, J. Nicolaï, T. Patil, and E. Raju, “Social media and crowdsourcing in disaster risk management: Trends, gaps, and insights from the current state of research,” *Risk Hazards Crisis Public Policy*, vol. 15, no. 2, pp. 104–127, Jun. 2024, doi: 10.1002/rhc3.12297.

[27] M. Ruoff, U. Gnewuch, A. Maedche, and B. Scheibehenne, “Designing Conversational Dashboards for Effective Use in Crisis Response,” *J Assoc Inf Syst*, vol. 24, no. 6, pp. 1500–1526, Jan. 2023, doi: 10.17705/1jais.00801.

[28] L. Dwarakanath, A. Kamsin, R. A. Rasheed, A. Anandhan, and L. Shuib, “Automated Machine Learning Approaches for Emergency Response and Coordination via Social Media in the Aftermath of a Disaster: A Review,” *IEEE Access*, vol. 9, pp. 68917–68931, 2021, doi: 10.1109/ACCESS.2021.3074819.

[29] M. Karimiziarani, “Social Media Analytics in Disaster Response: A Comprehensive Review,” Jul. 2023, Accessed: Nov. 21, 2024. [Online]. Available: https://arxiv.org/abs/2307.04046v1

[30] A. Lovari and S. A. Bowen, “Social media in disaster communication: A case study of strategies, barriers, and ethical implications,” *J Public Aff*, vol. 20, no. 1, p. e1967, Feb. 2020, doi: 10.1002/PA.1967.

[31] C. Zhang, C. Fan, W. Yao, X. Hu, and A. Mostafavi, “Social media for intelligent public information and warning in disasters: An interdisciplinary review,” *Int J Inf Manage*, vol. 49, pp. 190–207, Dec. 2019, doi: 10.1016/J.IJINFOMGT.2019.04.004.

[32] Y. Liu *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” Jul. 2019, Accessed: Jan. 03, 2025. [Online]. Available: https://arxiv.org/abs/1907.11692v1

[33] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, “Pre-trained Models for Natural Language Processing: A Survey,” *Sci China Technol Sci*, vol. 63, no. 10, pp. 1872–1897, Mar. 2020, doi: 10.1007/s11431-020-1647-3.

[34] C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to Fine-Tune BERT for Text Classification?,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11856 LNAI, pp. 194–206, 2019, doi: 10.1007/978-3-030-32381-3_16.

[35] M. T. H. K. Tusar and M. T. Islam, “A Comparative Study of Sentiment Analysis Using NLP and Different Machine Learning Techniques on US Airline Twitter Data,” *Proceedings of International Conference on Electronics, Communications and*

Information Technology, ICECIT 2021, Oct. 2021, doi: 10.1109/ICECIT54077.2021.9641336.

[36] "CrisisNLP." Accessed: Feb. 11, 2025. [Online]. Available: https://crisisnlp.qcri.org/lrec2016/lrec2016.html

[37] S. Pokhrel, S. Ganesan, T. Akther, and L. Karunarathne, "Building Customized Chatbots for Document Summarization and Question Answering using Large Language Models using a Framework with OpenAI, Lang chain, and Streamlit," *Journal of Information Technology and Digital World*, vol. 6, no. 1, pp. 70–86, Apr. 2024, doi: 10.36548/JITDW.2024.1.006.