FReCS: A First Responder Classification System

Ademola Adesokan[1*], Sanjay Madria[1**], and Long Nguyen[2***]

[1] Department of Computer Science, Missouri University of Science and Technology
[2] School of Applied Computational Sciences, Meharry Medical College
[*]aaadfg@mst.edu, [**]madrias@mst.edu [***]hlnguyen@mmc.edu

Abstract. In today's digital age, categorizing social media data, particularly from platforms like X, can be an effective strategy for identifying key first responders during emergencies, thereby improving overall emergency response efforts. In this study, we introduce a First Responder Classification System (FReCS), a framework that annotates and classifies disaster tweets from 26 crisis events. Our annotations cater for first reponders and their sub-layers. Furthermore, we proposed a classifier called RoBERTa-CAFÉ that integrates pre-trained RoBERTa with Cross-Attention and Focused-Entanglement components, improving the precision and reliability of classification tasks. The model is rigorously tested across publicly available disaster datasets. The RoBERTa-CAFÉ model outperformed state-of-the-art models in identifying relevant emergency communications, displaying its generalization, robustness, and adaptability. Our FReCS approach offers a pioneering technique for classifying first responders and enhances emergency management systems' operational capabilities, leading to more efficient and effective disaster responses. FReCS annotated dataset and code are available on GitHub[3]

Keywords: Data Annotation · Social Media · Emergency Management · First Responder · Transformer

1 Introduction

First responders are crucial in disaster response, playing a vital role in safeguarding lives, properties, and communities as a whole [1]. Their prompt response to emergencies is significant, enabling swift and effective action in disaster situations [2]. This immediacy is vital for mitigating the impact of disasters, potentially addressing immediate needs in order to achieve the aforementioned roles [3]. The predominant focus of research in disaster management has been on the individuals and communities directly affected by calamitous events, with comparatively less emphasis on the experiences of professionals such as police officers and firefighters.

Despite their importance, the effectiveness of first responders varies significantly across different types of disasters. For instance, the San Diego wildfires witnessed the effective deployment of first responders. Emergency managers and public health professionals played a crucial role in integrating their prevention and response efforts, effectively managing the significant disasters faced by the communities [4]. Similarly, the response to mental health calls by first responders following Hurricane Harvey provides insight into emergency service utilization

[3] github.com/abdul0366/FReCS

during the disaster. This study examines the effects of Hurricane Harvey on mental health calls to Emergency Medical Services (EMS) and the Houston Police Department [5], demonstrating the critical role of first responders in managing complex emergencies.

However, the response to Hurricane Harvey also exposed some critical challenges, particularly in the context of Graduate Medical Education (GME) disaster planning at Corpus Christi Medical Center (CCMC). This situation underscored the need for more robust and effective disaster planning within GME programs, highlighting gaps in preparedness and response capabilities [6]. Another significant challenge encountered in disaster management, particularly highlighted during Hurricane Harvey, is the need for accurate data classification for first responders. This issue led to miscommunication between users and responders or volunteers, as evidenced in the event [7]. This gap in clear and accurate information exchange impeded the effective coordination of emergency response efforts, showcasing the need for improved data classification and communication strategies in disaster response. Furthermore, the response to Hurricane Maria brought to light significant challenges in managing disaster complexities and data management. This highlighted an urgent need for an automatic first responder classification system and communication strategy improvements during disaster response, emphasizing enhanced tools and methodologies for managing large-scale emergencies [8].

The absence of a structured classification system for responders in disaster management can lead to significant inefficiencies and heightened risks during emergency responses. As noted by [9], without a clear delineation and classification of roles, first responders may face challenges in coordination and communication, potentially leading to delayed response times, misallocation of resources, and increased risks to responders and affected populations. This lack of organization can worsen the impact of the disaster and impede recovery efforts [9].

The utilization of Social Media (SM) platforms, particularly X (formerly known as Twitter), in disaster management has been increasingly recognized. An online survey conducted among X users who sought help through tweets during Hurricane Harvey revealed a significant finding: 91% of these users reported that X was a valuable tool for facilitating the rescue of affected victims [10]. This statistic reinstated the growing importance of SM platforms in emergency response. SM data can effectively supplement traditional systems like dispatch calls, mostly used in emergency services [11].

Integrating SM data to classify first responders offer several transformative advantages, thereby enhancing disaster response's overall efficiency and effectiveness [1], such as: **1. facilitating the efficient allocation of resources** for informed and effective response strategies [12], **2. fostering greater public engagement** by responding effectively in times of crisis as a critical service [11], **3. Scalability of the traditional system using SM platforms** to handle large volumes of data, enabling the monitoring and response to multiple incidents simultaneously, which is too complicated and complex for manual systems [1], **4. Stabilization** plays a pivotal role in providing immediate assistance, ranging from medical aid and rescue operations to initial damage assessment [13]. **5.**

Psychological support, in addition to physical assistance, first responders are instrumental in helping victims to calm, reassure, and assist individuals in shock or distress, which is essential in mitigating immediate psychological impacts [14].

The study aimed to provide answers to the following research questions:

- **Research Question 1:** How can we accurately classify the appropriate type of first responder for different emergency and crisis conditions?
- **Research Question 2:** Is it feasible to categorize first responders into sub-types that are customized to the specific needs and contextual demands of unique situations?

To address the challenges and questions mentioned earlier and explore the benefits of integrating X data for emergency response, we propose **FReCS**, a **F**irst **Re**sponder **C**lassification **S**ystem. The objective of FReCS is to re-annotate 26 crises for the purpose of first responder classification while also presenting a transformer-based model for classification tasks. Moreover, the model also includes a secondary classification to determine the specific sub-personnel required for crisis and emergency situations. Our FReCS system comprises four major classification tasks as shown in Figure 1: (1.) Relevancy, (2.) Disaster Type, (3.) First Responder, and (4.) Secondary Classification.

To achieve this goal, we employ a blend of advanced deep-learning techniques, including a pre-trained transformer model and multiple custom attention mechanisms. Hence, our classifier, RoBERTa-CAFÉ, comprised of a **RoBERTa** with **C**ross, **A**daptive, and **F**ocused-**E**ntanglement Components which we used to classify different tasks and events based on textual X data for binary and multiclass classification. The following are our major contributions to this work:

- We annotated 27,933 disaster tweets for first responders using the CrisisLexT26 dataset [15]. This framework introduces specific categories for first responders such as Police, EMS, and Firefighters. These classes enable more accurate analysis and classification of crisis-related tweets, thereby improving model training for disaster management. The framework enhances the practical use of SM data in crisis scenarios and strengthens the efficiency of emergency response coordination via digital platforms.
- To achieve specificity and clarity in response, we introduced a secondary annotation to add sub-layers to the first responder category. This process specifies sub-personnel roles such as Mobile Medical Units, Crime Prevention Teams, and Urban Search and Rescue teams for various crises. This detailed annotation allows for more precise resource deployment. It improves the dataset's utility for deeper analysis and modeling, enhancing the effectiveness of emergency management systems utilizing SM data.
- To ensure the reliability and accuracy of the dataset annotations, we used the Fleiss Kappa measure to assess the consistency of annotations among different annotators. Our inter-annotator agreement rating for first responder and secondary labels was 0.89 and 0.85, respectively.
- Our study introduces RoBERTa-CAFÉ, a modified pre-trained RoBERTa model enhanced with Cross-Attention and Focused-Entanglement Components to handle complex data better in crisis scenarios. This model incorpo-

rates Multi-Head Attention and an Adaptive Feed-Forward Network, which enhances its ability to filter and prioritize relevant SM information. The model demonstrates high effectiveness, with F1 and accuracy scores ranging from 86% to 100% across the four tasks. Our model significantly enhances the accuracy and reliability of automated disaster management systems in real-time applications.
– We validate the RoBERTa-CAFÉ model's effectiveness across diverse scenarios, showcasing its generalizability, consistency, robustness, and adaptability. The model effectively classifies data from various crises, tested on datasets such as CrisisLexT6 and CrisisBench, as well as specific events like the Nepal Earthquake and Queensland Floods. Its consistent high performance across different validation methods, including k-fold cross-validation, affirms its reliability as a tool for real-time crisis management.

2 Related Work

Recent evolutions in disaster management and first responder effectiveness have highlighted the role of integrating technology, policy development, and comprehensive training. These mutual efforts aim to improve response times, situational awareness, and overall outcomes during emergencies and disasters.

[16] identified a gap in real-time access to building system data for emergency responders, emphasizing the potential to significantly enhance situational awareness and reduce response times. They proposed a roadmap to overcome challenges in securely transmitting and processing building sensor data to first responders, emphasizing the need for a systemic approach to improve emergency response via informed decision-making.

Similarly, [1] introduced the ONSIDE, which leverages SM platforms to streamline disaster response coordination. By integrating Information-Centric Networking with a SM Engine, ONSIDE addresses the real-time analysis challenges of SM data, utilizing natural language processing to ensure rapid and relevant information delivery to first responders.

In aviation safety, [17] proposed an In-Time Aviation Safety Management System designed specifically for UAS and autonomous systems in emergency scenarios. This system emphasizes predictive modeling to proactively identify and mitigate risks, highlighting the need for scenario testing to identify new safety data requirements and operational hurdles.

Furthermore, emphasizing the role of education, [18] explored the impact of an Emergency First Response (EFR) training program at Tecnológico de Monterrey. Their findings underscored the importance of integrating emergency response training within higher education to enhance EFR skills across various disciplines, improving community and workplace safety.

Addressing the communication challenges in disaster scenarios, [19] presented ReDiCom, a resilient architecture designed to enhance first responder communication. By supporting network resilience and utilizing coded computation, ReDiCom facilitates efficient information dissemination and resource management, underscoring the potential of technological advancements in improving disaster management.

On the policy front, [20] examined state-level policies addressing first responder mental health. Their study categorized policies into workers' compensation-related and non-workers' compensation-related, highlighting legislative efforts to support first responders facing adverse mental health outcomes due to occupational trauma and the need for systematic evaluations to establish evidence-based mental health care practices.

Lastly, [21] developed SOSFloodFinder, a system utilizing NLP and GPS technologies to classify urgency in emergency communications from flood victims. This innovation demonstrates how technology can enhance the precision and efficiency of first responder activities during floods, contributing to the broader goal of improving disaster management and response.

These studies demonstrate a comprehensive approach to enhancing the efficiency and effectiveness of emergency management and first responder activities. These efforts aim to improve safety, efficiency, and outcomes in disaster response and emergency situations by leveraging technology, policy development, and targeted training. However, FReCS stands apart from existing studies by incorporating sub-types into the first responder category. This allows for identifying specialized personnel (such as those in the police units responsible for criminal activities). This approach enables customized responses to unique emergency situations instead of a generalized approach that treats all situations with the same protocol.

3 Our Approach

This section outlines our methodology for accurately categorizing tweets for emergency response coordination. This process involves dataset annotation and a multi-level classification framework, as shown in Figure 1. We further provide a detailed explanation of each step in the following subsection.

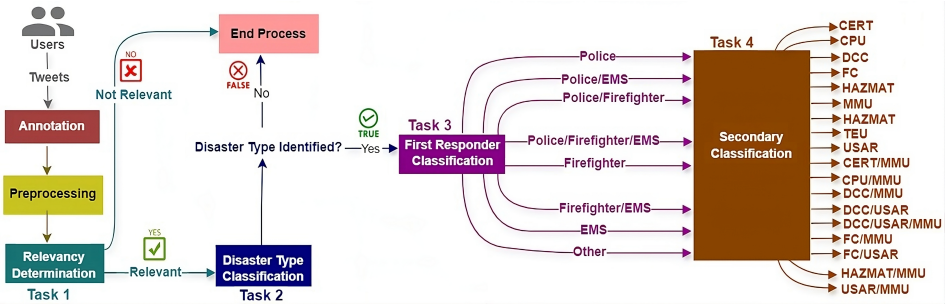

Fig. 1: ***FReCS Proposed System Framework.***

3.1 Dataset and Annotation Process

In this study, we utilized the CrisisLexT26 dataset [15], comprising about 28,000 tweets across 26 crisis events from 2012 and 2013, initially annotated by crowdsourced workers based on event types (e.g., Flood, Wildfire, Earthquake), informativeness, information types (e.g., caution and advice, infrastructure damage), and information sources (e.g., governments, NGOs). For detailed documentation on

the crowdsourced annotations, see [15]. Notably, the dataset lacked annotations for first responders and secondary classifications. To fill this gap, we conducted a detailed annotation process over two months with a team of three students (two annotators and one experienced moderator). Our initial primary annotation encompassed four label classes for first responders: Police, EMS, Firefighter, and Other, aligning with FEMA standards. While recognizing that some regions classify additional agencies as first responders, we maintained these three primary categories for consistency across different jurisdictions. For secondary annotations, we initially introduced nine label categories: Mobile Medical Unit (MMU), Community Emergency Response Team (CERT), Crime Protection/Prevention Unit (CPU), Dispatch Call Center (DCC), Traffic Enforcement Unit (TEU), Hazardous Materials (HAZMAT), Fire Control (FC), Urban Search and Rescue (USAR), and Other. Table 1 shows the class-label distribution.

First Responder Class Labels	**Secondary Layer Class Labels**
Police: 3953, **EMS:** 753, **Firefighter:** 1248, **Police/EMS:** 488, **Police/Firefighter:** 450, **Firefighter/EMS:** 181, **Police/Firefighter/EMS:** 290, **Other:** 20570	**FC/USAR:** 53, **MMU:** 481, **USAR/MMU:** 90, **USAR:** 236, **FC/MMU:** 65, **FC:** 634, **DCC/USAR/MMU:** 290, **DCC/MMU:** 385, **DCC/USAR:** 450, **DCC:** 2435, **CERT/MMU:** 158, **CERT:** 114, **CPU/MMU:** 103, **CPU:** 315, **HAZMAT:** 325, **HAZMAT/MMU:** 26, **TEU:** 80, **Other:** 20570

Table 1: ***Label Distribution for Task 3 and 4***

This classification allowed for a more nuanced assignment of resources, with DCC and TEU functioning as sub-layers of Police; CERT and MMU under EMS; and FC, HAZMAT, and USAR under Firefighter. This secondary classification aimed to enhance operational specificity and improve response efficiency by deploying the most suitable responder team to each unique emergency. During the annotation, it became evident that some tweets required the simultaneous deployment of multiple responder types. We addressed this complexity by assigning multiple labels where necessary, expanding the first responder and secondary classification labels from four to eight and nine to eighteen, respectively, as shown in Figure 1. Following the initial annotation phase, we engaged in a rigorous review process. This collaborative approach involved annotators actively verifying each other's work, with the experienced moderator resolving any disagreements. We then assessed the consistency of these annotations through the inter-annotator agreement process. Each annotator rated their agreement with a score of 1 or disagreement with a score of 0. We employed the Fleiss Kappa statistical measure [22] to gauge the level of consensus among annotators. The results revealed a high consistency rate, with a Fleiss Kappa score of 0.89 and 0.85 for the first responder and secondary labels, respectively, indicating substantial agreement and affirming the reliability and accuracy of our annotations.

Our **preprocessing** steps include normalizing text [23], removing duplicates and links [24], special characters and stopwords [25] to ensure accurate and reliable model training.

3.2 RoBERTa-CAFÉ Classifier

Our classifier model, RoBERTa-CAFÉ (Cross-Attention and Focused-Entanglement) include classification of Task 1 (Relevancy), Task 2 (Disaster Type), Task 3 (First Responder) and Task 4 (Secondary layers), which integrates advanced attention mechanisms and RoBERTa's contextual embeddings to accurately classify disaster tweets for different tasks. The model has several core components, as shown in Figure 2.

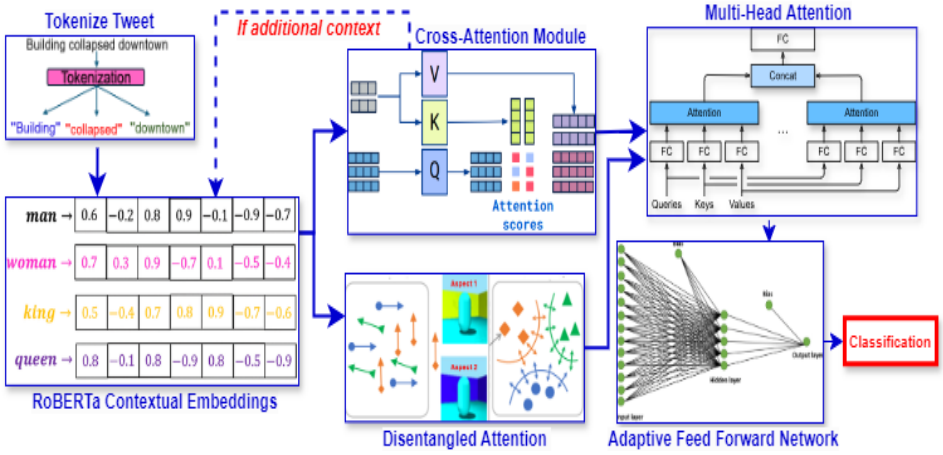

Fig. 2: ***RoBERTa-CAFÉ Classifier***

(1) RoBERTa contextual embeddings [26]: RoBERTa, an advanced version of BERT, forms the backbone of the RoBERTa-CAFÉ model, providing the ability to generate deep contextual embeddings. These embeddings enable the model to understand subtle language dynamics. The embeddings are expressed as follows:

$$\mathbf{E} = E_{\text{tokens}} + E_{\text{positions}} \tag{1}$$

We then process the input embeddings through multiple layers of transformer block, each block consisting of:

- Multi-Head Self-Attention, which allows the model to attend to different parts of the input sequence:

$$\mathbf{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(head_1, \ldots, head_h)W^O \tag{2}$$

- Additionally, a position-wise fully connected feed-forward network is applied to each position separately and identically in each layer:

$$\mathbf{FFN}(\mathbf{x}) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{3}$$

- To aid in stabilizing the learning process, residual connections and layer normalization are used. The output is obtained by adding the output of the sublayer operation to the input, followed by layer normalization:

$$\mathbf{output} = \text{LayerNorm}(x + \text{Sublayer}(x)) \tag{4}$$

  Where $\text{Sublayer}(x)$ is the operation by the multi-head attention or the feed-forward network.

(2) Cross-Attention Module (CAM): integrates external contextual information with RoBERTa's embeddings [27]. This mechanism allows the model to focus on specific parts of the text by considering the additional context provided, thus enhancing its ability to adapt to various situations and datasets. The module is an extension of the self-attention mechanism and is applied between two different sets of inputs: the main input x and the context input context. Our CAM module comprises multiple steps:

- A linear transformation of the x (query) and context (key and value) inputs into query, key, and value spaces:

$$\mathbf{Q} = W_q x + b_q, \quad \mathbf{K} = W_k \text{context} + b_k, \quad \mathbf{V} = W_v \text{context} + b_v \tag{5}$$

  Where W_q, W_k, W_v are weight matrices and b_q, b_k, b_v are biases for queries, keys, and values, respectively.
- Attention scores are computed by taking the dot product of the query with the key of each element in the context and dividing it by the square root of the dimension of keys to stabilize gradients during training:

$$\textbf{scores} = \frac{QK^T}{\sqrt{d_k}} \tag{6}$$

  Where d_k is the dimensionality of the keys.
- We applied the softmax function to the scores to obtain the attention weights:

$$\textbf{attention weights} = \text{softmax}(\text{scores}) \tag{7}$$

- The output is computed as a weighted sum of the values with the weights given by the attention weights:

$$\textbf{output} = \text{attention weights} \cdot V \tag{8}$$

- Finally, the output is summed with the input to let the layer perform as a residual connection:

$$\textbf{attended output} = \text{output} + x \tag{9}$$

(3) Disentangled Attention (DeA): This component divides the attention mechanism into various paths (aspects) to aid the model in separately and simultaneously learning different kinds of features from the data [28], such as semantic and syntactic features. This division helps capture the diverse nature of language used in disaster-related communications more efficiently. In the DeA module, we processed two separate aspects of the input using sigmoid-activated linear transformations. It allows the layer to concentrate on different input aspects or features independently. The output is obtained by combining the two aspects and multiplying it with the input. The mechanism is represented as:

$$\textbf{aspect}_1(x) = \sigma(W_1 x + b_1) \ \textit{and} \ \textbf{aspect}_2(x) = \sigma(W_2 x + b_2) \tag{10}$$

$$\textbf{output} = (\text{aspect}_1(x) + \text{aspect}_2(x)) \cdot x \tag{11}$$

Where σ is the sigmoid function, W_1, W_2 are the weight matrices, b_1, b_2 are bias vectors, and x is the input.

(4) Multi-Head Attention (MHA): is a mechanism that allows models to attend to different representation subspaces at different positions, enabling them to capture a variety of dependencies in the input [29]. With multiple 'heads,' the model can capture a variety of dependencies in the input, such as those between different key terms in disaster data, which is crucial for accurate classification. Our MHA module divides the model's attention into multiple 'heads,' allowing it to attend to different parts of the input simultaneously. We represent the breakdown of the multi-head as follows:

$$\mathbf{Q} = W_Q X,\ \mathbf{K} = W_K X,\ \text{ and } \mathbf{V} = W_V X \tag{12}$$

$$\textbf{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{13}$$

$$\textbf{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(head_1, \ldots, head_h)W^O \tag{14}$$

Where heads are the individual attention outputs and W^O is another learned parameter matrix.

(5) Adaptive Feed Forward Network (AFFN): This is composed of feed-forward layers that utilize gating mechanisms, such as GLUs, to regulate the flow of information. This network adapts by enhancing or reducing feature representations, enabling it to focus on relevant features while discarding less important data [30].

$$\mathbf{x_{new}} = \text{GLU}(W_i x + b_i) = (W_{i,1}x + b_{i,1}) \otimes \sigma(W_{i,2}x + b_{i,2}) \tag{15}$$

Where $W_{i,1}, W_{i,2}$ and $b_{i,1}, b_{i,2}$ are the weights and biases of the linear transformations, σ is the sigmoid activation function, and $\otimes$ denotes element-wise multiplication.

(6) Classification Layer: The final layer of the model is a linear layer that maps the enriched text representations to the output classes, which correspond to different classifications under different tasks. This makes the model a valuable tool for automated disaster response systems.

Uniqueness: Our RoBERTa—CAFÉ is unique as it integrate enhanced attention mechanisms (CAM and DeA) that allows the model to focus on what is being said and how different aspects of the information related to external contexts and internal text structures. It also provides robust feature processing (AFFN and MHA) that ensures the model can efficiently process a wide array of textual features, enhancing the classifier's accuracy and flexibility across diverse disaster-related datasets.

4 Results

This section outlines the experimental validation of the methods introduced previously. The performance of our models was rigorously tested through train/test splits and k-fold cross-validation, employing metrics such as accuracy, precision, recall, and F1 score. These experiments were performed on a robust computational system featuring dual NVIDIA® Tesla V100 GPUs and an Intel® Xeon®

Gold CPU, offering substantial computing power to meet the intensive processing requirements of our deep learning frameworks.

4.1 Model Training and Testing

Our RoBERTa-CAFÉ classifier employs advanced neural network architectures for accurate tweet classification. Recall in Section 3.2 that we use sophisticated attention mechanisms, including cross-attention, disentangled attention, multi-head attention, and an adaptive feed-forward network, to effectively handle complex textual dependencies.

Preprocessed tweets use RoBERTa's tokenizer to transform the text into token sequences. The tokens are managed by a custom PyTorch Dataset class for optimized batching and loading during training and testing. The RoBERTa-CAFÉ model is trained on a labeled dataset, split into different training and testing sets, using a DataLoader for efficient batch processing. We use *RAdam* and a learning rate scheduler to optimize training and achieve stable convergence in multiple epochs. The training process involves minimizing the loss for multi-class and binary classification tasks by adjusting model weights iteratively. Our hyperparameters are shown in Table 2

The model's performance is assessed using the metrics described, including classification reports, after thorough training. This evaluation process examines the model's classification accuracy and its capacity to generalize to unseen data, ensuring that the RoBERTa-CAFÉ classifier effectively learns from the training data and remains robust when faced with new datasets. More information about our model training and hyperparameters can be found in our code on this link.

Table 2: ***Hyperparameters and their values for our Classification model***

Hyperparameter	**Value**
Input Dimensions	768
Number of Heads in MHA	8
AFFN Dimension	2048
Depth of AFFN	2 layers
Dropout Rate	0.2
Number of Classes	Variable (as per task)
Batch Size	64
Optimizer	RAdam
Learning Rate	2×10^{-5}
Loss Function	CrossEntropyLoss/BCE
Learning Rate Scheduler	Step LR (gamma=0.1, step size=10)
Epochs	10
Max Length for Tokenization	128 characters

4.2 Train/Test Split Vs. K-fold Cross Validation

For all four tasks, our study tested different train/test splits (90/10, 80/20, 70/30, 60/40) and cross-fold validations (5, 10, and 15). Our analysis indicates that the performance metrics results for different train/test splits ranging from 90/10 to 60/40 are highly consistent. This uniformity is illustrated in Table 3 by the matched F1 score and accuracy results across all tasks. Our analysis suggests

that there are no significant differences in the performance metrics attributable to the proportion of the split, thus indicating the model's stability.

Moreover, the cross-fold validation results comprising 5, 10, and 15 folds, as demonstrated in Table 4, show comparable outcomes with negligible variances. The consistency of the results across various split ratios and cross-validation folds highlights the robustness of our model, which confirms its reliability irrespective of data segmentation methods.

Tasks	Train (90)/Test (10)							Train (80)/Test (20)							Train (70)/Test (30)							Train (60)/Test (40)						
	Macro Avg			Wei. Avg				Macro Avg			Wei. Avg				Macro Avg			Wei. Avg				Macro Avg			Wei. Avg			
	Pr	Re	F1	Pr	Re	F1	Acc	Pr	Re	F1	Pr	Re	F1	Acc	Pr	Re	F1	Pr	Re	F1	Acc	Pr	Re	F1	Pr	Re	F1	Acc
Task 1	0.83	0.83	0.83	0.94	0.94	0.94	0.94	0.85	0.80	0.82	0.94	0.94	0.94	0.94	0.86	0.78	0.81	0.93	0.94	0.93	0.94	0.86	0.80	0.82	0.93	0.94	0.94	0.94
Task 2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Task 3	0.71	0.70	0.70	0.87	0.87	0.87	0.87	0.71	0.72	0.71	0.87	0.87	0.87	0.87	0.67	0.71	0.68	0.87	0.87	0.87	0.87	0.68	0.69	0.68	0.86	0.86	0.86	0.86
Task 4	0.58	0.66	0.60	0.87	0.86	0.86	0.86	0.65	0.54	0.57	0.86	0.86	0.85	0.86	0.61	0.59	0.59	0.86	0.85	0.85	0.85	0.59	0.57	0.56	0.86	0.84	0.85	0.84

Table 3: ***Result of our 4 tasks under 4 train/test splits using R–CAFÉ.***

Tasks	5-Fold							10-Fold							15-Fold						
	Macro Avg			Wei. Avg				Macro Avg			Wei. Avg				Macro Avg			Wei. Avg			
	Pr	Re	F1	Pr	Re	F1	Acc	Pr	Re	F1	Pr	Re	F1	Acc	Pr	Re	F1	Pr	Re	F1	Acc
Task 1	**0.85**	**0.81**	0.82	**0.93**	**0.94**	**0.93**	**0.94**	**0.85**	**0.81**	**0.83**	**0.93**	**0.94**	**0.93**	**0.94**	**0.85**	**0.81**	**0.83**	**0.93**	**0.94**	**0.93**	**0.94**
Task 2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Task 3	0.68	0.72	0.70	0.87	0.86	0.86	0.86	0.69	0.72	0.70	0.88	0.87	0.87	0.87	0.67	0.71	0.69	0.87	0.86	0.87	0.86
Task 4	0.60	0.62	0.60	0.86	0.86	0.86	0.86	0.62	0.63	0.61	0.87	0.86	0.86	0.86	0.58	0.63	0.59	0.86	0.86	0.86	0.86

Table 4: ***Evaluation of our 4 tasks under different folds using R–CAFÉ.***

4.3 Task 1 (Relevancy) and Task 2 (Disaster Type)

Our RoBERTa-CAFÉ model outperforms [31] model in Task 1 as proven in Table 5, which focuses on relevancy. RoBERTa-CAFÉ achieved a recall improvement of 20% and an F1 score of 0.93, showing a 14% improvement over [31]'s F1 score. RoBERTa-CAFÉ's higher difference recall suggests that it is better at identifying relevant cases, while the increased precision suggests that its relevancy classification is more accurate. The higher F1 score confirms that RoBERTa-CAFÉ has a better balance of precision and recall, as shown in Table 5. Both

Table 5: ***Performance of Task 1 and 2***

Models	Task 1			Task 2		
	P	**R**	**F1**	**P**	**R**	**F1**
Burel et al. [31]	0.87	0.74	0.79	1.00	1.00	1.00
R-CAFÉ	**0.93**	**0.94**	**0.93**	1.00	1.00	1.00

the RoBERTa-CAFÉ and the [31] model achieved perfect scores in Task 2, which involves classifying disaster types. This indicates that both models are highly effective in accurately identifying all relevant cases without false positives or negatives. Furthermore, there is no significant difference in performance metrics between the two models for this task.

4.4 Task 3 (First Responder) and Task 4 (Secondary Classification)

Due to the uniqueness of our annotation labels, which creates a gap of not having related work to compare with, we evaluated the performance of the RoBERTa-CAFÉ model against four baseline classifiers: Decision Tree - DT, Naïve Bayes - NB, Support Vector Machine - SVM, and Logistic Regression - LR in Tasks 3 and

4, which involved first responder and secondary classification. The implementation of the four baselines is similar to the work of [32]. The results from Table 6 and 7 showed that RoBERTa-CAFÉ provided a competitive approach with high and consistent scores across metrics, demonstrating its robustness in handling the tasks' unique requirements.

The model's performance was particularly notable for maintaining high accuracy and recall, which are crucial for reliable disaster response applications. RoBERTa-CAFÉ performed well in the recall, which is a critical factor in secondary classification tasks. It showed effectiveness in handling complex classification tasks, as demonstrated by its high F1 scores across all splits. The model's accuracies were also consistently high, indicating reliable performance across different data partitions.

Models	Train (90)/Test (10)							Train (80)/Test (20)							Train (70)/Test (30)							Train (60)/Test (40)						
	Macro Avg			Wei. Avg				Macro Avg			Wei. Avg				Macro Avg			Wei. Avg				Macro Avg			Wei. Avg			
	Pr	Re	F1	Pr	Re	F1	Acc	Pr	Re	F1	Pr	Re	F1	Acc	Pr	Re	F1	Pr	Re	F1	Acc	Pr	Re	F1	Pr	Re	F1	Acc
DT	0.34	0.21	0.23	0.69	0.77	0.70	0.77	0.32	0.19	0.21	0.67	0.76	0.68	0.76	0.39	0.20	0.22	0.68	0.76	0.68	0.76	0.33	0.19	0.21	0.67	0.76	0.68	0.76
NB	0.64	0.38	0.45	0.80	0.81	0.80	0.81	0.59	0.36	0.41	0.79	0.80	0.78	0.80	0.65	0.37	0.42	0.80	0.81	0.79	0.81	0.64	0.34	0.39	0.79	0.80	0.78	0.80
SVM	0.68	0.53	0.59	0.82	0.84	0.83	0.84	0.68	0.53	0.58	0.82	0.84	0.82	0.84	0.70	0.53	0.59	0.82	0.84	0.82	0.84	0.69	0.52	0.58	0.82	0.83	0.82	0.83
LR	0.68	0.52	0.58	0.82	0.84	0.83	0.84	0.68	0.51	0.57	0.82	0.83	0.82	0.83	0.68	0.49	0.55	0.82	0.83	0.82	0.83	0.69	0.48	0.55	0.81	0.83	0.81	0.83
R–CAFÉ	**0.71**	0.70	0.70	**0.87**	**0.87**	**0.87**	**0.87**	**0.71**	**0.72**	**0.71**	**0.87**	**0.87**	**0.87**	**0.87**	0.67	0.71	0.68	**0.87**	**0.87**	**0.87**	**0.87**	0.68	0.69	0.68	0.86	0.86	0.86	0.86

Table 6: ***R–CAFÉ comparison with baseline models for Task 3 under 4 splits.***

Models	Train (90)/Test (10)							Train (80)/Test (20)							Train (70)/Test (30)							Train (60)/Test (40)						
	Macro Avg			Wei. Avg				Macro Avg			Wei. Avg				Macro Avg			Wei. Avg				Macro Avg			Wei. Avg			
	Pr	Re	F1	Pr	Re	F1	Acc	Pr	Re	F1	Pr	Re	F1	Acc	Pr	Re	F1	Pr	Re	F1	Acc	Pr	Re	F1	Pr	Re	F1	Acc
DT	0.19	0.15	0.15	0.66	0.77	0.69	0.77	0.24	0.13	0.14	0.65	0.76	0.68	0.76	0.28	0.13	0.14	0.67	0.76	0.68	0.76	0.21	0.12	0.13	0.65	0.75	0.68	0.75
NB	0.50	0.22	0.27	0.79	0.81	0.78	0.81	0.45	0.22	0.26	0.76	0.80	0.77	0.80	0.40	0.21	0.25	0.76	0.80	0.77	0.80	0.42	0.20	0.24	0.76	0.80	0.76	0.80
SVM	0.60	0.44	0.48	0.82	0.84	0.82	0.84	0.61	0.40	0.46	0.81	0.83	0.81	0.83	0.60	0.38	0.45	0.81	0.83	0.81	0.83	0.61	0.38	0.44	0.80	0.83	0.81	0.83
LR	0.67	0.44	0.51	0.82	0.83	0.82	0.83	0.61	0.36	0.43	0.81	0.83	0.81	0.83	0.59	0.33	0.40	0.81	0.83	0.81	0.83	0.61	0.32	0.40	0.80	0.82	0.80	0.82
R–CAFÉ	0.58	**0.66**	**0.60**	**0.87**	**0.86**	**0.86**	**0.86**	**0.65**	0.54	0.57	0.86	**0.86**	0.85	**0.86**	0.61	0.59	0.59	0.86	0.85	0.85	0.85	0.59	0.57	0.56	0.86	0.84	0.85	0.84

Table 7: ***R–CAFÉ comparison with baseline models for Task 4 under 4 splits.***

4.5 Ablation Study

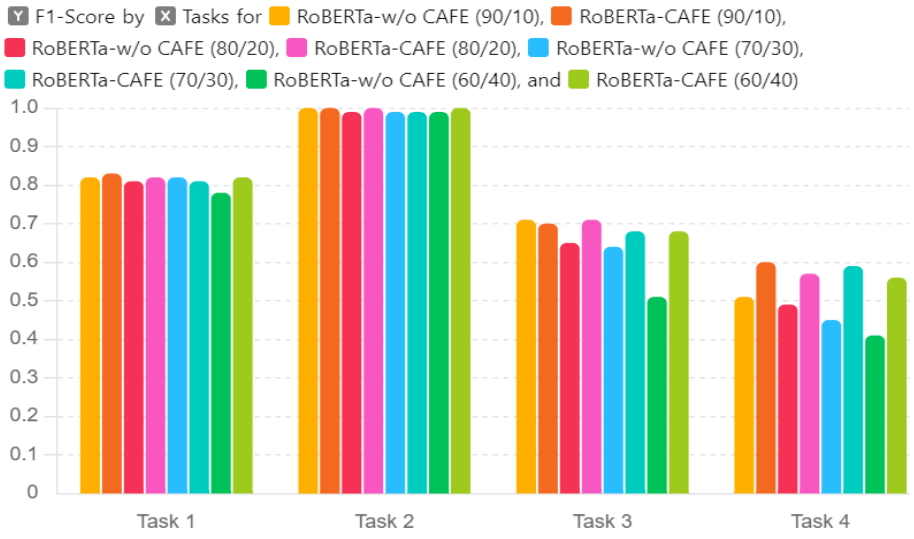

Fig. 3: ***RoBERTa–CAFÉ Vs finetuned RoBERTa without CAFÉ***

In our experiments, we analyzed the impact of custom attention layers on the performance of the RoBERTa-CAFÉ classifier. We compared the RoBERTa-CAFÉ with a version without custom attention layers (CAFÉ) across four tasks and train/test splits, using F1 scores as the benchmark. The results showed that incorporating custom attention layers improved the RoBERTa-CAFÉ's performance, as shown in Figure 3.

The RoBERTa-CAFÉ model consistently outperformed the RoBERTa w/o CAFÉ model in Task 1 for all train/test dataset splits. For Task 2, both models performed well, with the CAFÉ layers not showing a significant performance enhancement; this is attributed to the distinguishable characteristics inherent in this classification task. Task 3 demonstrated a noticeable improvement in performance for the RoBERTa-CAFÉ model compared to the RoBERTa w/o CAFÉ model, with the most significant increase in F1 score observed in the 90/10 data split. Task 4 showed a robust enhancement in the RoBERTa-CAFÉ model's performance, particularly in the 60/40 split, where the CAFÉ layers resulted in a 15% increase in F1 score over the RoBERTa w/o CAFÉ model.

4.6 Beyond FReCS: CrisisLexT6, CrisisBench, NEQ and QFL

In our study, we extended the RoBERTa-CAFÉ model to analyze its performance on various publicly available crisis-related datasets to evaluate its effectiveness and generalizability beyond the FReCS dataset. We considered four datasets: CrisisLexT6, CrisisBench, Nepal Earthquake (NEQ), and Queensland Flood (QFL). Subsequently, we summarize our findings and highlight the comparison with existing models, thereby demonstrating the robust capabilities of RoBERTa-CAFÉ across diverse crisis communication scenarios.

NEQ dataset from Table 9, RoBERTa-CAFÉ, significantly outperformed [34] by increasing all metrics by 17 points, achieving 0.79 across these metrics. Similarly, in the QFL dataset, RoBERTa-CAFÉ's performance outperformed [34] by achieving an impressive score difference of 0.17 for precision, 0.16 for recall and F1 score, marking substantial improvements.

Table 8: ***Comparison with CrisisBench Dataset.***

Models	**Informativeness**				**Humanitarian**			
	P	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**
Alam et al. [33]	0.88	0.88	0.88	0.88	0.79	0.78	0.78	0.78
R-CAFÉ	0.88	0.88	0.88	0.88	0.79	0.78	0.78	0.78

Table 9: ***Comparison with NEQ and QFL Dataset.***

Models	**NEQ**			**QFL**		
	P	**R**	**F1**	**P**	**R**	**F1**
Alam & Imran [34]	0.65	0.65	0.65	0.93	0.94	0.94
R-CAFÉ	**0.79**	**0.79**	**0.79**	**0.97**	**0.96**	**0.96**

Regarding the CrisisLexT6 dataset from Table 10, RoBERTa-CAFÉ achieved an accuracy and macro-F1 score of 0.95, matching the performance of the best existing model by [35] and slightly outperforming [36] on accuracy. As for the CrisisBench dataset in Table 8, RoBERTa-CAFÉ mirrored the performance of [33] across all metrics.

Table 10: ***Comparison with CrisisLexT6 Dataset***

Models	**Acc**	**Models**	**M-F1**
Jaoa [36]	0.94	Li et al. [37]	0.90
Li et al. [35]	**0.95**	Li et al. [35]	**0.95**
R-CAFÉ	**0.95**	R-CAFÉ	**0.95**

These results affirm the versatility of our RoBERTa-CAFÉ model in handling a range of crisis-related communications with high precision and reliability. Its ability to adapt and maintain high performance across various datasets stresses its potential as a powerful crisis management and response tool.

5 Conclusion and Future Work

The effectiveness of FReCS, a First Responder Classification System that utilizes the advanced capabilities of the RoBERTa-CAFÉ model to scrutinize and classify SM data for emergency response purposes, has been demonstrated in this study. Our findings reveal that integrating refined custom attention mechanisms into the pre-trained RoBERTa model significantly enhances FReCS's precision and speed in identifying relevant emergency-related communications. The system's robust performance across various datasets highlights its potential to revolutionize the landscape of disaster management by providing timely and accurate information crucial for first responder deployment efficiency.

In future research, we plan to enhance the system's applicability and reliability across different geographical and cultural contexts; we intend to expand the dataset to include a more extensive range of languages. Additionally, integrating multimedia data, such as images and videos to enrich the system's contextual understanding and response accuracy.

Acknowledgement

This research project received support from the NSF - USA CNS-2219615, CNS-2219614, and the Kummer Institute for Student Success, Research, and Economic Development at the Missouri University of Science and Technology through the Kummer Innovation and Entrepreneurship Doctoral Fellowship.

References

1. V. Mittal, M. Jahanian, and K. K. Ramakrishnan, "Online Delivery of Social Media Posts to Appropriate First Responders for Disaster Response," in *ACM International Conference Proceedings Series*, 2021, doi: 10.1145/3427477.3429272.
2. S. Khatoon *et al.*, "Development of social media analytics system for emergency event detection and crisis management," *Computers, Materials and Continua*, vol. 68, no. 3, 2021, doi: 10.32604/cmc.2021.017371.
3. K. Wu, J. Wu, and M. Ye, "A review on the application of social media data in natural disaster emergency management," *Progress in Geography*, vol. 39, no. 8, 2020, doi: 10.18306/dlkxjz.2020.08.014.
4. B. W. Clements and J. A. P. Casani, *Disasters and Public Health: Planning and Response: Second Edition*, 2016, doi: 10.1016/C2014-0-01322-6.

5. J. Saunders *et al.*, "Emergency mental health calls to first responders following a natural disaster: Examining the effects from Hurricane Harvey," *International Journal of Academic Medicine*, vol. 7, no. 1, 2021, doi: 10.4103/IJAM.IJAM_71_20.
6. B. Newman and C. Gallion, "Hurricane Harvey: Firsthand Perspectives for Disaster Preparedness in Graduate Medical Education," *Academic Medicine*, vol. 94, no. 9, 2019, doi: 10.1097/ACM.0000000000002696.
7. L. Zou *et al.*, "Social Media for Emergency Rescue: An Analysis of Rescue Requests on Twitter during Hurricane Harvey," *International Journal of Disaster Risk Reduction*, vol. 85, 2023, doi: 10.1016/j.ijdrr.2022.103513.
8. M. M. Kress, K. F. Chambers, D. D. Hernandez Abrams, and S. K. McKay, "Principles for data management, visualization, and communication to improve disaster response management: Lessons from the Hurricane Maria response mission," *Journal of Emergency Management*, vol. 19, no. 8, 2021, doi: 10.5055/jem.0658.
9. G. D. Haddow, J. A. Bullock, and D. P. Coppola, *Introduction to Emergency Management*, 2020, doi: 10.1016/B978-0-12-817139-4.01001-0.
10. V. V. Mihunov, N. S. N. Lam, L. Zou, Z. Wang, and K. Wang, "Use of Twitter in Disaster Rescue: Lessons Learned from Hurricane Harvey," *International Journal of Digital Earth*, vol. 13, no. 12, 2020, doi: 10.1080/17538947.2020.1729879.
11. R. H. Kirby, M. Reams, and N. S. N. Lam, "The Use of Social Media by Emergency Stakeholder Groups: Lessons Learned from Areas Affected by Hurricanes Isaac and Sandy," *Journal of Homeland Security and Emergency Management*, vol. 20, no. 2, 2023, doi: 10.1515/jhsem-2021-0031.
12. R. Koshy and S. Elango, "Utilizing Social Media for Emergency Response: A Tweet Classification System Using Attention-Based BiLSTM and CNN for Resource Management," *Multimedia Tools and Appl.*, 2023, doi: 10.1007/s11042-023-16766-z.
13. N. Ein *et al.*, "Physical and psychological challenges faced by military, medical and public safety personnel relief workers supporting natural disaster operations: a systematic rev.," *Curr. Psych.*, vol. 43, no. 2, 2024, doi: 10.1007/s12144-023-04368-9.
14. L. Dong and J. Bouey, "Public Mental Health Crisis during COVID-19 Pandemic, China," *Emerg. Infec. Dis.*, vol. 26, no. 7, 2020, doi: 10.3201/eid2607.200407.
15. A. Olteanu, S. Vieweg, and C. Castillo, "What to expect when the unexpected happens: Social media communications across crises," in *CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing*, 2015, doi: 10.1145/2675133.2675242.
16. D. G. Holmberg, M. A. Raymond, and J. Averill, "Delivering Building Intelligence to First Responders," *Nat. Institute of Standards and Tech. Technical Note*, 2013.
17. D. Kirkman *et al.*, "Informing New Concepts for UAS and Autonomous System Safety Management using Disaster Management and First Responder Scenarios," in *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*, vol. 2021-October, 2021, doi: 10.1109/DASC52595.2021.9594356.
18. G. Dieck-Assad, O. I. González Peña, and J. M. Rodríguez-Delgado, "Evaluation of emergency first response's competency in undergraduate college students: Enhancing sustainable medical education in the community for work occupational safety," *International Journal of Environmental Research and Public Health*, vol. 18, no. 15, 2021, doi: 10.3390/ijerph18157814.
19. K. K. Ramakrishnan, M. Yuksel, H. Seferoglu, J. Chen, and R. A. Blalock, "Resilient Communication for Dynamic First Responder Teams in Disaster Management," *IEEE Communications Magazine*, 2022, doi: 10.1109/MCOM.003.2200015.
20. K. O'Dare, A. Mathis, R. Tawk, L. Atwell, and D. Jackson, "State Level Policies on First Responder Mental Health in the U.S.: A Scoping Review," *Admin. and Pol. in Mental Health and Mental Health Serv. Res.*, 2024, doi: 10.1007/s10488-024-01352-8.

21. S. H. Kamal, A. A. Aziz, and W. A. Mustafa, "SOSFloodFinder: A Text-Based Priority Classification System for Enhanced Decision-Making in Optimizing Emergency Flood Response," *Jour. of Aut. Intel.*, vol. 7, no. 1, 2024, doi: 10.32629/jai.v7i1.874.
22. J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, 1971, doi: 10.1037/h0031619.
23. A. Adesokan, S. Madria, and L. Nguyen, "HatEmoTweet: low-level emotion classifications and spatiotemporal trends of hate and offensive COVID-19 tweets," *Social Netw. Analysis and Mining*, vol. 13, pp. 136, 2023, doi: 10.1007/s13278-023-01132-6.
24. A. Adesokan and S. Madria, "NeuEmot: Mitigating Neutral Label and Reclassifying False Neutrals in the 2022 FIFA World Cup via Low-Level Emotion," in *Proceedings of the 2023 IEEE International Conference on Big Data*, 2023, pp. 578-587.
25. A. Adesokan, S. Madria, and L. Nguyen, "TweetACE: A Fine-grained Classification of Disaster Tweets using Transformer Model," in *Proceedings of the IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2023, pp. 1-9.
26. N. A. Semary, W. Ahmed, K. Amin, P. Pławiak, and M. Hammad, "Improving sentiment classification using a RoBERTa-based hybrid model," *Frontiers in Human Neuroscience*, vol. 17, 2023, doi: 10.3389/fnhum.2023.1292010, issn: 16625161.
27. M. Gheini, X. Ren, and J. May, "Cross-Attention is All You Need: Adapting Pretrained Transformers for Machine Translation," in *Proceedings of EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, doi: 10.18653/v1/2021.emnlp-main.132.
28. P. He, X. Liu, J. Gao, and W. Chen, "DEBERTA: Decoding-Enhanced BERT with Disentangled Attention," in *Proceedings of ICLR 2021 - 9th International Conference on Learning Representations*, 2021.
29. A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, issn: 10495258.
30. N. Shazeer, "GLU Variants Improve Transformer," *preprint arXiv:2002.05202*, 2020.
31. G. Burel, H. Saif, M. Fernandez, and H. Alani, "On Semantics and Deep Learning for Event Detection in Crisis Situations," presented at the Workshop on Semantic Deep Learning (SemDeep), 2017.
32. S. Das, K. Bhattacharyya, and S. Sarkar, "Performance Analysis of Logistic Regression, Naive Bayes, KNN, Decision Tree, Random Forest and SVM on Hate Speech Detection from Twitter," *International Research Journal of Innovations in Engineering and Technology*, vol. 07, no. 03, 2023, doi: 10.47001/irjiet/2023.703004.
33. F. Alam, H. Sajjad, M. Imran, and F. Ofli, "CrisisBench: Benchmarking Crisis-related Social Media Datasets for Humanitarian Information Processing," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, 2021, doi: 10.1609/icwsm.v15i1.18115.
34. F. Alam, S. Joty, and M. Imran, "Domain adaptation with adversarial training and graph embeddings," in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 1, 2018, doi: 10.18653/v1/p18-1099.
35. H. Li, D. Caragea, and C. Caragea, "Combining Self-Training with Deep Learning for Disaster Tweet Classification," in *Proceedings of the International ISCRAM Conference*, vol. 2021-May, 2021, issn: 24113387.
36. R. S. João, "On Informative Tweet Identification for Tracking Mass Events," in *Proceedings of ICAART 2021 - 13th International Conference on Agents and Artificial Intelligence*, vol. 2, 2021, doi: 10.5220/0010392712661273.
37. H. Li, X. Li, D. Caragea, and C. Caragea, "Comparison of Word Embeddings and Sentence Encodings as Generalized Representations for Crisis Tweet Classification Tasks," in *Proceedings of the ISCRAM Asian Pacific Conference*, Nov 2018.